«КАЗАНСКИЙ (ПРИВОЛЖСКИЙ) ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт Вычислительной Математики и Информационных Технологий

Ахметов И.З, Галимянов А.Ф

ЭЛЕМЕНТЫ МЕТОДОВ ОПТИМИЗАЦИИ

КАЗАНЬ – 2023

Оглавление

Введение

В данной учебном пособии будет попытка последовательно, понятно изложить симплекс-метод. По опыту авторов книги простое заучивание симплекс-метода как некой таблицы с кучей разных условий и алгоритмов крайне неэффективно. В этом учебном пособии предложен другой подход к его понимаю: с помощью понятия градиента. Так же в книге дано подробное описание линейной регрессии с различных сторон, рассмотрены теоретические проблемы ее применения.

Книга не претендует на абсолютную математическую строгость и точность. Главной целью было сделать материал доступным для понимания студентами. Материал снабжен минимально необходимой теорией, и, насколько это возможно, освобожден от излишней математизации. Акцент сделан на понимание интуиции и идей, стоящих за линейной регрессией и симплекс-методом.

Данная версия является черновой. Возможно, в ней имеются опечатки и ошибки. В дальнейшем в нее будут вноситься дополнительные главы и исправления. В книге имеются приводятся доказательства некоторых утверждений, теорем. При возникновении трудностей при первом чтении их можно пропустить, обращая внимания лишь на формулировки.

В книге имеются так же упражнения, большая часть которых приводится с решениями. Рекомендуется сначала попытаться решить их самостоятельно прежде, чем смотреть решение.

От студентов для понимания большей части книги требуются прочные знания по математическому анализу, линейной алгебре и аналитической геометрии за 1 курс. Для понимания отдельных глав, связанных с линейной регрессией так же необходимо знание теории вероятностей и математической статистики. Пособие может быть полезно студентам-бакалаврам, а так же магистрам и всем тем, кто интересуется линейной регрессией и симплекс-методом. Так же оно может быть использовано как дополнительный материал при изучения основ машинного обучения.

Глава 1

Линейная модель аппроксимации.

1.1 Линейная регрессия. Постановка задачи.

Линейная регрессия - одна из широко распространенных моделей зависимости одной переменной от другой либо набора других переменных, которую можно свести к аппроксимации одного вектора другим либо набором других векторов.

Пусть нас имеется некая величина Y, которая, как предполагается, линейно зависит от одной либо большего числа переменных $X_1, X_2 ... X_m$. К примеру, пусть Y - это цена на квартиру. Она зависит от X_1 - количество квадратных метров, X_2 - удаленность от центра, X_3 - население города, ... X_m - положение звезд на небе. Важно, чтобы $\forall i \; X_i \in R$. Допустим, что у нас имеются основания полагать, что эта зависимость линейная, то есть

$$Y = a_0 + a_1 X_1 + a_2 X_2 + ... + a_m X_m + \varepsilon \tag{1.1}$$

где $\mathbf{a_0}$ - некое неизвестное число - константа, а ε - случайная величина, ошибка с математическим ожиданием μ, равным нулю и постоянной дисперсией σ^2, то есть имеющая нормальное распределение $\varepsilon \sim N(0, \sigma^2)$. Это очень важно, т.к в противное означает, что линейная регрессионная модель не способна корректно аппроксимировать данную зависимость, ее нельзя здесь применять. Пока что не будем на этом подробно останавливаться, в этом разделе предполагаем, что ε удовлетворяет необходимым свойствам и более не упоминаем. Также имеется n значений Y и n значений соответствующих переменных X_1, X_2, ... X_m, единичный вектор $\vec{B}$, умноженный на постоянную a_0. Допустим, эти данные получены в ходе наблюдений цен на квартиры в произвольном городе. То есть в данном случае у нас имеются данные о ценах квартир Y и соответствующих им признаков X_1, X_2, ... Xm. Пусть у нас имеется $\mathbf{n}$ таких наблюдений. Обычно $\mathbf{n}$ намного больше чем $\mathbf{m}$, т.е количество наблюдений гораздо больше количества признаков, которыми можно описать объект. В реальной жизни это естественно, ведь квартир сотни миллионов, а существенных признаков, по которым можно судить об их цене - десятки.

Всё это можно записать следующим образом

$$\begin{aligned} Y_1 &= a_0 + a_1 x_{11} + a_2 x_{12} + ... + a_m x_{1m} \\ Y_2 &= a_0 + a_1 x_{21} + a_2 x_{22} + ... + a_m x_{2m} \\ Y_3 &= a_0 + a_3 x_{31} + a_2 x_{32} + ... + a_m x_{3m} \\ &\vdots \\ Y_n &= a_0 + a_1 x_{n1} + a_2 x_{n2} + ... + a_m x_{nm} \end{aligned} \tag{1.2}$$

или, в векторном виде

$$\vec{Y} = a_0 \vec{A_0} + a_1 \vec{X_1} + a_2 \vec{X_2} + ... + a_m \vec{X_m} \tag{1.3}$$

либо, в матричном виде

$$Y = Xa \tag{1.4}$$

где a - вектор коэффициентов:

$$a = (a_0, a_1, a_2, ...a_m)$$

а X - матрица значений

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & .... & x_{1m} \\ 1 & x_{21} & x_{22} & .... & x_{2m} \\ ..... & & & & \\ 1 & x_{n1} & x_{n2} & .... & x_{nm} \end{pmatrix}$$

Столбцы $x_{ij}, j \in [1:m]$ представляют из себя вектор значений переменных $X_1, X_2 ... X_m$ соответственно. Первый столбец состоящий из единиц введен в связи с наличием постоянного значения a_0 и соответствующего ему вектора, все значения которого равны 1. Вообще, единичный вектор, умноженный на коэффициент a_0 вводится в связи с отсутствием центрированности данных, т.к средние значения вектора Y и зависимых от него переменных X_1, X_2, ... X_m в общем случае не равны нулю. Если же это не так, то коэффициент a_0 будет равен нулю. Но вероятность этого практически нулевая. Значение a_0 здесь - это так называемый коэффициент смещения (bias). Если его не вводить, то линейная регрессия не сможет корректно отображать смещенные относительно начала координат данные (например, если все значения искомой величины Y положительны, а не разбросаны более-менее равномерно около нуля). Другими словами, в случае равенства коэффициента a_0 нулю поверхность $Y = a_0 + a_1 X_1 + a_2 X_2 + ... + a_m X_m$ будет обязательно проходит через начало координат $(0, 0....0)$, что негативно скажется на способности линейной модели предсказывать смещенные относительно начала координат данные.

Замечание: матричная форма записи (1.4) может быть так же записана в виде

$$Y = aX \tag{1.5}$$

если положить, что a - это вектор-строка (1.1) коэффициентов, либо в форме

$$Y = a^T X \tag{1.6}$$

если a - это вектор-строка $\in R^{1,m+1}$

$$a = (a_0, a_1, a_2, ...a_m)$$

В обоих альтернативных способах представления матрица данных будет X представлена в виде

$$X = \begin{pmatrix} 1 & 1 & .... & 1 \\ x_{11} & x_{12} & .... & x_{1n} \\ x_{21} & x_{22} & .... & x_{2n} \\ ..... & & & \\ x_{m1} & x_{m2} & .... & x_{mn} \end{pmatrix}$$

Такой способ представления так же оправдан, поскольку обычно оператор, "воздействующий"на объект ставят слева от объекта. Оператор - это некое обобщение понятия функции. Так же как и функция, оператор отображает множество одних объектов на другое множество. В нашем случае оператор - это вектор коэффициентов **a**. (1.1), а объект - данные X.

Если количество данных наблюдений, т.е число n не равно числу зависимых переменных m (обычно оно много больше чем m) то очевидно, что решения, скорее всего, нет. Чтобы убедиться в этом, достаточно ради примера взять систему из трех уравнений и двух переменных, такую, что две строки в ней линейно независимы.

$$\begin{aligned} x - 5y &= 2 \\ 2x + 3y &= 5 \\ -3x + 7y &= 8 \end{aligned} \tag{1.7}$$

После упрощения получим

$$\begin{aligned} x - 5y &= 2 \\ 13y &= 1 \\ -8y &= 4 \end{aligned} \tag{1.8}$$

Очевидно, что y не может равняться одновременно $-\frac{1}{2}$ и $\frac{1}{13}$ и система не имеет решения. При большем количестве уравнений решения не будет подавно.

Теорема Кронекера-Капелли:

Система линейных алгебраических уравнений совместна тогда и только тогда, когда ранг её основной матрицы равен рангу её расширенной матрицы.

При количестве уравнений, много большем количества переменных ранг основной матрицы (без Y) почти наверняка будет больше ранга расширенной матрицы, т.к систему всегда можно привести к виду, где количество ненулевых элементов в векторе свободных коэффициентов Y не больше 1 (объясните почему). При этом ранг основной матрицы, состоящих из коэффициентов признаков равен 1 только если все строки прямо пропорциональны какой-то одной строке, т.е существует строка с индексом t, такая

что для любой другой строки с индексом i существует число $k \in R : \forall j\, X_{ij} = kX_{tj}$, вероятность чего при случайных данных практически нулевая. Поэтому решения у такой системы не будет.

Поскольку это естественно для реальных экспериментов, наблюдений, (количество объектов в данных много больше количества признаков) то именно такой случай мы и будем рассматривать.

1.2 Минимизация невязки

Невязкой в случае системы (1.18) называется разность правой и левой частей системы уравнений. Иными словами, это вектор, равный разности исходного Y и его аппроксимации векторами $X_1, X_2, ...X_m$. Конечно, данное уравнение как было показано раннее не имеет ни единственного, ни вообще какого либо решения при $n >> m$, но можно подобрать коэффициенты $a_1, a_2...a_m$ таким образом, чтобы найти "наиболее похожее на правду"решение, при котором значение нормы невязки будет минимально возможным для данной системы, то есть

$$\arg\min_a ||Y - Xa|| \tag{1.9}$$

Здесь и далее я буду использовать понятие нормы вектора вместо понятия длина. Почему? Норма - это широкораспространенное понятие в функциональном анализе. Оно обобщает понятия длина вектора, объем какого-либо объекта (например шара) и т.д. Это как-бы способ количественного измерения объекта. Норма всегда неотрицательна. Ее введение продиктовано также тем, что "количественное значение"объекта может измеряться по разному, в зависимости от специфики задачи. Например, длину двумерного вектора можно измерить как

$$\sqrt{x^2 + y^2}$$

А можно и так

$$\|x\| + \|y\|$$

Последнее - это так называемое манхэттенское расстояние. Оно обусловлено тем, что идя в городе по улице из точки А в точку Б нельзя физически проходить сквозь здания по диагонали, можно идти только вдоль домов, по прямоугольной сетке Oxy.

Вернемся к основной теме. Эту задачу можно рассмотреть в многомерном евклидовом пространстве, как проекцию одного вектора на другой, либо на плоскость, образованное двумя векторами. В общем случае на подпространство того пространства, в котором определен вектор Y. Это подпространство образовано векторами $X_1, ...X_m$, представляющими его базис. Число m может быть равно 1, а может быть и сколько угодно большим. Другими словами, стоит задача "выразить Y через векторы $X_1, ...X_m$ в виде суммы их линейной комбинации. Норма разности между исходным вектором Y и вектором Y, и выраженным через $X_1, ...X_m$ и будет здесь невязкой. Очевидно, для наилучшего приближения необходимо сделать так, чтобы аппроксимированный вектор Y

отличался от исходного как можно меньше, чтобы потерять как можно меньше исходной информации об Y. Это достигается минимизацией невязки. Для минимизации невязка должна быть ортогональна плоскости, на которую проецируется вектор Y. Докажем это, введя пару определений.

Определение 1. Пусть P - оператор, проецирующий вектор Y на подпространство X, образованное векторами $X_1, ...X_m$. Ортогональной проекцией вектора Y на подпространство X называется вектор P(Y) такой, что вектор разности $Y - P(Y)$ ортогонален $\forall x \in X$, то есть скалярное произведение $(Y - P(Y), x), \forall x \in X$, а вектор $Y - P(Y)$ ортогонален всему подпространству X. P(Y), очевидно, в данном случае является линейной комбинацией $X_1, ...X_m$.

Определение 2. Невязка проекции - вектор, равный разности вектора Y и его проекции P(Y) на подпространство X. В общем случае не обязана быть ортогональной X.

Утверждение 1. Ортогональная проекция P(Y) вектора Y на подпространство X определяется единственным образом.

Доказательство: действительно, пусть существует два различных вектора G(Y), P(Y), таких, что их невязки ортогональны X: $Y - P(Y) \perp x, Y - G(Y) \perp x, \forall x \in X$. В данном случае G - другой оператор, проекции Y на X. Рассмотрим скалярное произведение $(P(Y) - G(Y), x) = (P(Y) - G(Y) + Y - Y, x) = (Y - G(Y) - (Y - P(Y)), x) = 0 - 0 = 0$. Это возможно для $\forall x \in X$, только если $P(Y) = G(Y)$. Следовательно, проекция единственна.

Утверждение 2. Пусть P(Y) - ортогональная проекция Y на подпространство X. Тогда норма(в нашем случае это просто евклидова длина вектора) невязки $Y - P(Y)$ является минимальным расстоянием от вектора Y до подпространства X.

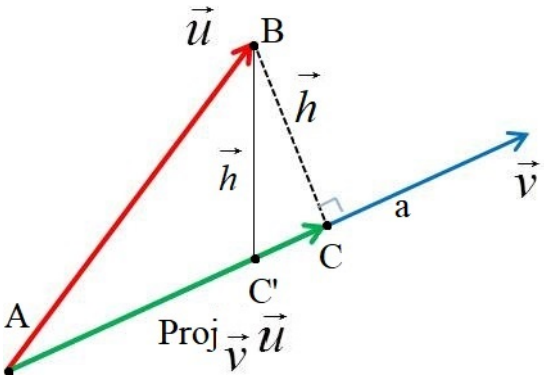
Доказательство: Пусть x - произвольный вектор из X. Расстояние между вектором Y и подпространством X, как известно, определяется выражением, минимизирующим норму разности вектора Y и произвольного $x \in X : ||X - Y|| = argmin_x||Y - x||$. Норма в евклидовом пространстве определяется через скалярное произведение как $\sqrt{(\vec{a})(\vec{a})}$. Докажем, что $\sqrt{(Y - P(Y))(Y - P(Y))}$ меньше $\sqrt{(Y - x)(Y - x)}, \forall x \in X$.

Напомним, что скалярное произведение вектора на себя всегда неотрицательно. Рассмотрим выражением $(Y - x)(Y - x) = (Y - P(Y) + P(Y) - x)(Y - P(Y) + P(Y) - x) = (Y - P(Y))(Y - P(Y)) + 2(Y - P(Y))(P(Y) - x) + (P(Y) - x)(P(Y) - x)$. Значение $(Y - P(Y))(P(Y) - x)$ обращается в ноль, так как $P(Y) \in X$, следовательно, $P(Y) - x \in X$, а $Y - P(Y)$ ортогонально любому элементу из X. Поскольку $(P(Y) - x)(P(Y) - x)$ не может быть отрицательным, то $(Y - P(Y))(Y - P(Y))$ не может быть меньше $(Y - x)(Y - x)$. Что и требовалось доказать.

Таким образом, мы доказали в общем случае, что норма разности вектора Y и его проекции $P(Y)$ на подпространство X, образованное векторами $X_1, ...X_m$ минимально тогда, когда эта разность $Y - P(Y)$ ортогональна подпространству X, а значит, и любому вектору $x \in X$. В основу данных доказательств взяты материалы из книги [5] на стр. 25-26.

Для иллюстрации доказательства и графической иллюстрации рассмотрим несколь-

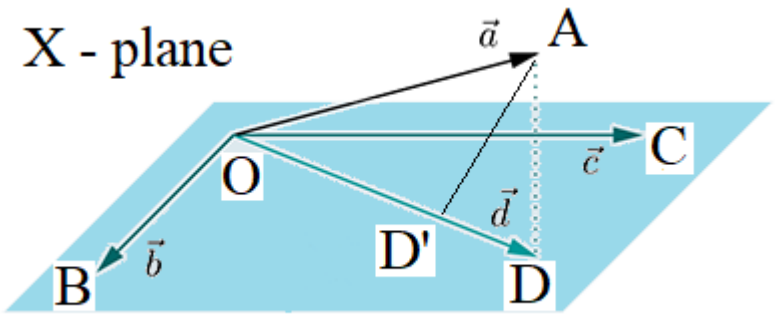
ко примеров. Начнем со случая проекции одного вектора на другой.



Упражнение: Пользуясь рисунком выше, докажите, что кратчайшее расстояние от точки **B** до прямой **a** есть длина отрезка **BC**, ортогонального **a**, а не, например, **BC'**. Здесь проекцией отрезка **AB** на прямую **a** будет отрезок **AC**. Подсказка: Для этого достаточно геометрии за 8 класс.

Решение: рассмотрим треугольник **BCC'**. Он является прямоугольным, так как отрезок **BC** ортогонален прямой **a**. Так как гипотенуза **CC'** не может быть меньше катета, то получаем, что длина отрезка **BC** есть наименьшее расстояние. Если внимательно посмотреть, то любопытно заметить, что схема этого доказательства такая же, как и для абстрактного доказательства в общем случае, приведенного выше.

В данном случае вектор $\vec{u}$ проецируется на вектор $\vec{v}$. Фактически подпространство X, на которое проецируется вектор $\vec{u}$ в этом случае состоит из одного вектора $\vec{v}$. Обозначим эту проекцию как $Proj_{\vec{v}}\vec{u}$ Изображение схематично, оба вектора могут находиться в многомерном пространстве и иметь размерность, например, 100, т.е принадлежать пространству R^{100}. Двумерное пространство выбрано для иллюстрации. Невязка в данном случае - это разность вектора $\vec{u}$ и его проекции на вектор $\vec{v}$. В этом случае она ортогональна вектору $\vec{v}$, хотя в целом невязка не обязана быть ортогональна. Обозначим невязку как вектор $\vec{h} = u \vec{-} v$, а норма невязки - т.е расстояние от $\vec{u}$ до $\vec{v}$, т.е длина вектора $\vec{h}$. Достаточно просто доказать, что для того, чтобы норма данной невязки была минимальна, необходимо, чтобы невязка была ортогональна вектору $\vec{v}$, а не $\vec{u}$, и вообще минимум будет только в случае ортогональности к $\vec{v}$ вектора, исходящего из конца вектора $\vec{u}$. Рекомендуется попробовать доказать это для частного случая евклидова пространства, размерность при этом не имеет значения (можно взять двумерное пространство), нужно лишь воспользоваться понятием длины вектора и скалярного произведения.



Упражнение: Докажите, что ортогональный к плоскости **X** отрезок **AD** есть кратчайшее расстояние от точки **A** до плоскости **X**, т.е что любой другой отрезок, исходящий из A и заканчивающийся на плоскости **X**, например **AD'** будет длиннее чем **AD**. В данном случае проекцией отрезка **OA** на плоскость X будет отрезок **OD**.

Решение: аналогично этой же задаче в двумерном случае.

Перейдем теперь к случаю многих переменных. Как видно из данного рисунка, вектор $\vec{a}$) аппроксимируется на подпространство, образуемое неколлинеарными векторами $\vec{b}$, $\vec{c}$. Его проекцией на эту плоскость является вектор $\vec{d}$. Данное изображение так же схематично, векторы могут находиться многомерном евклидовом пространстве, например R^{5000}, но векторы $\vec{b}$, $\vec{c}$ при этом могут образовывать свое подпространство, на которое можно проецировать любой вектор, не принадлежащий ему. Так же в многомерном пространстве возможно аппроксимировать вектор не только на подпространство, образованную двумя векторами (в случае трехмерного пространства это будет привычная нам плоскость), но и на многомерное подпространство, базисом которого является, например, 50 векторов, с размерностью как минимум на 1 меньшей чем размерность пространства исходного.

Вообще, векторы $\vec{b}$, $\vec{c}$ не обязаны быть перпендикулярны, главное, чтобы они были линейно независимы. В данном случае разность векторов $\vec{a}$ и $\vec{d}$ является невязкой проекции. Ее модуль равен расстоянию от вектора $\vec{a}$ до плоскости X. Несложно доказать, что для того, чтобы ее длина была минимальна, она должна быть ортогональна к плоскости, образуемой векторами $\vec{b}$, $\vec{c}$. В этом случае ее модуль принимает наименьшее значение. А если невязка ортогональна к плоскости, то на ортогональна ко всем векторам данной плоскости. Это известно еще со школьного курса геометрии. А если она ортогональна ко всем векторам плоскости, значит, скалярное произведение невязки на векторы плоскости равно нулю. В таком случае она так же ортогональна $\vec{b}$ и $\vec{c}$. Данный вывод можно записать в следующем виде:

$$\begin{aligned}(\vec{a}-\vec{d},\vec{b})=0\\(\vec{a}-\vec{d},\vec{c})=0\end{aligned} \tag{1.10}$$

Теперь, после поясняющих рисунков вернемся к проблеме (1.9). Вектора Y и $X_1, X_2, ...X_m \in R^n$, где n - сколько угодно большое число (обычно оно равно количеству наблюдений, экспериментов). В данном случае многомерный вектор Y проецируется на плоскость, образуемую векторами $X_1, X_2, ...X_m$, m - количество таких векторов.

Нам нужно представить вектор Y с помощью векторов $X_1, X_2, ...X_m$ и вектора коэффициентов $a = (a_0, a_1, a_2, ...a_m)$ в виде:

$$Y = a_0 + \sum_{i=1}^{m} X_i a_i$$

Как отмечено выше, разность между Y и его проекцией, т.е невязка ортогональна к плоскости, образованной векторами $X_1, X_2, ...X_m$, а значит, и ко всем векторам $X_1, X_2, ...X_m$. Это условие можно записать как:

$$\begin{aligned} &(\vec{X_1}^T, \vec{Y} - Xa) = 0 \\ &(\vec{X_2}^T, \vec{Y} - Xa) = 0 \\ &..... \\ &(\vec{X_m}^T, \vec{Y} - Xa) = 0 \end{aligned} \tag{1.11}$$

или, короче,

$$(X^T, Y - Xa) = 0 \tag{1.12}$$

Решим уравнение (1.12). Пользуясь свойством линейности скалярного произведения:

$$\begin{aligned} &(X^T, Y - Xa) = (X^T, Y) - (X^T, Xa) = 0 \\ &(X^T X)a = X^T Y \\ &a = (X^T X)^{-1} X^T Y \end{aligned} \tag{1.13}$$

Таким образом, мы получили формулу для аналитического способа нахождения коэффициентов, удовлетворяющих условию (1.9). Здесь $a = (a_1, a_2, ...a_m, b)$ - столбец коэффициентов, X - выражение (1.8) - массив данных, Y - вектор значений аппроксимируемой переменной.

Интересно отметить, что таким же образом можно ортогонально проектировать не только какой-нибудь вектор $y \in R^n$ на подпространство, образованное наборов других векторов, которые так же принадлежат R^n, но и функцию $f(x_1, ..., x_n)$ на набор функций $g_i(x_1, ..., x_n), i \in 1, 2, ..., k$, где k - количество функций, из которых образуется подпространство, на которое проектируется функция $f(x_1, ..., x_n)$, n - количество переменных, от которых зависят функции. Значение n может быть равно 1, а может быть и очень большим. В итоге получается ортогональная проекция, являющаяся наилучшим приближением функции $f(\vec{x})$ с помощью набора функций $g_i(\vec{x})$.

$$f(\vec{x}) = \sum_{i=1}^{k} a_i g_i(\vec{x}) + b$$

Данная идея является основой проекционных методов, применяемых для численного решения дифференциальных уравнений. Но это уже выходит за рамки данной книги.

Упражнение 1. Проанализируйте формулу (1.13). Что будет, если какие-либо 2 признака (т.е какие-нибудь 2 столбца матрицы X) будут линейно зависимы. Возможно ли будет тогда вычислить **a**?

Решение: Предположим, что имеется k признаков, $k >= 2$. Пусть X_i - i-й вектор-столбец матрицы. Допустим, что первый и второй признаки линейно зависимы, т.е $\exists\lambda \in R, \lambda \neq 0 : X_2 = \lambda X_1$. Обозначим X_iX_j как скалярное произведение i-го и j-го столбцов. Найдем значение (X^TX)

$$\begin{pmatrix} X_1 & X_2 & ... & X_k \end{pmatrix}^T \begin{pmatrix} X_1 & X_2 & ... & X_k \end{pmatrix} = \begin{pmatrix} X_1^2 & \lambda X_1^2 & X_1X_3 & ... & X_1X_k \\ \lambda X_1^2 & \lambda^2 X_1^2 & \lambda X_1X_3 & ... & \lambda X_1X_k \\ X_3X_1 & X_3X_2 & X_3^2 & ... & X_3X_k \\ ... & & & & \\ X_kX_1 & X_kX_2 & X_kX_3 & ... & X_k^2 \end{pmatrix}$$

Как видно, первая и вторая стока получившейся матрицы линейно зависимы. Элементарные преобразования матрицы, как известно, не меняют определитель. Вычитая из второй строки первую умноженную на λ мы получим нулевую строку. Доказано, что определитель матрицы с нулевой строкой равен нулю. Определитель матрицы (X^TX) будет равен нулю, следовательно у этой матрицы не будет обратной и найти коэффициенты линейной регрессии будет невозможно.

нет, т.к определитель матрицы (X^TX) будет равен нулю.

Упражнение 2. Проанализируйте формулу (1.13). Предположим, что в какие-либо 2 признака (т.е какие-нибудь 2 столбца матрицы X) будут сильно коррелированы с коэффициентом корреляции, равным по модулю ~ 0.95, то есть "почти"линейно зависимы. Пусть мы нашли коэффициенты **a**. Что произойдет с **a**, если добавить немного данных, т.е увеличить матрицу X небольшим количеством новых строк? Сильно ли изменится их значение коэффициентов **a**? Как можно охарактеризовать такую линейную модель?

Решение: В случае сильной коррелированности каких-либо двух признаков определитель матрицы (X^TX) будет близок к нулю. Тогда даже при малом изменении массива данных определитель матрицы $(X^TX)^{-1}$ очень сильно изменит свое значение, а вместе с ним и значения коэффициентов линейной регрессии. Такая модель будет неустойчивой, так как при малом изменении входных данных коэффициенты будут изменяться очень сильно.

1.3 Примеры

1.3.1 Случай одной переменной №1

В ходе эксперимента, который повторили 3 раза, исследователи, взглянув на график, предположили, что величина y (вектор правой части уравнения ниже) зависит от x (первый вектор в правой части) следующим образом

$$\begin{pmatrix} 5 \\ 5 \\ 9 \end{pmatrix} = a \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} - b \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \tag{1.14}$$

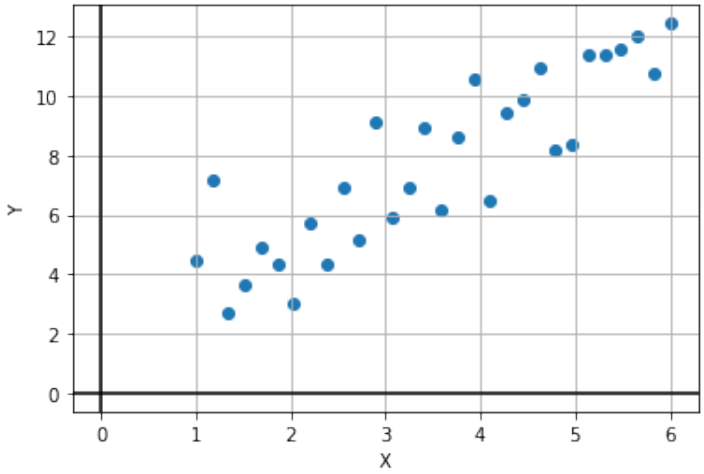
Очевидно, что система не имеет точного решения. Но можно найти приближенное. Найдем его в виде $y = ax + b$. Умножим систему скалярно сначала на вектор при коэффициенте a, затем на константный вектор $\vec{b}$ в правой части. Получим систему из двух уравнений. Таким образом, мы проектируем вектор в левой части на вектор в правой, равный сумме вектора $\vec{x}$ и константного $\vec{b}$. Получим систему:

$$\begin{aligned} 42 &= 14a_1 - 6b \\ 19 &= 6a_1 - 3b \end{aligned} \tag{1.15}$$

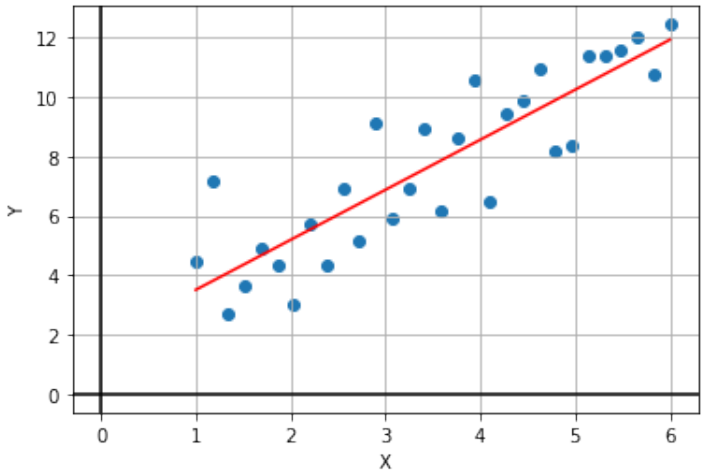
решая эту систему, получим ответ: $y = -7/3 + 2x$ - наилучшее приближение

1.3.2 Случай одной переменной №2

Пусть в результате эксперимента получили вот такую зависимость некой величины Y от X. Всего у нас 30 наблюдений



Здесь явно прослеживается некая линейная зависимость. В реальных экспериментах к значениям датчиков часто добавляется некий шум. Здесь не будут показаны утомительные массивы чисел. Представить их вы можете и сами, взглянув на график. Покажу лишь аппроксимацию, полученную в итоге по формуле (1.13).



Как видно, зависимость аппроксимирована линейной функцией. Ее формула: $Y = 1.82 + 1.68X$.

1.3.3 Случай многомерной регрессии

В данном примере предположим, что мы рассматриваем зависимость стоимости акций Y от двух переменных: индекса интереса к ним X_1 и процента безработицы X_2

$Y = (1464, 1394, 1357, 1293, 1256, 1254, 1234, 1195, 1159, 1167, 1130, 1075)$,

$X_1 = (2.75, 2.5, 2.5, 2.5, 2.5, 2.5, 2.5, 2.25, 2.25, 2.25, 2, 2)$,

$X_2 = (5.3, 5.3, 5.3, 5.3, 5.4, 5.6, 5.5, 5.5, 5.5, 5.6, 5.7, 5.9)$.

Перепишем это в матричном виде, добавив константу b. Получим

Данное уравнение, очевидно, так же согласно теореме Кронекера-Капелли не имеет решения. Но можно найти приближенное, минимизирующее невязку. Перепишем в матричном виде :

$$\begin{pmatrix} 1464 \\ 1394 \\ 1357 \\ 1293 \\ 1256 \\ 1254 \\ 1234 \\ 1195 \\ 1159 \\ 1167 \\ 1130 \\ 1075 \end{pmatrix} = \begin{pmatrix} 1 & 2.75 & 5.3 \\ 1 & 2.5 & 5.3 \\ 1 & 2.5 & 5.3 \\ 1 & 2.5 & 5.3 \\ 1 & 2.5 & 5.4 \\ 1 & 2.5 & 5.6 \\ 1 & 2.5 & 5.5 \\ 1 & 2.25 & 5.5 \\ 1 & 2.25 & 5.5 \\ 1 & 2.25 & 5.6 \\ 1 & 2 & 5.7 \\ 1 & 2 & 5.9 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} \tag{1.16}$$

или, коротко:

$$\Big(Y = Xa\Big) \tag{1.17}$$

где X - матрица, столбцы которой есть векторы $\vec{X_1}, \vec{X_2}, \vec{col_o nes}$, где $\vec{col_o nes}$ - вектор-столбец, состоящий из единиц, a - вектор-столбец коэффициентов (a_0, a_1, a_2), которые необходимо найти.

Пользуясь формулой (1.13) получим решение: $a = (1818.17, 294.87, -231.32,)$, т.е ответ: $Y = 1818.17 + 294.87X_1 - 231.32X_2$

Упражнение: на любом известном вам языке программирования напишите программу для линейной аппроксимации вектора Y набором m векторов $X_1, X_2 ... X_m$, принимающую сначала на вход количество признаков m и наблюдений n, затем массив, состоящий из m + 1 столбцов и n строк. Первый столбец - значения Y. Последущие стобцы - значения X_1, X_2, .. X_m по порядку. На выходе программа должна возвращать вектор коэффициентов $(a_0, a_1, a_2, ..., a_m)$ через пробел с точностью до двух знаков, с помощью которых вектор Y аппроксимируется векторами $X_1, X_2 ... X_m$. Отладьте программу на юнит-тестах, к примеру можете воспользоваться данными из этого раздела. Для нахождения коэффициентов воспользуйтесь формулой (1.13).

Пример ввода-вывода. Входные данные (из примера 1).

1 3
5 1
5 2
9 3

Вывод: -2.33 2

1.4 Полиномиальная регрессия.

Пусть у нас так же имеется n наблюдений величины Y, зависящей от n наблюдений величины X. Допустим, что по каким-то причинам (например, судя по графику) имеется основание предположить, что зависимость Y от X нелинейная, а задается с помощью полинома. Что делать в данном случае?

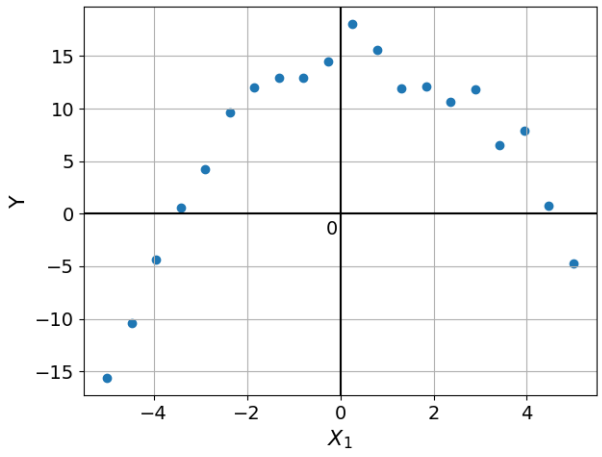
Для примера рассмотрим случай, когда у нас дана зависимая переменная Y, независимая X, и некий набор их значений. Допустим, у нас есть основания думать, что зависимость квадратичная, то есть $Y = b + a_1X + a_2X^2$. При этом имеется лишь массив наблюдений Y и X, но не X^2. В данном случае нужно добавить новый массив X^2, который получается путем поэлементного возведения в квадрат элементов массива X. На языке программирования Python это можно было бы записать вот так: X^2 = [element*element for element in X]. Далее, чтобы получить значения коэффициентов (a_1, a_2, b) нужно действовать так же, как и с обычной линейной регрессией, то есть найти нужные коэффициенты по формуле (1.13). Будет понятнее на примере.

Пусть у нас имеется массив данных Y и X:

Y = [-14.12, -10.15, 0.33, -0.63, 3.94, 10.19, 7.95, 10.77, 13.96, 14.01, 12.81, 14.76, 13.87, 14.66, 8.43, 8.11, 9.02, 7.08, -3.56, -3.71],

X = [-5. , -4.47, -3.95, -3.42, -2.89, -2.37, -1.84, -1.32, -0.79, -0.26, 0.26, 0.79, 1.32, 1.84, 2.37, 2.89, 3.42, 3.95, 4.47, 5.],

Данная зависимость представлена на графике. Интуитивно понятно, что она представляет собой параболу. В таком случае, нужно добавить X^2 для уравнения параболы, прямая линия для аппроксимации в данном случае не подойдет.



Обозначим X как X_1, искусственно создадим еще одну переменную $X_2 = X_1 * X_1 = x * x : x \in X_1$, которая является по сути множеством значений из X, возведенных в квадрат

Далее, используя массивы данных Y, X_1, X_2, запишем подробно $Y = Xa$

$$\vec{Y} = [X_1, X_2, b]\vec{a} \tag{1.18}$$

или, в еще более подробном виде

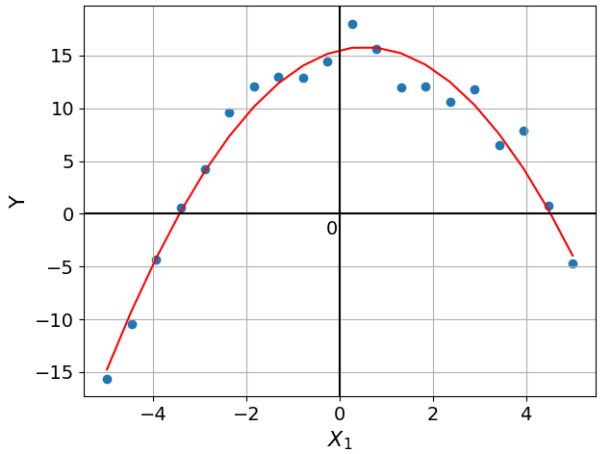
$$\begin{pmatrix} -14.12 \\ -10.15 \\ 0.33 \\ -0.63 \\ 3.94 \\ 10.19 \\ 7.95 \\ 10.77 \\ 13.96 \\ 14.01 \\ 12.81 \\ 14.76 \\ 13.87 \\ 14.66 \\ 8.43 \\ 8.11 \\ 9.02 \\ 7.08 \\ -3.56 \\ -3.71 \end{pmatrix} = \begin{pmatrix} 1 & -5 & 25 \\ 1 & -4.47 & 20.01 \\ 1 & -3.95 & 15.58 \\ 1 & -3.42 & 11.7 \\ 1 & -2.89 & 8.38 \\ 1 & -2.37 & 5.61 \\ 1 & -1.84 & 3.39 \\ 1 & -1.32 & 1.73 \\ 1 & -0.79 & 0.62 \\ 1 & -0.26 & 0.07 \\ 1 & 0.26 & 0.07 \\ 1 & 0.79 & 0.62 \\ 1 & 1.32 & 1.73 \\ 1 & 1.84 & 3.39 \\ 1 & 2.37 & 5.61 \\ 1 & 2.89 & 8.38 \\ 1 & 3.42 & 11.7 \\ 1 & 3.95 & 15.58 \\ 1 & 4.47 & 20.01 \\ 1 & 5 & 26 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ b \end{pmatrix} \tag{1.19}$$

Решая данную систему с помощью уже знакомой формулы (1.13), получим ответ: $(a_0, a_1, a_2, b) = [14.45, 1.05, -1.03]$. Мы получили полиномиальную регрессию следующего вида:

$$y = 14.45 + 1.05x - 1.03x^2$$

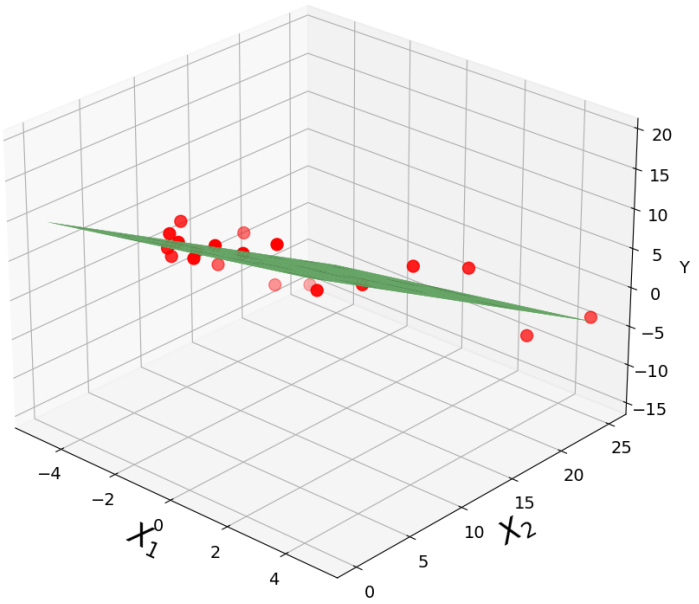
Очень близко к истинной зависимости $y = 15 + x - x^2$. Данные y выше были получены путем добавления к значениям y некоторого шума ε, имеющего нормальное распределение ε $N(0, \sqrt{2})$. Реальные данные практически всегда приходят с небольшой ошибкой из-за помех, несовершенства точности измерений и в целом вклада случайных непредвиденных событий.

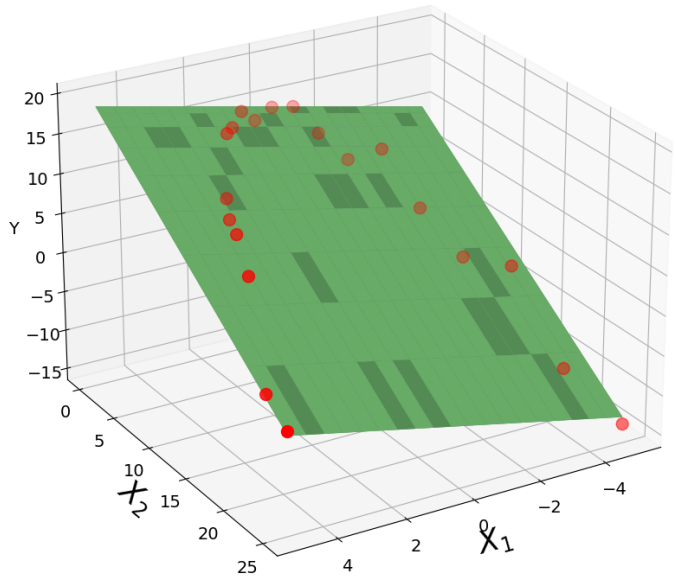
График аппроксимации:



При необходимости можно попробовать добавить и другие признаки, к примеру такие как X в третей и любой другой степени (в данном случае добавление этого признака не нужно и даже вредно, т.к создает дополнительный шум, под который регрессия будет подстраиваться, игнорируя общие закономерности).

То что мы проделали выше можно представить как переход из пространства переменных (X, Y) в пространство $(X, X*X, Y)$ с идеей, что возможно в новом пространстве существует линейная функция зависимости между Y и нашими данными. Действительно, построим график полученной аппроксимации в трехмерном пространстве с двух разных ракурсов:





Как видно, в трехмерном пространстве массив Y отлично аппроксимируется линейной функцией, что было невозможно в двумерном пространстве.

Совершенно аналогично, если у нас имеется массив значений целевой функции Z и зависимых переменных X, Y мы можем аппроксимировать Z формулой $Z = a_1X^2 + a_2XY + a_3Y^2 + a_4X + a_5Y + b$ путем поэлементного возведения в степень значений X и Y. XY получается путем поэлементного умножения элементов $x_i \in X$ на соответствующие элементы y_i.

В данном разделе мы рассмотрели полиномиальную регрессию, которая так же является линейной моделью.

Замечание: конечно, для того, чтобы график прошел в точности через N точек можно увеличить степернь полинома до N-1 (см интерполяционный многочлен Лагранжа) . Однако таким образом аппроксимация получится слишком заточенной под набор имеющихся данных. При этом обобщающая способность модели, т.е корректно предсказывать значение $y = f(x)$ для неизвестных x, т.е для тех x, которые не были использованы при построении аппроксимации будет ужасной В таких случаях говорят, что модель аппроксимации "переобучилась"на обучающих данных. По нашему субъективному представлению термин "переобуаться"здесь не совсем подходит. В англоязычной литературе используется "overfitted model что означает: слишком сильно "подстроилась "под обучающие данные.

Упражнение 1:

Пусть у нас есть данные x = [-2. , -1.79, -1.58, -1.37, -1.16, -0.95, -0.74, -0.53, -0.32, -0.11,

0.11, 0.32, 0.53, 0.74, 0.95, 1.16, 1.37, 1.58, 1.79, 2.]
y = [-67.94, -49.68, -35.62, -23.35, -10.8 , -5.5 , -2.96, -0.99, 2.46, 3.92, 8.1 , 5.23, 8.62, 10.07, 12.28, 15.58, 18.17, 24.6 , 23.53, 31.68]
y зависит от x как полином некоторой степени. Установите наиболее подходящую степень многочлена для аппроксимации этих данных, найдите его коэффициенты. Помните, что более высокая степень многочлена не обязательно является лучшей моделью аппроксимации. Можете отложить часть имеющихся данных для тестирования модели, чтобы видеть, какая аппроксимация лучше всего работает для неизвестных ей x. Очевидно, что на тестовых данных настраивать аппроксимацию нельзя. Можно выбрать разные метрики для оценки качества. Остановимся на среднем квадрате ошибки. Для тестирования выделяют обычно $20-30\%$ имеющихся данных. Иными словами, вам нужно найти такой многочлен $f(x)$, что $argmin_{f(x)} \frac{\sum((f(x_i)-y_i)^2}{N}, x_i \in test\_data, N -$

Решение: воспользовавшись программой для нахождения линейной регрессии, перебрать полиномы вплоть до некоторой степени k и сравнить среднюю ошибку на тестовых данных, выбрав ту модель, для которой средняя ошибка будет наименьшей. Каким будет k - решать вам.

Упражнение 2:
Пусть у нас есть данные x = [-3. , -2.68, -2.37, -2.05, -1.74, -1.42, -1.11, -0.79, -0.47, -0.16, 0.16, 0.47, 0.79, 1.11, 1.42, 1.74, 2.05, 2.37, 2.68, 3.]
y = [-3. , -2.68, -2.37, -2.05, -1.74, -1.42, -1.11, -0.79, -0.47, -0.16, 0.16, 0.47, 0.79, 1.11, 1.42, 1.74, 2.05, 2.37, 2.68, 3.]
z = [17.13, 15.82, 9.87, 5.78, 6.24, 3.4 , 4.62, 1.56, -3.94, 4.02, -0.05, 8.65, 11.67, 10.23, 12.41, 16.2 , 25.6 , 28.9 , 34.16, 41.51]
z зависит от x, y как полином некоторой степени. Установите наиболее подходящую степень многочлена для аппроксимации этих данных, найдите его коэффициенты. Подсказка: в случае, если это многочлен степени k в нем могут отсутствовать некоторые слагаемые со степенью меньше k. Решение: то же, что и в первой задаче, но уже для функции двух переменных. Возможно, значение коэффициентов при некоторых переменных будет близко к нулю. Это скорее всего означает, что данные переменные несущественны, коэффициент возник из-за шума в данных.

1.5 Линеаризация некоторых нелинейных задач

Здесь и далее предполагаем, что $\varepsilon \sim N(0, \sigma^2)$ - случайная нормально распределенная ошибка с постоянной дисперсией и нулевым математическим ожиданием, X, Y - переменные, α, β, a_1, a_2 - постоянные коэффициенты.

Пусть у нас имеется зависимость вида:

$$Z = a_1 X + a_2 Y^2 + \varepsilon$$

Здесь можно сделать замену переменной $Y' = Y^2$ и тем самым получить линейную

регрессию

$$Z = a_1X + a_2Y' + \varepsilon$$

В случае ниже можно произвести замену $\frac{1}{Y} = Y'$

$$Z = a_1X + a_2\frac{1}{Y} + \varepsilon = a_1X + a_2Y' + \varepsilon$$

Если один из признаков - это логарифм от переменной, то аналогично с помощью замены $Y' = ln(Y)$ получаем

$$Z = a_1X + a_2ln(Y) + \varepsilon = a_1X + a_2Y' + \varepsilon$$

Мультипликативную модель вида

$$Y = \alpha X^{\beta}\varepsilon'$$

где ε' - случайная величина, имеющая логнормальное распределение, такое, что $ln(\varepsilon') = \varepsilon$, можно линеаризовать, беря натуральный логарифм от левой и правой частей. Получим:

$$ln(Y) = ln(\alpha) + \beta ln(X) + \varepsilon$$

Если у нас имеется обратная экспоненциальная модель вида

$$Y = \frac{1}{1 + \alpha e^{\beta_1 X + \varepsilon}}$$

То ее тоже можно довольно просто линеаризовать. Для начала представим в виде

$$\frac{1}{Y} - 1 = \alpha e^{\beta_1 X + \varepsilon}$$

Затем возьмем логарифм от правой и левой частей. Получаем

$$ln(\frac{1}{Y} - 1) = ln(\alpha) + \beta_1 X + \varepsilon$$

Переобозначив переменные в левой и правой частях получим линейную регрессию.

Для мультипликативной и экспоненциальной модели в конце после получения ответа с помощью линеаризованной модели надо делать обратное преобразование для Y. Например, для экспоненциальной модели нужно взять экспоненту от полученного значения $ln(Y)$, т.к $e^{ln(Y)} = Y$

1.6 Общий случай линейной модели

Аппроксимация вида $Y = a_0 + a_1X + a_2X^2$ на самом деле тоже является линейной регрессионной модель. Возможно, вы возразите: ведь x^2 уже не линейная функция! Но если мы переобозначим X^2 как, например, Z, то получим уже классический случай линейной регрессии от двух переменных: $Y = a_0 + a_1X + a_2Z$. Фактически, если

недостаточно X и нужно ввести дополнительную переменную X^2, то мы просто переходим в другое пространство (X, X^2) и уже там пытаемся аппроксимировать Y линейной моделью.

В общем случае переменные при коэффициентах a_i не обязательно должны быть полиномами какой-либо степени. Это может быть, например, синус $sin(x)$, функция Гаусса $\frac{1}{2\pi k}e^{\frac{(x-\mu)^2}{k^2}}$, что угодно. Важно, чтобы эти элементарные функции были линейно независимы (подумайте, почему). То есть, мы можем вместо $Z = a_1X + a_2Y + b$ искать решение в виде $Z = a_0 + a_1X + a_2sin(x) + Y^2 + a_3\frac{1}{2\pi k}e^{\frac{(Y-\mu)}{k^2}}$, если это будет давать лучшую аппроксимацию. Как выбрать элементарные функции - общего рецепта нет, зависит от данных, от задачи. Как видите, это тоже можно считать линейной моделью, так как коэффициенты $a_1, a_2...a_n$ при элементарных функциях постоянны, а значения элементарной функции $f(x)$ в точке x можно переобозначить как x'.

Главное, чтобы функция аппроксимации зависела от неизвестных коэффициентов линейно. Например функция вида $y = a^x + b$, где **a** , **b** - неизвестные коэффициенты уже не является линейной функцией, т.к она зависит от коэффициента **a** нелинейно. Обобщая, можно сказать, что линейная модель аппроксимации - это такая модель, которая зависит от своих коэффициентов линейно. Для случая одной переменной она выглядит так:

$$y = \sum_{i=1}^{n} a_i g_i(x)$$

где a_i - постоянные коэффициенты.

1.7 Матричные производные.

В этом разделе будет дано очень краткое введение в матричные производные и их применение. В математике в связи с наличием векторных и матричных уравнений, неравенств есть необходимость в правилах вычисления производных таких выражений от скаляров, векторов, матриц. Все это можно назвать просто матричными производными, поскольку вектор-столбец и скаляр можно воспринимать соответственно как матрицу размерности (n, 1) и (1,10) соответственно. Существует 2 широкораспространенных соглашения правил вычисления таких производных: это так называемые Numerator-layout notation и Denominator-layout notation. Приведем краткий обзор этих соглашений. Здесь и далее будут использованы следующие обозначения: x, y - скаляры. **x**, **y** - векторы, **X**, **Y** - матрицы.

Numenator-layoyt:

$$\frac{\partial y}{\partial \mathbf{x}} = \left[\frac{\partial y}{\partial x_1}; \frac{\partial y}{\partial x_2}; ... \frac{\partial y}{\partial x_n}\right]$$

$$\frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \frac{\partial y_2}{\partial x} \\ \vdots \\ \frac{\partial y_m}{\partial x} \end{bmatrix}$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

$$\frac{\partial y}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \frac{\partial y}{\partial x_{21}} & \cdots & \frac{\partial y}{\partial x_{p1}} \\ \frac{\partial y}{\partial x_{12}} & \frac{\partial y}{\partial x_{22}} & \cdots & \frac{\partial y}{\partial x_{p2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial x_{1q}} & \frac{\partial y}{\partial x_{2q}} & \cdots & \frac{\partial y}{\partial x_{pq}} \end{bmatrix}$$

Denominator-layoyt.

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{bmatrix} \tag{1.20}$$

$$\frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x}; \frac{\partial y_2}{\partial x}; \ldots \frac{\partial y_m}{\partial x} \end{bmatrix}$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_n} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

$$\frac{\partial y}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \frac{\partial y}{\partial x_{12}} & \cdots & \frac{\partial y}{\partial x_{1q}} \\ \frac{\partial y}{\partial x_{21}} & \frac{\partial y}{\partial x_{22}} & \cdots & \frac{\partial y}{\partial x_{2q}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial x_{p1}} & \frac{\partial y}{\partial x_{p2}} & \cdots & \frac{\partial y}{\partial x_{pq}} \end{bmatrix}$$

Любое выражение, полученное с помощью Numenator-layoyt может быть легко приведено к выражению, полученного с помощью Denominator-layoyt с помощью транспонирования выражения, полученного применением Numenator-layoyt. В обратную сторону данный факт также справедлив. Но всегда стоит обозначать, какую именно нотацию вы используете для нахождения матричных производных. И ни в коем случае не смешивать Denominator-layoyt и Numenator-layoyt!

К примеру, взять производную по вектору от матричного уравнения в Numenator-layoyt, затем от получившегося выражения взять производную по скаляру в Denominator-layoyt, это будет грубой ошибкой. Сначала следует привести вычисления к соответствию с одним из соглашений. В данном примере следует после взятия производной по вектору в Numenator-layoyt сначала привести выражение тому, что получилось бы согласно Denominator-layoyt с помощью транспонирования выражения, и только затем брать производную по скаляру. В дальнейшем будет использоваться Denominator-layoyt.

Теперь выведем уравнение линейной регрессии с помощью матричных производных. Нам нужно минимизировать следующее выражение:

$$L = \|\hat{Y} - Y\|^2 = \|X\omega - Y\|^2 \tag{1.21}$$

Как вы могли заметить, здесь мы минимизируем квадрат невязки. Так надо, так как по теореме Маркова-Гаусса именно минимизация квадрата невязки а не абсолютного значения даст лучшую оценку параметров модели. Здесь $\hat{Y}$ - это аппроксимация, Y - значения. То есть нужно найти наиболее подходящий вектор коэффициентов ω, минимизирующий выражение:

$$argmin_\omega \|X\omega - Y\|^2$$

Пусть у нас n строк данных и m признаков. Тогда $X \in R^{n\times(m+1)}$ - матрица, $Y \in R^n$, - вектор-столбец, $\omega \in R^{m+1}$ - вектор столбец. $+1$ добавляется в связи с наличием свободного коэффициента. Несложно проверить, что $L \in R$.

Находить оптимальный вектор ω будем в пространстве, в котором определены скалярное произведение и норма. В нашем случае это просто конечномерное евклидово пространство. Причем норма определена через скалярное произведение. Таким образом, нам нужно минимизировать следующее выражение

$$argmin_\omega((X\omega - Y)^T \cdot (X\omega - Y)) \tag{1.22}$$

Прежде чем продолжить, следует выделить несколько свойств, которые нам пригодятся при выводе ω используя denominator-layout.

1. Транспонированное значение скаляра равно скаляру. Свойство очевидно следует из свойств чисел $\in R$.

2. $\mathbf{X}'_y = 0$ если $\mathbf{X}$ не зависит от y. Доказать самостоятельно.

3.
$$\frac{\partial \mathbf{x}^T \mathbf{a}}{\partial \mathbf{x}} = \mathbf{a}$$

Здесь $\mathbf{x}^T$ - вектор-строка, а $\mathbf{a}$ - вектор-столбец. Произведение $\mathbf{x}^T\mathbf{a}$ равно скаляру $\Sigma_{i=1}^n a_i x_i$. Учитывая правило нахождения производной скаляра по вектору (1.20) получаем требуемое равенство.

4.
$$\frac{\partial \mathbf{x}^T \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}\mathbf{x}$$

Здесь $\mathbf{A}$ - симметричная матрица (это важно) $\in R^{n\times n}$. В таком случае данное выражение будет скаляром. Как и раннее, $\mathbf{x}$ - вектор-столбец $\in R^n$. Значение скаляра

$S = \Sigma_{j=1}^{n}\Sigma_{i=1}^{n} a_{ij}x_i x_j$. Без ограничения общности возьмем производную от S например по x_1. Легко проверить, что $S'_{x_1} = \Sigma_{i=1}^{n} a_{1j}x_j + \Sigma_{j=1}^{n} a_{i1}x_i$. Поскольку в силу симметрии $a_{ij} = a_{ji}$ пользуясь правилом нахождения скаляра по вектору (1.20) получаем требуемое равенство.

Теперь перейдем к решению задачи (1.22). Получаем

$$\begin{gathered} L = (X\omega - Y)^T \cdot (X\omega - Y) \\ = (\omega^T X^T - Y^T) \cdot (X\omega - Y) \\ = (\omega^T X^T X\omega - \omega^T X^T Y - Y^T X\omega + Y^T Y) \\ = (\omega^T X^T X\omega - \omega^T X^T Y - Y^T X\omega + Y^T Y) \\ = (\omega^T X^T X\omega - 2\omega^T X^T Y + Y^T Y) \\ = (\omega^T X^T X\omega - 2\omega^T X^T Y)Y^T Y \\ L'_\omega = 2X^T X\omega - 2X^T Y = 0 \\ => \omega = (X^T X)^{-1} X^T Y \end{gathered} \tag{1.23}$$

В данных выкладках использовано то, что так как $Y^T X\omega$ - скаляр, то $Y^T X\omega = (Y^T X\omega)^T = \omega^T X^T Y$ и тот факт, что выражение $Y^T Y$ не зависит от ω. Интересно, что формула получилась такой же, как и при выводе другим методом (1.13). Вообще это не единственный способ. Можно еще вывести с помощью метода наименьших квадратов. В случае функции одной переменной это просто. В случае функции произвольного количества переменных уже значительно сложнее.

Производные по векторам и матрицам играют важную роль в машинном обучении, например с их помощью удобно записывать выражения для алгоритма обратного распространения ошибки при обучении нейросети, известного как backpropagation error. Так же они нужны в регрессионном анализе (выше с их помощью было выведена формула коэффициентов линейной регрессии), статистике. Там с их помощью исследуют многомерные случайные величины, распределения. К примеру, нормальное распределение в трехмерном пространстве.

В следующих трех упражнениях будет использоваться denominator-layout.

Упражнение 1: докажите что для произвольной матрицы $\mathbf{A} \in R^{n\times n}$ и вектора-столбца $\mathbf{x} \in R^n$.:

$$\frac{\partial \mathbf{x}^T \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A}^T + \mathbf{A})\mathbf{x}$$

Решение:

Начало аналогично симметричной матрице $\mathbf{A}$. Без ограничения общности возьмем производную по x_1 от выражения $S = \Sigma_{j=1}^{n}\Sigma_{i=1}^{n} a_{ij}x_i x_j$. Получим $S'_{x_1} = \Sigma_{i=1}^{n} a_{1j}x_j + \Sigma_{j=1}^{n} a_{i1}x_i$. Легко увидеть в первом слагаемом произведение первой строки на $\mathbf{x}$, а во втором произведение первого столбца на $\mathbf{x}$. Собирая все n строк вместе, получим требуемое утверждение.

Упражнение 2: докажите что для произвольного вектора-столбца $\mathbf{x} \in R^n$:

$$\frac{\partial \mathbf{x}^T \mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}$$

Решение: Составим выражение $S = \sum_{i=1}^{n} x_i x_i$. По правилу производной скаляра от вектора в denominator-layout здесь должен получиться вектор-столбец. Для i-й координаты получаем $\frac{\partial S}{\partial x_i} = 2x_i$, тем самым получая ответ.

Упражнение 3: Проблему в случае с сильной корреляцией признаков можно отчасти решить с помощью регуляризации. В таком случае функцию потерь L представляют в виде

$$L = \|X\omega - Y\|^2 + \lambda\|\omega\|^2$$

Здесь λ - это положительное вещественное число. Это так называемая регуляризация Тихонова. Возьмите производную от данного выражения по ω и тем самым выведите формулу нахождения оптимальных коэффициентов ω при регуляризации.

Решение: производную $\lambda\|\omega\|^2$ по ω можно представить в виде $2\lambda E\omega$. Очевидно, что это эквивалентно $2\lambda\omega$. Сложим уже известную производную для линейной регрессии без регуляризации с $\lambda 2E\omega$ и приравняем нулю. Получим

$$2X^T X\omega - 2X^T Y + 2\lambda E\omega = 0$$
$$=> (X^T X + \lambda E)\omega = X^T Y$$
$$=> \omega = (X^T X + \lambda E)^{-1} X^T Y$$

Упражнение 4: Иногда бывает так, что данных очень много (например 100ГБ), все данные не помещаются в оперативную память компьютера, поэтому воспользоваться явной формулой нахождения коэффициентов линейной регрессии не получится. В таком случае нужно извлекать вектора данных по одному или небольшими подмножествами (batch), находить производные по параметрам и изменять значения параметров в направлении градиента по параметрам от функции потерь.

Пусть дана система линейных уравнений $\mathbf{y} = \mathbf{X}\mathbf{a} + \mathbf{b}$, где $\mathbf{a}$ - вектор-столбец коэффициентов линейной регрессии (без свободного коэффициента), $\mathbf{a} \in R^k$, $\mathbf{X} \in R^{n,k}$, $\mathbf{b}$ - вектор-столбец $\in R^n$, все компоненты которого постоянны и равны некоторому числу $b \in R$, $\mathbf{y} \in R^n$ - значения из экспериментов. Обозначим как $\hat{\mathbf{y}} = \mathbf{X}\mathbf{a} + \mathbf{b}$ предсказания линейной модели, $L = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y})^2$ - функция потерь, среднее разности в квадрате между предсказанными и эмпирическими значениями $\mathbf{y}$. Найдите производные от L по $\mathbf{a}$ и по b. Параметры модели при таком методе будут изменяться итеративно в направлении уменьшения значения функции потерь: $\mathbf{a} = \mathbf{a} - \lambda\frac{dL}{d\mathbf{a}}$, $b = b - \lambda\frac{dL}{db}$. Здесь λ - положительная вещественная константа, размер шага прироста коэффициентов. Если все требования к линейной регрессии выполнены то при таком методе коэффициенты обязательно сойдутся к точке, где функция потерь минимальна.

Указание: вам понадобится производная скаляра по вектору и правило цепочки при дифференцировании.

Решение: Пользуясь правилом цепочки:

$$\frac{dL}{d\mathbf{a}} = \frac{dL}{d\hat{\mathbf{y}}}\frac{d\hat{\mathbf{y}}}{d\mathbf{a}}$$

Теперь вспомним, что производная скаляра по вектору - вектор-столбец, вектора по вектору - матрица. Несложно доказать, что $\frac{d\mathbf{Ax}}{d\mathbf{x}} = A^T$.

Производная $\frac{dL}{d\mathbf{a}} = \frac{1}{n}(\hat{\mathbf{y}} - \mathbf{y})$ - вектор-столбец. Собирая все вместе, получим

$$\frac{dL}{d\mathbf{a}} = \frac{1}{n} X^T (\hat{\mathbf{y}} - \mathbf{y})$$

В конце важно проверять, чтобы размерности были согласованы, т.е если $\mathbf{a}$ - вектор-столбец $\in R^{k,1}$, то производная скаляра по этому вектору так же должна быть вектором-столбцом $\in R^{k,1}$. Производная $\frac{dL}{db}$ будет равна:

$$\frac{dL}{db} = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)$$

Полезные ссылки:

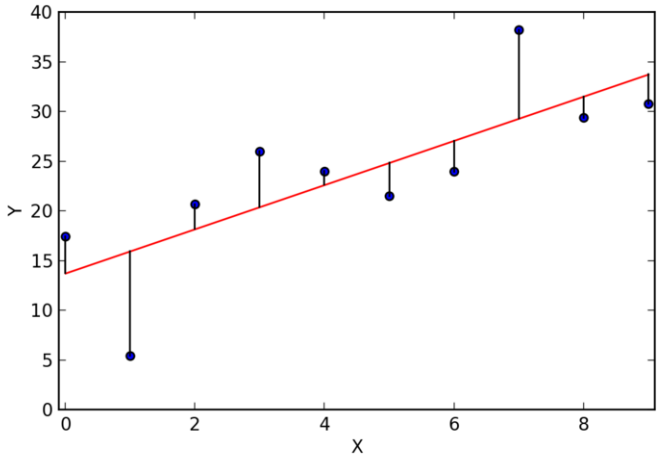
- Методичка для нахождения производных по векторам и матрицам: `https://en.wikipedia.org/wiki/Matrix_calculus`
- Краткий обзор основ линейной алгебры и матричных производных: `https://jonathan-hui.medium.com/machine-learning-linear-algebra-a5b1658f0151`

1.8 Метод наименьших квадратов.

В этой и особенно следующей главах очень желательно обладать базовыми знаниями теории вероятностей и математической статистики.

Метод наименьших квадратов, сокращенно МНК - часто используемый метод для нахождения параметров неизвестной функции при известных значениях в некоторых точках. На самом деле с помощью данного метода мы вывели уравнение линейной регрессии в прошлой главе, но там мы использовали матричные производные. Здесь же будет дан вывод беря производные поэлементно.

Как вы скоро увидите, в этом тоже есть смысл.



На рисунке выше проиллюстрирован смысл метода наименьших квадратов. Нужно найти такую функцию, что сумма значений расстояний (здесь это черные отрезки

между синими точками и красной прямой) между значениями искомой функции $f(x)$ и данными y в точках x была наименьшей. Вернее, ищут сумму квадратов расстояний, т.к доказано, что именно они, а не абсолютные значения определяют оптимальную кривую.

Выведем с помощью этого метода уравнение коэффициентов линейной регрессии. Начнем с одномерного случая. Пусть у нас даны пары значений $x_i, y_i, i \in [1 : n]$ и есть основания полагать, что лучше всего аппроксимирует зависимость x от y прямая $y = a_0 + a_1 x$. Тогда для произвольного i-го наблюдения выполняется:

$$y_i = a_0 + a_1 x_i + \varepsilon_i$$

Здесь ε_i - случайные величины, ошибки, которые должны удовлетворять определенным свойствам, чтобы линейная регрессия корректно аппроксимировала данные. Пока что будет считать, что удовлетворяет, подробно о них будет рассказано позже.

Запишем функцию потерь L от переменных (a_0, a_1) как:

$$L(a_0, a_1) = \sum_{i=1}^{n} (y_i - (a_0 + a_1 x_i))^2$$

Далее нужно взять производную от этой величины по неизвестным параметрам (a_0, a_1) и приравнять к нулю. Получим систему:

$$\begin{array}{l} \frac{\partial L}{\partial a_0} = -2 \sum_{i=1}^{n} (y_i - (a_0 + a_1 x_i)) = 0 \\ \frac{\partial L}{\partial a_1} = -2 \sum_{i=1}^{n} x_i (y_i - (a_0 + a_1 x_i)) = 0 \end{array} \tag{1.24}$$

Перепишем систему в виде

$$\begin{array}{c} a_0 n + a_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{array}$$

Обозначим $\frac{\sum_{i=1}^{n} x_i}{n}$ как $\overline{x}$, $\frac{\sum_{i=1}^{n} y_i}{n}$ как $\overline{y}$, $\frac{\sum_{i=1}^{n} x_i^2}{n}$ как $\overline{x^2}$, $\frac{\sum_{i=1}^{n} x_i y_i}{n}$ как $\overline{xy}$. Перепишем систему выше в виде

$$\begin{array}{c} a_0 n + a_1 n \overline{x} = n \overline{y} \\ a_0 n \overline{x} + a_1 \overline{x^2} = n \overline{xy} \end{array}$$

В этой системе неизвестны только (a_0, a_1). Ее легко можно решить например методом Крамера. Получим ответ:

$$\begin{array}{c} a_0 = \overline{y} = a_1 \overline{x} \\ a_1 = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} \end{array}$$

Таким образом, мы получили формулу для нахождения коэффициентов линейной функции одной переменной. Обсудим "физический смысл"полученного выражения.

Видно, что коэффициент a_0 будет равен нулю, если данные центрированы, т.е если математические ожидания зависимой переменной y и независимой x равны нулю. Этот коэффициент выражает так называемое смещение (bias) данных.

Коэффициент a_1 представляет собой в сущности значение линейной корреляции между y и x, умноженное на $\frac{\sqrt{D(Y)}}{\sqrt{DX}}$. Действительно, знаменатель равен значению дисперсии x, так как $D(X) = E(X^2) - E(X)^2$, а дисперсия, как известно, неотрицательна. Выражение в числителе - это ковариация переменных x и y, $cov(x, y) = E((X - E(X))(Y - Y(Y)))$. Если вспомнить что $corr(x, y) = \frac{cov(x,y)}{D(X)D(Y)}$, получим требуемое утверждение. Интересно, что если y и x независимы, то этот коэффициент равен нулю, так как значение ковариации в таком случае будет равно нулю.

Можно доказать, что перечисленные свойства справедливы и для функции многих переменных.

Известно, что равенство нулю первых производных необходимое, но недостаточное условие экстремума. Теперь докажем, что вектор (a_0, a_1) действительно доставляет минимальное значение функции L. Для этого нужно найти матрицу вторых производных функции L по (a_0, a_1), так называемый Гессиан, и доказать, что его детерминант строго больше нуля.

Напоминаем, что для достижения дважды дифференцируемой функции **k** переменных экстремума в точке $(x_0, x_1, ...x_k)$ необходимо и достаточно:

1. Равенство нулю всех частных производных функции в данной точке.
2. Детерминант Гессиана в точке должен быть строго больше нуля (для минимума) либо строго меньше нуля (для максимума)

Это можно доказать, используя формулу Тейлора для произвольного числа переменных. Вернемся к нашей задаче. Требуется найти следующую матрицу:

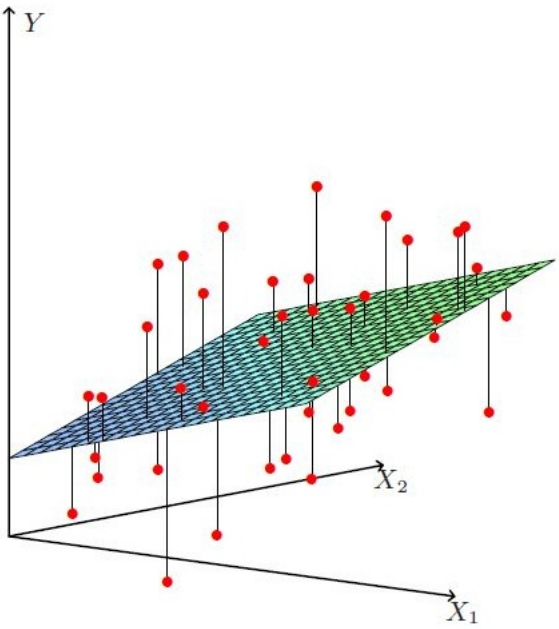
$$H(L) = \begin{pmatrix} \frac{\partial^2 L}{\partial a_0^2} & \frac{\partial^2 L}{\partial a_0 \partial a_1} \\ \frac{\partial^2 L}{\partial a_0 \partial a_1} & \frac{\partial^2 L}{\partial a_1^2} \end{pmatrix}$$

Для этого воспользуемся первыми производными, найденными в выражении (1.24). Получим:

$$H(L) = \begin{pmatrix} 2n & 2n\overline{x} \\ 2n\overline{x} & 2n\overline{x^2} \end{pmatrix}$$

Детерминант этой матрицы будет равен $det(H(L)) = 4n^2(\overline{x^2} - \overline{x}^2) = 4n^2 D(X)$. Полагая, что массив X - не константа, и зная, что дисперсия всегда неотрицательна получаем требуемое утверждение.

Для трехмерной задачи вы наверное догадываетесь, что вместо прямой будет уже плоскость. Остальное то же самое.



Теперь перейдем к выводу формулы линейной регрессии с помощью МНК для произвольного числа переменных. Предположим у нас **k** переменных. В этот раз нужно найти функцию с параметрами $(a_0, a_1, ...a_k)$ удовлетворяющую условиям:

$$y_i = a_0 + \sum_{j=1}^{k} x_{ij} a_j + \varepsilon_i$$

ε_i - случайные ошибки, который должны обладать теми же свойствами, что и для функции одной переменной.

Теперь функция потерь L будет функцией от **k+1** переменных: $L(a_0, a_1, ...a_k)$.

$$L(a_0, a_1) = \sum_{i=1}^{n} (y_i - (a_0 + \sum_{j=1}^{k} x_{ij} a_j))^2$$

Беря от L производные по параметрам $(a_0, a_1, ...a_k)$ получаем систему:

$$\begin{gathered}
\frac{\partial L}{\partial a_0} = -2 \sum_{i=1}^{n} (y_i - (a_0 + \sum_{j=1}^{k} x_{ij} a_j)) = 0 \\
\frac{\partial L}{\partial a_1} = -2 \sum_{i=1}^{n} x_{i1} (y_i - (a_0 + \sum_{j=1}^{k} x_{ij} a_j)) = 0 \\
\frac{\partial L}{\partial a_2} = -2 \sum_{i=1}^{n} x_{i2} (y_i - (a_0 + \sum_{j=1}^{k} x_{ij} a_j)) = 0 \\
... \\
\frac{\partial L}{\partial a_k} = -2 \sum_{i=1}^{n} x_{ik} (y_i - (a_0 + \sum_{j=1}^{k} x_{ij} a_j)) = 0
\end{gathered}$$

Перепишем в виде

$$\begin{gathered}
a_0 n + \sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij} a_j = \sum_{i=1}^{n} y_i \\
a_0 \sum_{i=1}^{n} x_{i1} + \sum_{i=1}^{n} x_{i1} \sum_{j=1}^{k} x_{ij} a_j = \sum_{i=1}^{n} x_{i1} y_i \\
a_0 \sum_{i=1}^{n} x_{i2} + \sum_{i=1}^{n} x_{i2} \sum_{j=1}^{k} x_{ij} a_j = \sum_{i=1}^{n} x_{i2} y_i \\
a_0 \sum_{i=1}^{n} x_{i3} + \sum_{i=1}^{n} x_{i3} \sum_{j=1}^{k} x_{ij} a_j = \sum_{i=1}^{n} x_{i3} y_i \\
... \\
a_0 \sum_{i=1}^{n} x_{ik} + \sum_{i=1}^{n} x_{ik} \sum_{j=1}^{k} x_{ij} a_j = \sum_{i=1}^{n} x_{ik} y_i
\end{gathered} \tag{1.25}$$

Пусть у на $X \in R^{n,k+1}$ - матрица данных (первый столбец состоит из единиц), $Y \in R^n$ - вектор-столбец значений, $a \in R^{jk+1}$ - вектор-столбец коэффициентов. Тогда можно увидеть, что правую часть системы (1.25) можно записать в виде X^TY.

В левой части выражение $\sum_{i=1}^{n}(a_0 + \sum_{j=1}^{k} x_{ij}a_j)$ можно записать в виде Xa, $Xa \in R^n$, после чего можно увидеть, что произведение x_{ij} на $\sum_{i=1}^{n}(a_0 + \sum_{j=1}^{k} x_{ij}a_j)$ представляет из себя X^TXa. В итоге получаем уравнение

$$X^TXa = X^TY$$

из которого получаем ответ

$$a = (X^TX)^{-1}X^TY$$

Здесь в отличии от способа, где мы использовали матричные производные во многом приходилось опираться на интуицию и догадки по поводу того, как можно представить получившееся выражение в виде произведения матриц, в чем и состоит один из недостатков этого метода, когда мы считаем производные не по матрицам а по отдельным скалярам. К тому же эти вычисления довольно громоздкие. Гессиан в этой точке тоже отрицателен, что так же доказано. Здесь доказательство приводится не будет.

Вообще метод наименьших квадратов подходит не только для линейных уравнений, но и для многих других типов функций, например для зависимости вида

$$y = e^{\mathbf{a}x}$$

если даны пары значений x_i, y_i. Здесь нужно найти неизвестный параметр-скаляр $\mathbf{a}$. Определим функцию потерь L

$$L = \sum_{i=1}^{n}(y_i - e^{\mathbf{a}x_i})^2$$

Здесь так же нужно взять производную от функции потерь L по параметрам, найти стационарную точку и доказать, что в ней Гессиан отрицателен.

1.9 Теоретические аспекты линейной регрессии

1.9.1 Постановка задачи

Линейная регрессия очень хорошо изучена с точки зрения теории. В данном разделе обсудим лишь часть того, что известно, в частности особенности применения линейной регрессии, теоретические аспекты, связанные с ней, а так же саму линейную регрессию с точки зрения теории вероятностей.

Пусть у нас имеется вектор-столбец значений искомой функции $Y \in R^n$, вектор параметров линейной регрессии $a \in R^{k+1}$, матрица данных $X \in R^{n,k+1}$, вектор столбец ошибок $\varepsilon \in R^n$. Все это можно записать в матричном виде

$$Y = Xa + \varepsilon$$

Требуется найти вектор a, который "максимально правдоподобно"удовлетворяет этой системе.

Всего у нас наблюдений **n**. Для каждого i-го наблюдения можно записать следующее выражение:

$$y_i = a_0 + \sum_{j=1}^{k} a_i X_{ij} + \varepsilon_i$$

Здесь ε_i - случайные величины, ошибки, которые должны удовлетворять следующим свойствам (иначе корректной аппроксимации не будет):

1. $E(\varepsilon_i) = 0$, равенство нулю математических ожиданий ожиданий.
2. $D(\varepsilon_i) = \sigma^2$, равенство дисперсии - гомоскедастичность.
3. $\varepsilon_i \sim N(0, \sigma^2)$, нормальное распределение.
4. $Cov(\varepsilon_j, \varepsilon_i) = 0, i \neq j$, равенство нулю ковариации, т.е линейная независимость.
5. Отсутствие выбросов

Выбросы - это аномально большие значения, сильно отклоняющиеся от нормы. К примеру, во множестве [1, 5, 2, 100500, 3] вы наверное догадываетесь, какое число является выбросом.

Напомним еще одно определение. Пусть у нас имеется оценка $\hat{a}$ параметра a некой случайной величины. Эта оценка называется несмещенной, если

$$E[a] = \hat{a}$$

Это определение нам важно, так как мы ищем именно несмещенные оценки параметров линейной регрессии. Саму модель линейной регрессии можно рассматривать как случайную величину, так как при разных данных X ее параметры a будут немного другими, даже если выполнены все требования к данным и ошибкам. Это обусловлено тем, что мы не можем использовать всевозможные данные во вселенной об исследуемой модели, а лишь только некоторое их подмножество X.

Оптимальной оценкой называется такая, что ее дисперсия минимальна, т.е если a_1 - оптимальная оценка параметра a, то при $\forall a_2$ $E[(a - a_1)^2] <= E[(a - a_2)^2]$

Оценка, которая является одновременно оптимальной и несмещенной называется эффективной.

1.9.2 Вывод уравнения

Теорема Маркова-Гаусса:

Рассматривается модель линейной регрессии k переменных в которой Y_i связаны с переменными X_{ij} зависимостью следующего вида: $Y_i = \sum_{j=1}^{k} X_{ij} a_j + a_0 + \varepsilon_i$. На основе n выбранных наблюдений находятся оценочные параметры уравнение регрессии $\hat{Y}_i = \sum_{j=1}^{k} X_{ij} \hat{a_j} + \hat{a_0}$.

Если данные обладают следующими свойствами:

1. Модель данных правильно специфицировала.
2. $\forall j$ все X+X_{ij} детерминированы и не равны между собой
3. $E(\varepsilon_i) = 0$, равенство нулю математических ожиданий ожиданий, т.е нет систематичности в ошибках.
4. $D(\varepsilon_i) = \sigma^2$, равенство дисперсии - гомоскедастичность.
5. $\varepsilon_i \sim N(0, \sigma^2)$, нормальное распределение ошибок.
6. $Cov(\varepsilon_j, \varepsilon_i) = 0, i \neq j$, равенство нулю ковариации, т.е линейная независимость случайных ошибок.

то в этих условиях МНК дает оптимальные несмещенные оценки линейной регрессии.

Здесь первое условие означает, что переменные линейно независимы (по крайней мере корреляция между ними не слишком большая), что имеется свободный коэффициент смещения a_0 и все необходимые признаки X_j для аппроксимации Y. К примеру, может потребоваться введение квадрата какой-либо переменной для корректной аппроксимации, что мы рассматривали при изучении полиномиальной регрессии.

Второе условие означает, что данные не должны быть константами.

Если все эти условия выполнены, то теорема гласит, что оптимальные коэффициенты $\hat{a}$ определяются с помощью выражения

$$\hat{a} = argmin_a |Y - Xa|^2$$

Мы уже несколько раз до этого выводили формулу нахождения $\hat{a}$. Напомним, что:

$$\hat{a} = (X^T X)^{-1} X^T Y$$

1.9.3 Метод максимального правдоподобия

Метод максимального правдоподобия - популярный метод нахождения неизвестных параметров случайной величины. Напомним его суть. Пусть у нас имеется случайная величина Y и параметры θ. Y может принимать различные значения. Допустим, у нас имеется n наблюдений, в каждом из которых случайная величина приняла определенное значение Y_i при условии, что имеются параметры a и данные X_i. Требуется найти параметры $\hat{a}$ таким образом, чтобы минимизировать следующее выражение

$$P(Y|X, \hat{a}) = argmin_a(\prod_{i=1}^{n} P(Y_i|X_i, a))$$

Правда, обычно минимизируют не $P(Y|X, a)$, а логарифм этого выражения. Это связано с тем, что произведение большого количества чисел $p : 0 <= p =< 1$ может быстро стать настолько малым, что компьютер попросту не сможет поддерживать количество знаков после запятой, чтобы хранить такое значение. Так как логарифм - функция монотонная и строго возрастающая, то $ln(f(x))$ не изменяет значение точек x, в которых

функция $f(x)$ достигает своих экстремальных значений, благодаря чему достаточно найти максимум логарифма данного выражения с помощью производных по параметрам.

Рассмотрим с его помощью взгляд на уравнение линейной регрессии с точки зрения теории вероятностей. Модель та же

$$Y = Xa + \varepsilon$$

и предположение относительно ошибок ε то же самое

$$\varepsilon \sim N(0, \sigma^2)$$

Тогда можно записать вероятность того, что Y примет значение y_i при имеющихся данных X_i

$$p(y_i|X_i, a) = \sum_{j=1}^{k} X_{ij}a_j + a_0 + N(0, \sigma^2) = N(\sum_{j=1}^{k} X_{ij}a_j + a_0, \sigma^2)$$

Согласно теореме Маркова-Гаусса наблюдения независимы, т.к ошибки ε_i не коррелированы. Тогда можно записать логарифм функции правдоподобия для значений y_i

$$\begin{aligned} ln(p(Y|X, a)) &= ln(\textstyle\prod_{i=1}^{n} N(\sum_{j=1}^{k} X_{ij}a_j + a_0, \sigma^2)) \\ &= \textstyle\sum_{i=1}^{n} ln(N(\sum_{j=1}^{k} X_{ij}a_j + a_0, \sigma^2)) \\ &= -\frac{n}{2}ln(2\pi\sigma^2) - \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - X_i a) \end{aligned}$$

Выражение для оценки вектора неизвестных коэффициентов a можно записать в виде

$$\hat{a} = argmax_a ln(p(Y|X, a))$$

В данном выражении можно отбросить все, что не зависит от a. Получаем:

$$\begin{aligned} \hat{a} &= argmax_a ln(p(Y|X, a)) \\ &= argmax_a(-\tfrac{1}{\sigma^2}\textstyle\sum_{i=1}^{n}(y_i - X_i a)) \\ &= argmax_a(-\tfrac{1}{\sigma^2}|Y - Xa|^2) \end{aligned}$$

Если взять производную по параметрам от логарифма правдоподобия, получим:

$$\begin{aligned} \tfrac{dln(p(Y|X,a))}{da} &= -\tfrac{1}{\sigma^2}\tfrac{d|Y - X\vec{a}|^2}{da} = 0 \\ <=> \tfrac{d|Y - X\vec{a}|^2}{da} &= 0 \end{aligned}$$

Следовательно, максимизация правдоподобия данных при имеющихся наблюдениях это то же самое, что минимизация ошибки аппроксимации методом МНК. Это доказывает, что то, что лучшая функция потерь L для нахождения оптимальных коэффициентов a линейной регрессии является квадратом невязки - это следствие того, что ошибки распределены нормально и обладают всеми перечисленными раннее свойствами, а не наоборот.

1.9.4 Вероятностная природа коэффициентов линейной регрессии и предсказаний модели

Как уже было отмечено до этого, мы не обучаем модель линейной регрессии на всевозможных в природе данных. Мы используем для этого лишь некоторое доступное подмножество этих данных. И в зависимости от того, какие данные нам попадутся коэффициенты линейной регрессии a_i могут незначительно отличаться при условии выполнения всех требований к линейной регрессии. Поскольку ответ линейной регрессии $\hat{y(x)}$ на входные данные x зависит от ее параметров a_i, представляющих собой случайные величины, то логично предположить, что и ответ $\hat{y(x)}$ - это некоторая случайная величина, имеющая матожидание и дисперсию.

Рассмотрим эти вопросы. Начнем со случая Линейной регрессии одной переменной.

Предварительно введем некоторые определения. Пусть у нас имеется n наблюдений, y_i - имеющиеся эмпирические значения, $\hat{y_i}$ - предсказанные моделью.

$$RSS = \sum_{i=1}^{n} (\varepsilon_i)^2 = \sum_{i=1}^{n} (y_i(x) - \hat{y_i(x)})^2$$

Это сумма остатков - разностей между опытными и предсказанными моделью значениями, иначе говоря сумма ошибок. RSS - Residual Sum of Squares.

$$S^2 = \frac{RSS}{n-2}$$

Это несмещенная оценка дисперсии ошибок наблюдений ε

Если $ES^2 = \sigma^2$, то σ^2 - истинная дисперсия ошибок предсказаний. В таком случае, случайная величина

$$Z_S = \frac{S^2(n-2)}{\sigma^2} \sim \chi(n-2)$$

имеет распределение хи-квадрат с $n-2$ степенями свободы. Эта СВ нужна для определения интервала, в котором может находиться истинное значение σ^2

Допустим, что выполнены все требования теоремы Маркова-Гаусса. У нас имеется модель линейной регрессии

$$\hat{y(x)} = \hat{a_0} + \hat{a_1}x$$

Тогда коэффициенты простой линейной регрессии обладают следующими свойствами:

1. $E[\hat{a_0}] = a_0$ - несмещенность свободного коэффициента
2. $E[\hat{a_1}] = a_1$ - несмещенность коэффициента при x.
3. $D[\hat{a_0}] = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2/n}{nD_x^*}$
4. $D[\hat{a_1}] = \frac{\sigma^2}{nD_x^*}$

5. $\hat{a_0} \sim N(a_0, \sqrt{D[\hat{a_0}]})$
6. $\hat{a_1} \sim N(a_1, \sqrt{D[\hat{a_1}]})$

Здесь D_x^* - выборочная несмещенная оценка дисперсии x, σ^2 - истинная дисперсия ошибок наблюдений. Распределения неизвестных параметров предполагаются нормально распределенными, так как ошибки распределены нормально.

Тогда следующая статистика имеет распределение Стьюдента с $n-2$ степенями свободы. Замечание: при больших значениях n распределение Стьюдента можно считать нормальным. Обычно его считают нормальным при числе степеней свободы больше 30, т.е в данном случае если $n-2 > 30$. Значение $n-2$ обусловлено тем, что у нас имеется 2 коэффициента связи a_0 и a_1. Поскольку значение σ нам неизвестно, то мы используем выборочное среднеквадратичное отклонение ошибок S.

$$Z_{a_i} = \frac{\sigma}{S}\frac{\hat{a}_i - a_i}{\sqrt{D[\hat{a}_i]}} \sim T(n-2)$$

Используя значения Z_{a_i} и Z_S, мы можем построить доверительный интервал для параметров a_i и σ^2 с уровнем значимости α:

$$a_0 : \hat{a_0} \pm t_{1-\frac{\alpha}{2}}(n-2)S\sqrt{\frac{\sum_{i=1}^{n} x_i^2/n}{nD_x^*}} \tag{1.26}$$

$$a_1 : \hat{a_1} \pm t_{1-\frac{\alpha}{2}}(n-2)S\sqrt{\frac{1}{nD_x^*}} \tag{1.27}$$

$$\sigma^2 : (\frac{S^2(n-2)}{\chi^2_{1-\frac{\alpha}{2}}(n-2)}, \frac{S^2(n-2)}{\chi^2_{\frac{\alpha}{2}}(n-2)}) \tag{1.28}$$

Для определения того, значимы ли коэффициенты a_0 и a_1 нужно провести статистический тест, в котором нулевая гипотеза H_0 - значение коэффициента a_i равно нулю, т.е текущее значение a_i - случайно, реальной связи между y и x нет. Альтернативная гипотеза - H_1 - связь есть. α обычно выбирают 0.95.

К примеру, для коэффициента a_1 если значение 0 попадает в интервал (1.27), то можно принять нулевую гипотезу, что значение этого коэффициента - случайность, реальной связи между переменными y и x нет с уровнем значимости α. Вероятность ложно отвергнуть нулевую гипотезу в случае, если она на самом деле верна при этом равна $1-\alpha$.

Теперь вернемся к доверительному интервалу для $\hat{y(x)}$. Напомним, что $\hat{y(x)} = \hat{a_0} + \hat{a_1}x$ Поскольку при выполнении теоремы Гаусса-Маркова коэффициенты $\hat{a_0}$ и $\hat{a_1}$ - оптимальные и несмещенные, то справедливы следующие свойства:

1. $E[\hat{y(x)}] = E[a_0 + a_1]x = E[a_0] + E[a_1]x = \hat{a_0} + \hat{a_1}x$
2. $D[\hat{y(x)}] = \frac{\sigma^2}{n}(1 + \frac{(\overline{x}-x)^2}{D_x^*})$
3. $\hat{y(x)} \sim N(a_0 + a_1x, \sqrt{D[\hat{y(x)}]})$

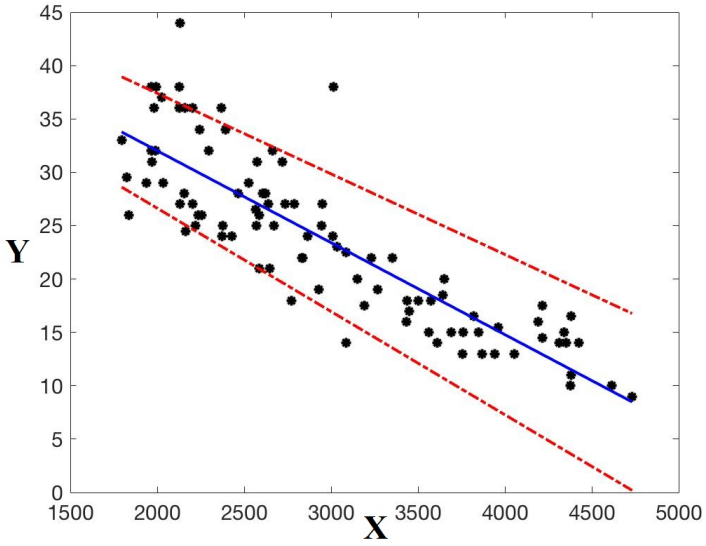
Тогда случайна величина Z_y имеет распределение Стьюдента с $n-2$ степенями свободы:

$$Z_y = \frac{\sigma}{S}\frac{\hat{y(x)} - (a_0 + a_1 x)}{\sqrt{D\hat{y(x)}}} \sim T(n-2)$$

В таком случае реальное значение $y(x)$ будет попадать в следующим доверительный интервал:

$$\hat{y(x)} \pm t_{1-\frac{\alpha}{2}}(n-2)S\sqrt{\frac{1 + \frac{(\overline{x}-x)^2}{D_x^*}}{n}}$$

В качестве примера рассмотрим следующую зависимость $y(x) = a_0 + a1x$



Это пример того, как может выглядеть доверительный интервал для значений. Черными точками обозначены наблюдения. Синяя линия - предсказания $\hat{y(x)}$ для x. Все, что между красными линиями - доверительный интервал с уровнем значимости α. Пусть $\alpha = 0.99$. Это означает, что истинное значение $y(x)$ с вероятностью 0.99 находится где-то между этими линиями при каждом конкретном значении x.

Теперь разберем случай множественной линейной регрессии. В целом тут все аналогично простой линейной регрессии, но с поправкой на то, что здесь вместо матожидания и дисперсии одного признака нужно находить вектор матожиданий и матрицу ковариации для вектора признаков соответственно.

Пусть имеется система уравнений

$$Y = Xa + \varepsilon$$

где a - неизвестный вектор истинных коэффициентов, ε - вектор ошибок наблю-

дений. X - матрица данных:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & ... & x_{1k} \\ 1 & x_{21} & x_{22} & ... & x_{2k} \\ ... & & & & \\ 1 & x_{n1} & x_{n2} & ... & x_{nk} \end{pmatrix}$$

Требования теорема Гаусса-Маркова выполнены. С помощью этой системы мы получили модель

$$\hat{y} = \hat{a_0} + \sum_{i=1}^{k} \hat{a_i} x_i$$

где $\hat{a_i}$ - оценки неизвестных параметров a_i. Покажем, что оценка $\hat{a}$ является несмещенной. Действительно, т.к $\hat{a} = (X^T X)^{-1} X^T Y$ и $Y = Xa + \varepsilon$, то $\hat{a} - a = (X^T X)^{-1} X^T (Xa + \varepsilon) - a = (X^T X)^{-1} X^T \varepsilon$. Тогда $E[(X^T X)^{-1} X^T \varepsilon] = \vec{0}$, так как матожидание ε равно нулю.

Теперь найдем матрицу ковариации, а вместе с ней и дисперсию оценок $\hat{a}$. $D[\hat{a}] = E[(\hat{a} - a)(\hat{a} - a)^T] = E[(X^T X)^{-1} X^T \varepsilon \varepsilon^T X (X^T X)^{-1}] = \sigma^2 (X^T X)^{-1}$

А сейчас запишем свойства оценок S, $\hat{a}$ и в матричном виде:

1. $E[\hat{a}] = a$ - несмещенность коэффициентов
2. $D[\hat{a}] = \sigma^2 (X^T X)^{-1}$
3. $\varepsilon \sim N(\vec{0}, \sigma^2 I_n)$
4. $Z_S = \frac{S^2(n-2)}{\sigma^2} \sim \chi^2(n-k-1)$
5. $\hat{a} \sim N(0, \sigma^2 (X^T X)^{-1})$
6. $Z_{a_i} = \frac{\hat{a_i} - a_i}{S\sqrt{[(X^T X)^{-1}]_{i,i}}} \sim T(n-k-1)$

Здесь k - количество признаков, от которых зависит линейная регрессия. Доверительный интервал для оценок σ^2, $\hat{a}$ по аналогии будет выглядеть следующим образом:

$$a_i : \hat{a_i} \pm t_{1-\frac{\alpha}{2}}(n-k-1) S \sqrt{[(X^T X)^{-1}]_{i,i}}$$

$$\sigma^2 : \left(\frac{S^2(n-k-1)}{\chi^2_{1-\frac{\alpha}{2}}(n-k-1)}, \frac{S^2(n-k-1)}{\chi^2_{\frac{\alpha}{2}}(n-k-1)}\right)$$

Введем случайную величину для $y(x)$, имеющую распределение Стьюдента с $n - k - 1$ степенями свободы.

$$Z_y = \frac{\hat{y(x)} - y(x)}{S\sqrt{x^T (X^T X)^{-1} x}},$$

Здесь $x = (1, x_1, x_2, ..x_k)$ - вектор данных. Доверительный интервал с уровнем значимости α для этой величины будет определяться числами

$$\hat{y(x)} \pm t_{1-\frac{\alpha}{2}}(n-2) S \sqrt{x^T (X^T X)^{-1} x}$$

На этом заканчивается наше краткое изложение введения в теорию линейной регрессии.

1.9.5 Проблемы, встречающиеся при применении линейной регрессии

В этой главе обсудим, что бывает, когда нарушаются какие-либо предположения о характере распределения ошибок ε_i.

Пусть нам нужно аппроксимировать истинную функцию $y\_true = 2x + 3$. Ее истинный вид на не известен. Мы получили ее опытные значения y, искаженные случайной ошибкой - шумом ε. Всего наблюдений $n = 30$, y зависит всего от одной переменной x.

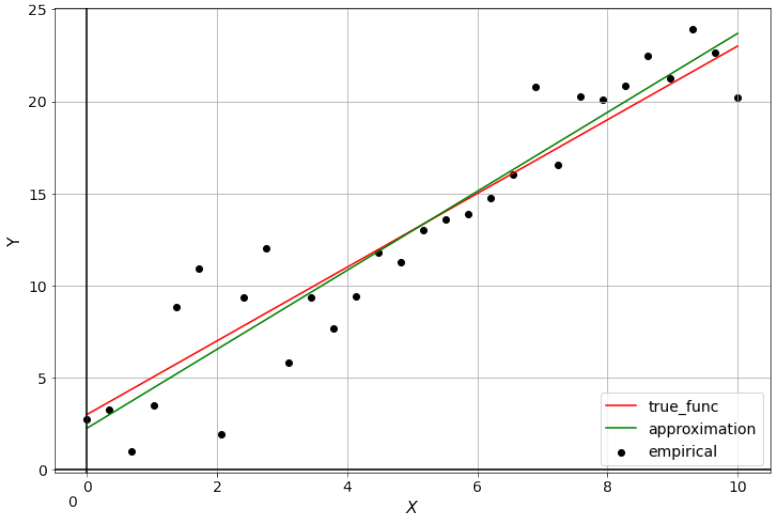
$$y = Xa + \varepsilon$$

На основе этих данных мы построили аппроксимацию неизвестной функции $\hat{y}$, а так же вычислили значение остатков $residuals = y - \hat{y}$. Из равенства $\varepsilon = y - Xa$ очевидно, что эти остатки должны обладать всеми свойствами случайной ошибки, перечисленными раннее:

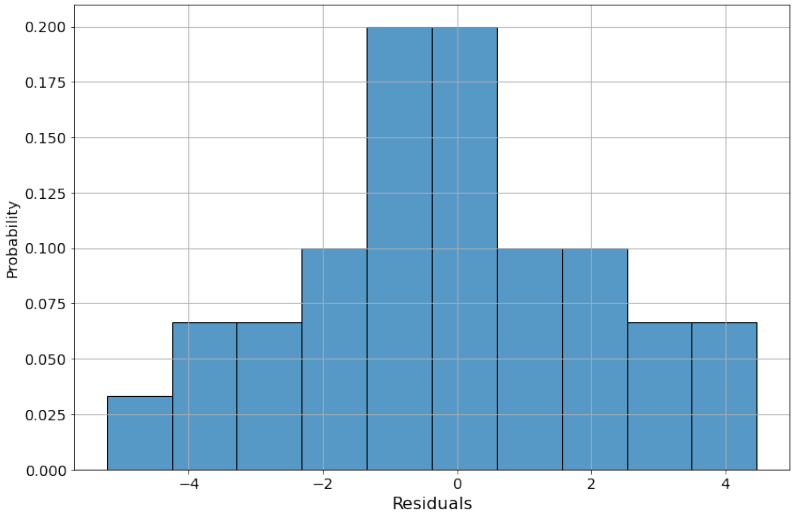
1. $E(\varepsilon_i) = 0$, равенство нулю математических ожиданий ожиданий.
2. $D(\varepsilon_i) = \sigma^2$, равенство дисперсии - гомоскедастичность.
3. $\varepsilon_i \sim N(0, \sigma^2)$, нормальное распределение.
4. $Cov(\varepsilon_j, \varepsilon_i) = 0, i \neq j$, равенство нулю ковариации, т.е линейная независимость.
5. Отсутствие выбросов

Посмотрим, что будет, если некоторые из этих свойств будут нарушены.

Начнем со случая, когда все требования к ошибкам выполнены, модель правильно специфицирована. Посмотрим, как все выглядит в случае верно подобранной модели. Ошибки имеют распределение $\varepsilon \sim N(0, 2^2)$.

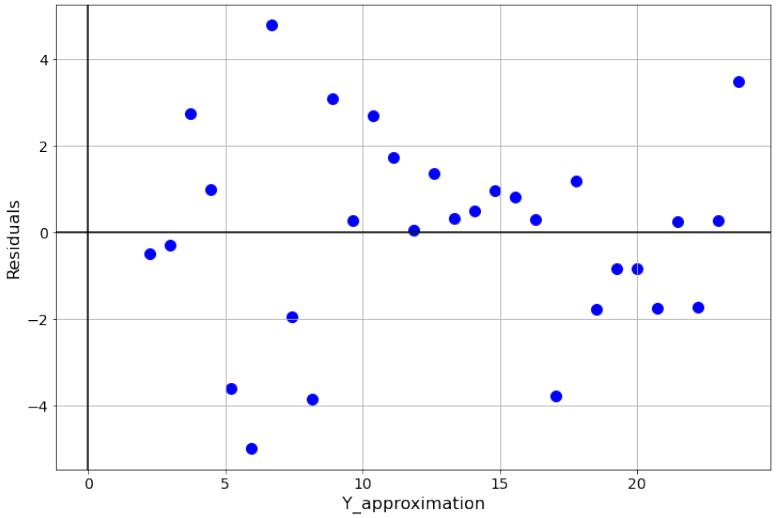


Мы видим, что аппроксимация почти полностью совпадает с истинной функцией. Посмотрим на распределение остатков.

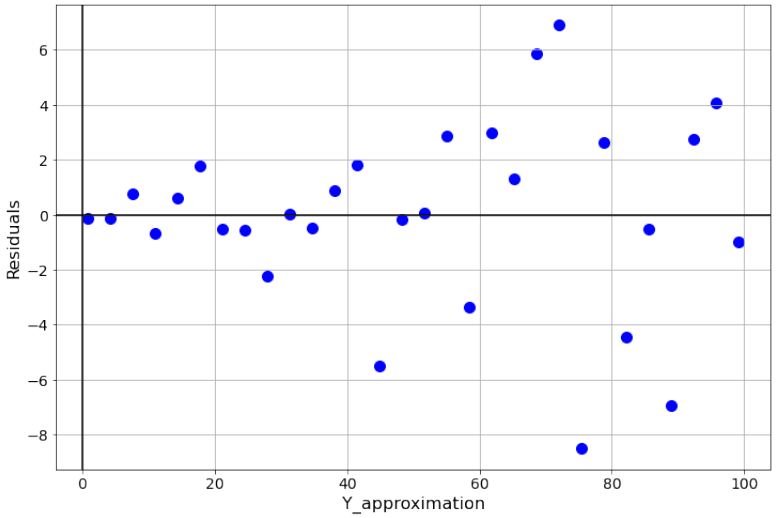


Видно, что оно похоже на нормальное. Чтобы в этом убедится, можно провести тест на нормальность, например Шапиро-Уилкоксона, либо Колмогорова-Смирнова. Для Шапиро-Уилкоксона в данном случае pvalue будет равно $pvalue \approx 0.86$, что дает основания принять нулевую гипотезу о нормальном распределении.

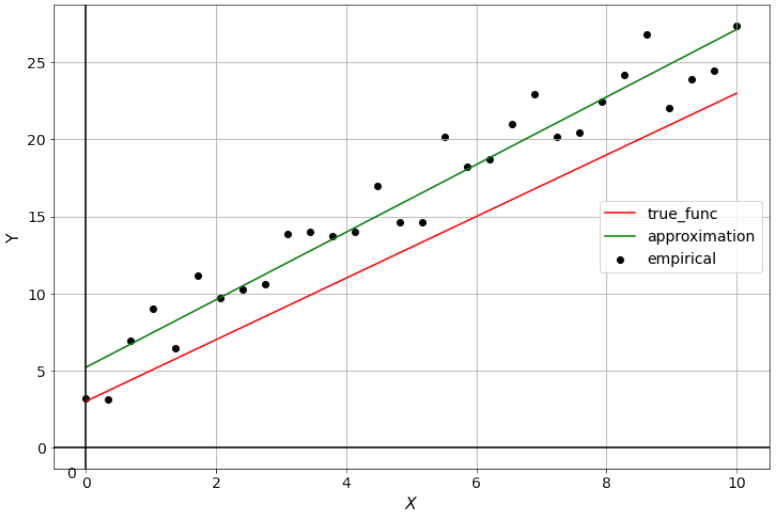
Так же может быть полезно взглянуть на график зависимости остатков от значений аппроксимации либо признаков x. По этому графику так же можно судить о выполнении требований к ошибкам.



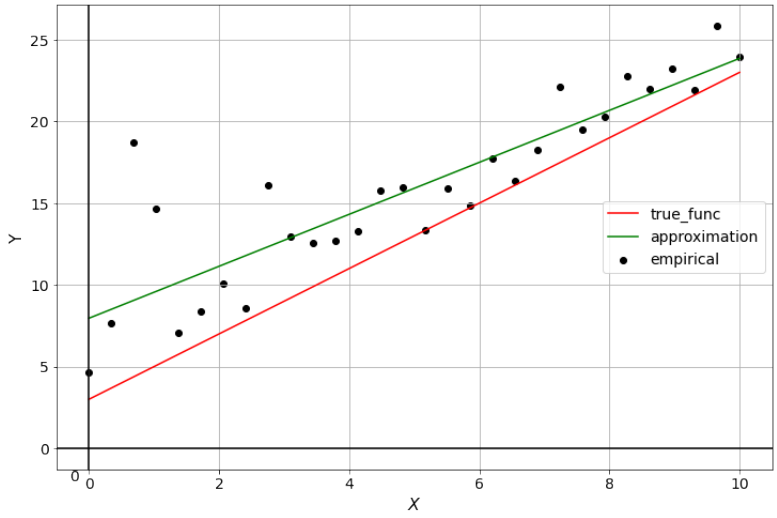
Теперь посмотрим, что будет в случае нарушения гомоскедастичности.



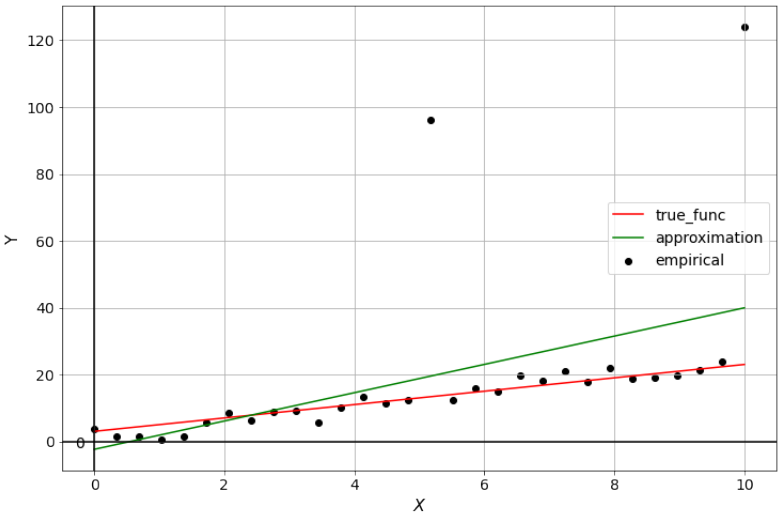
Если дисперсия ошибок непостоянна, то это может говорить о неправильно выбранной спецификации модели. Если же матожидание ошибок не равно нулю, то аппроксимировать мы будем скорее какую-то другую функцию, смещенную относительно истинной, в общем аппроксимировать не то, что надо.



В случае, если случайные ошибки распределены не нормально, а как-то по другому, то опять же аппроксимировать истинную функцию не выйдет. В примере ниже ошибки распределены экспоненциально с параметром $\lambda = 3$.



При наличии выбросов (outliers) параметры аппроксимированной функции могут быть сильно искажены, как видно на этом графике. Поэтому для линейной регрессии крайне важно перед аппроксимацией удалить все выбросы. Выбросы могут присутствовать по разным причинам. К примеру, если кто-то сделал опечатку и ввел на знак больше.



В случае неверной спецификации модели, т.е когда способ аппроксимации выбран неверно можно получить вот это

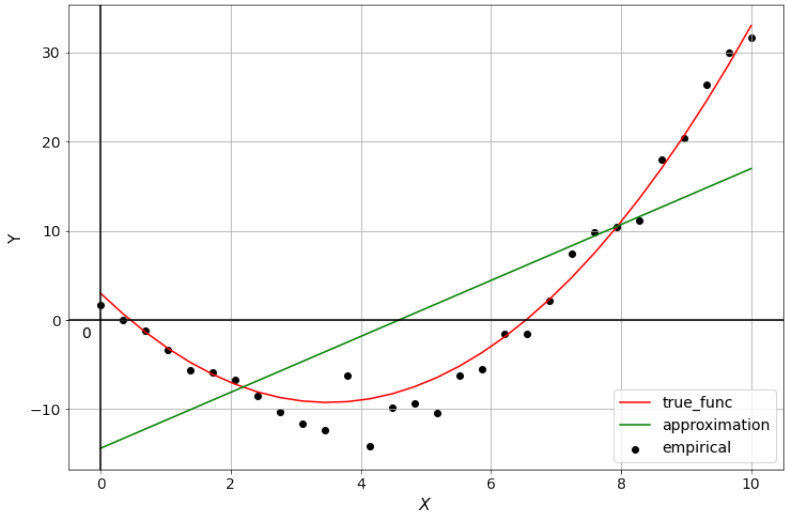
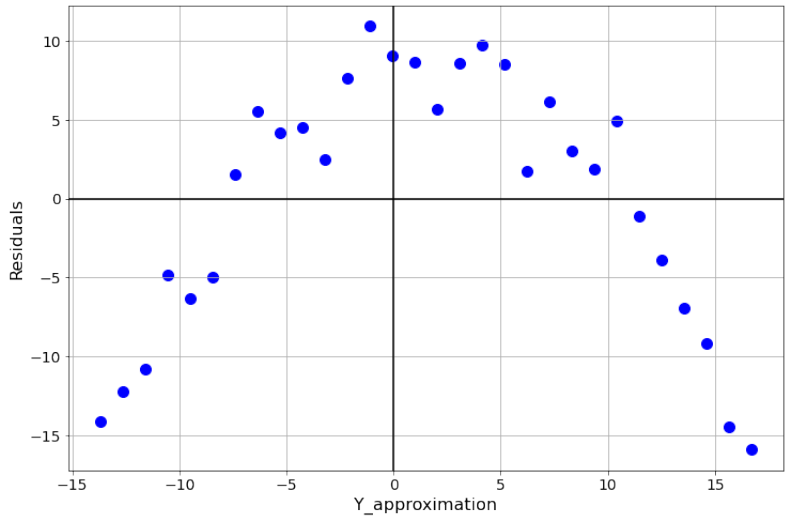
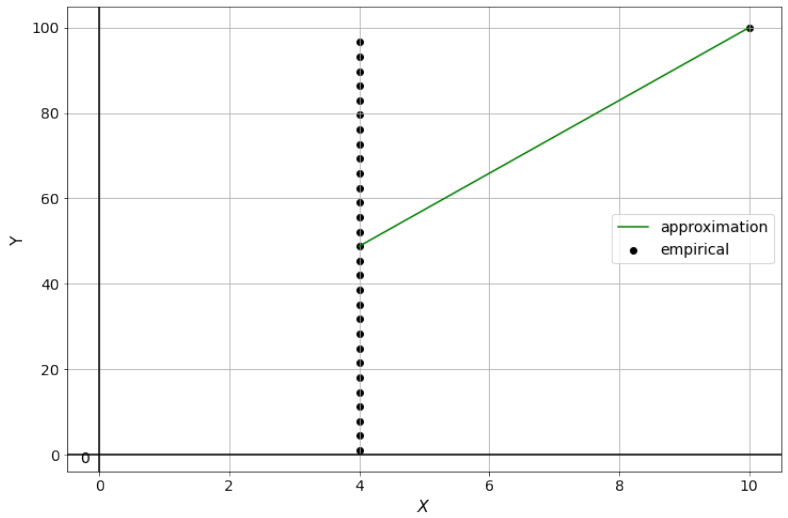


График распределения остатков в таком случае выглядит следующим образом:



Тут проблему может исправить добавление новой переменной, равной x^2.

Если предварительно не проанализировать свойства распределения переменной x или хотя бы не посмотреть на график распределения x, y то можно увидеть зависимость там, где ее нет.



Такие переменные следует просто отбрасывать

Так же важно, чтобы независимые переменные, если их больше одной были попарно не сильно коррелированы. Пороговое значение коррелированности, при котором переменные считаются недопустимо сильно коррелированы выбирают по-разному. Часто это значения 0.6, 0.7, 0.8 по модулю.

Таким образом, здесь мы перечислили некоторые проблемы, встречающиеся при построении линейной регрессии. На них следует обращать внимание, в противном случае модель будет ненадежной, и вы получите аппроксимацию чего угодно, но не той функции, которую ищете.

Полезные ссылки:

- Отличный бесплатный онлайн-курс по математической статистике: `https://stepik.org/course/326`
- Линейные модели в машинном обучении:`https://habr.com/ru/company/ods/blog/323890/`

Глава 2

Понятие градиента

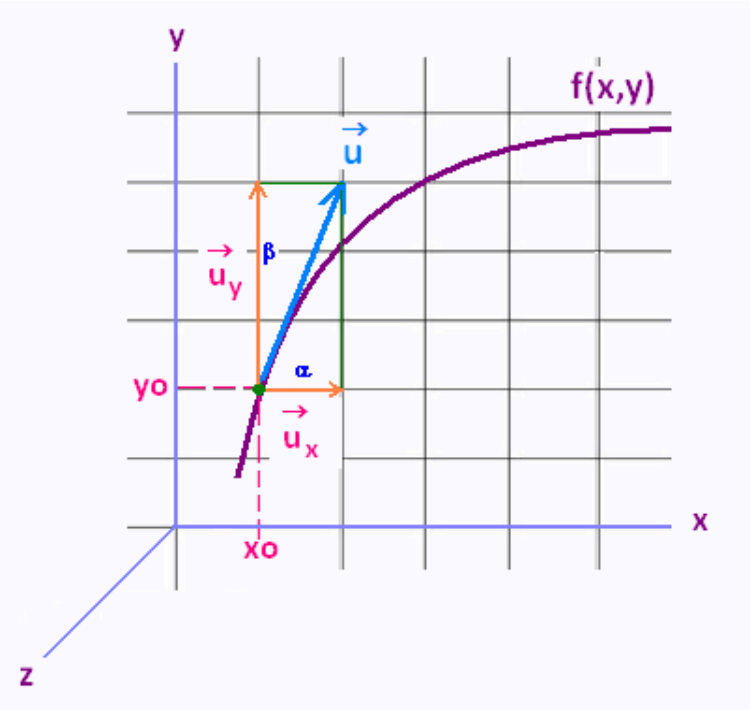
2.1 Производная по направлению

Нижеприведенный вывод формулы производной по направлению вектора $\vec{u}$ первом чтении можно опустить, и просто принять, что

$$\frac{df}{du} = f_y(x,y)cos(\alpha) + f_x(x,y)cos(\beta) \tag{2.1}$$

Здесь α, β - это углы, которые составляет вектор $\vec{u}$ с осями x, y. Сумма квадратов их косинусов равна 1. Рассмотрена функция f в двумерном пространстве, но формулу можно легко обобщить для произвольной размерности. Вывод формулы так же будет дан для двумерной функции, что сделано для простоты, так как в многомерном случае рассуждения аналогичны.

Рассмотрим функцию двух переменных $z = f(x,y)$ в декартовой системе координат. Дадим x, y очень малые приращения $\delta x, \delta y$, такие, что $\vec{\delta x} + \vec{\delta y} = \vec{\delta u}$. По теореме Пифагора $(\delta x)^2 + (\delta y)^2 = (\delta u)^2$. Это можно сделать, например, как на поясняющем рисунке ниже. Здесь вектор $\vec{\delta u}$ - направление, вдоль которого мы берем производную. Полагаем, что функция дифференцируема в точке x, y.



Тогда значение производной по направлению $\vec{u}$ мы можем записать в следующем виде:

$$\frac{df}{du} = \lim_{|u|\to 0} \frac{f(x+\delta x, y+\delta y) - f(x,y)}{|u|} \tag{2.2}$$

Преобразуем верхнюю часть уравнения (2.2).

$$\begin{aligned} &\frac{df}{du} = \lim_{|u|\to 0} \frac{f(x+\delta x,y+\delta y)-f(x,y)+f(x+\delta x,y)-f(x+\delta x,y)}{|u|} = \\ &= \lim_{|u|\to 0} \frac{(f(x+\delta x,y+\delta y)-f(x+\delta x,y))\frac{\delta y}{\delta y}+(f(x+\delta x,y)-f(x,y))\frac{\delta x}{\delta x}}{|u|} = \\ &= \lim_{|u|\to 0} \frac{f_y(x,y)\delta y+f_x(x,y)\delta x+o(\delta x)+o(\delta y)}{|u|} \end{aligned} \tag{2.3}$$

Устремляя предел $|u|$ в выражении (2.3) к нулю, получаем:

$$\begin{aligned} &\tfrac{df}{du} = f_y(x,y)\tfrac{\delta y}{|u|} + f_x(x,y)\tfrac{\delta x}{|u|} = \\ &= f_y(x,y)cos(\alpha) + f_x(x,y)cos(\beta) \end{aligned} \tag{2.4}$$

Здесь α, β - направляющие косинусы, как на рисунке выше. Нетрудно видеть, что сумма квадратов этих косинусов равна единице, т.е (α, β) - единичный вектор.

2.2 Градиент

Заметим, что производную по направлению можно представить как скалярное произведение векторов (f_x, f_y) и (α, β). Вектор из частных производных назовем градиентом:

$$\nabla f = (f_x, f_y) \tag{2.5}$$

Найдем, в каком случае она максимальна. Обозначим вектор направляющих косинусов как $\vec{l}$. Итак

$$\frac{df}{du} = (\nabla f, \vec{l}) = |\nabla f||\nabla l|cos(\alpha) = |\nabla f|cos(\alpha) \tag{2.6}$$

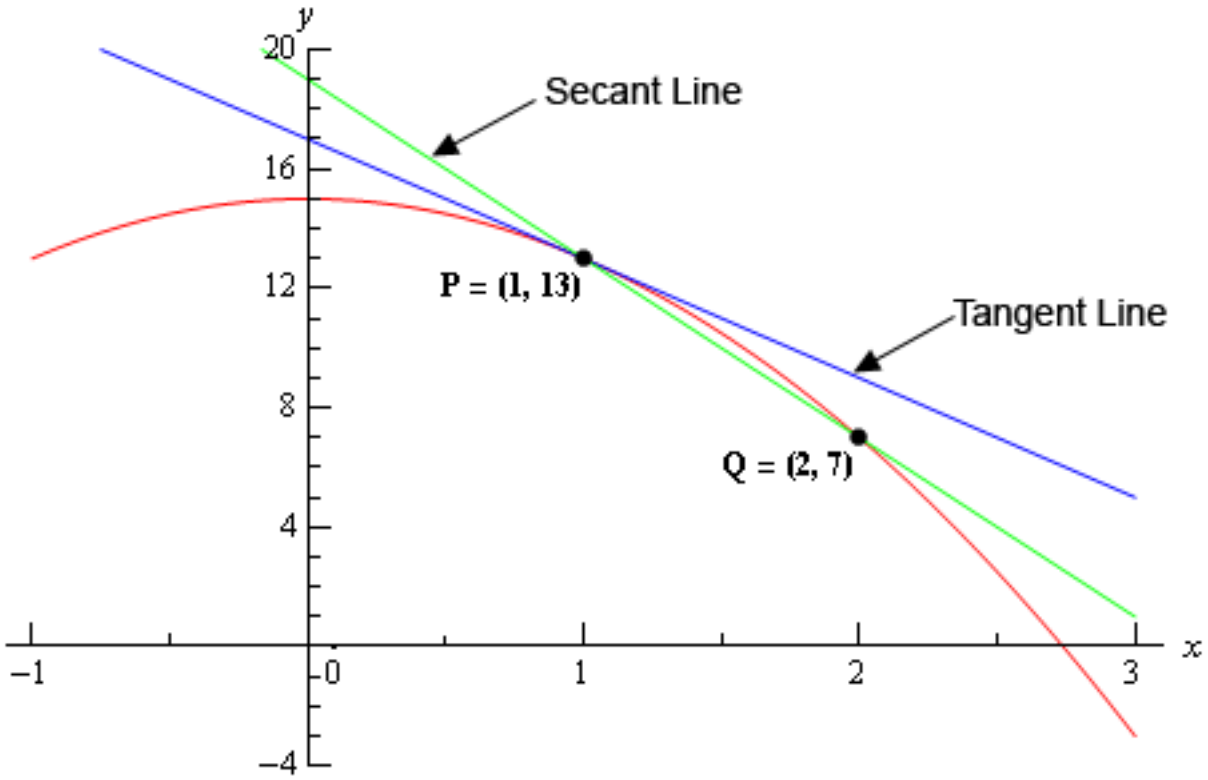
Здесь α - угол между направляющим вектором и градиентом. Поскольку $\nabla l| = 1$, то очевидно, что если $\alpha = 0$, то производная по направлению достигает своего наибольшего значения. Таким образом, градиент - это направление наискорейшего роста функции в данной точке (x,y). От точки к точке значение градиента в общем случае разное. Отмечу, что приведенные рассуждения могут быть легко обобщены на функцию сколько угодно большого числа аргументов, лишь бы функция была дифференцируема. Двумерный случай выбран для наглядности.

Упражнение: аналогично вышеприведенным рассуждениям найдите направление наискорейшего убывания функции.

Решение: в кратце если угол между направляющим вектором и градиентом будет равен π, то производная по направлению достигнет своего наименьшего значения. Интересно заметить, что это направление прямо противоположно градиенту. Это так называемый антиградиент.

2.3 Свойства градиента функции

Для начала нужно напомнить некоторые определения. Пусть $r(\vec{t})$ - кривая, зависящая от параметра t, дифференцируемая в точке t_0 в n-мерном пространстве. Тогда ее производная в точке t_0 будет направляющим вектором касательной к данной кривой как предельное положение секущей, проходящей через точку $r(\vec{t}_0)$. В двумерном случае это наиболее очевидно. Чтобы проиллюстрировать это, рассмотрим рисунок с графиком функции $y = 15 - 2x^2$ в точке $x = 1$:



Здесь если двигать зеленую прямую так, что правая точка пересечения этой прямой с графиком будет все ближе к левой, то в пределе зеленая прямая совместится с синей, являющейся касательной. Теперь перейдем в общему случаю в пространстве R^n. Действительно, пусть имеется дифференцируемая функция f от вектора x: $x, f(x) \in R^n$ проходящая через точки x_0 и x_1: $x_0, x_1 \in R^n$. Тогда следующее выражение будет направляющим вектором прямой, проходящей через точки данные точки:

$$\vec{sec} = \frac{f(x_0) - f(x_1)}{x_0 - x_1}$$

Устремляя x_1 к x_0, получим касательную к кривой f, проходящую через точку x_0: $sec(t) = tf'(x_0) + tf(x_0)$, $t \in R$.

Рассмотрим уравнение следующее уравнение:

$$F(x, y, z) = 0 \tag{2.7}$$

Оно определяет некоторую поверхность в трехмерном пространстве. Возьмем некоторую точку $p_0 = (x_0, y_0, z_0)$ такую, что F дифференцируема в этой точке по всем переменным. Проведем всевозможные кривые $r(t) = (x(t), y(t), z(t))$ через данную точку p_0 при условии, что все эти кривые принадлежат поверхности F, т.е выполняется равенство $F(r(t)) = 0$. Тогда так же существует значение параметра t t_0: $r(t_0) = p_0$. Поскольку кривые $r(t)$ лежат на F, то выражение (2.7) можно записать в виде

$$F(x(t), y(t), z(t)) = 0 \tag{2.8}$$

Возьмем полный дифференциал (2.7) в точке p_0. Поскольку оно равно нулю, то и дифференциал dF будет равен нулю.

$$F_x(p_0)x'_t(t_0) + F_y(p_0)y'_t(t_0) + F_z(p_0)z'_t(t_0) = 0 \tag{2.9}$$

Как как, как было показано ранее вектор $r'(\vec{t}) = (x'_t(t_0), y'_t(t_0), z'_t(t_0))$ является направляющим вектором касательной к вектор-функции $r(\vec{t}) = (x(t), y(t), z(t))$ в точке p_0, то выражение (2.9) представляет собой скалярное произведение в ортогональном пространстве. А значит, что вектор $(F_x(p_0), F_y(p_0), F_z(p_0))$ ортогонален касательной к любой прямой $r(\vec{t}) = (x(t), y(t), z(t))$: $r(\vec{t}) \in F(x, y, z)$ в точке p_0. Значит, все эти касательные лежат в одной плоскости. Если бы это было не так, т.е одна существовала бы прямая, пересекающая данную плоскость в точке p_0, то она образовывала бы с вектором нормали $(F_x(p_0), F_y(p_0), F_z(p_0))$ непрямой угол, что противоречит условию (2.9). Все вместе эти касательный прямые образуют так называемую касательную плоскость к поверхности (2.7) в точке p_0.

Определение. **Касательной плоскостью** к поверхности F в точке p_0 называется плоскость, в которой лежат касательные к всевозможным кривым $r(\vec{t})$ в данной точке: $r(\vec{t}) in F$.

Уравнение произвольной плоскости, проходящей через точку $p_0 = (x_0, y_0, z_0)$ и имеющей вектор нормали (A, B, C), как известно из курса аналитической геометрии имеет вид

$$A(x - x_0) + B(y - y_0) + C(z - z_0) = 0 \tag{2.10}$$

Как доказано выше вектор $(F'_x(p_0), F'_y(p_0), F'_z(p_0))$ является нормалью к касательной плоскости, проходящей через p_0. В таком случае уравнение касательной плоскости к функции F в точке p_0 будет иметь виде.

$$F'_x(p_0)(x - x_0) + F'_y(p_0)(y - y_0) + F'_z(p_0)(z - z_0) = 0 \tag{2.11}$$

Ну а поскольку градиент функции в точке p_0 так же представляет собой вектор $(F'_x(p_0), F'_y(p_0), F'_z$ то из этого следует, что **градиент функции F в точке p_0 ортогонален касательной плоскости функции в данной точке**.

Замечание 1. Мы рассматривали по сути двумерную поверхность, определяемую уравнением (2.7) в трехмерном пространстве. Все вышеприведенные рассуждения могут быть легко обобщены до произвольного R^n пространства, где поверхность имеет размерность $n - 1$.

Таким образом, градиент функции $G = F(x_1, x_2, ...x_n)$ в точке p_0 является направлением нормали к касательной плоскости поверхности, проходящей через точку p_0 и заданной уравнением $F(x_1, x_2, ...x_n) = 0$. Но стоит отметить, что мы доказали это только для ортогонального пространства R^n, так как опирались на то, что из равенства нулю суммы попарного произведения координат векторов следует ортогональность базисных векторов.

Докажем, что градиент ортогонален касательной к поверхности уровня функции в любой точке, где функция дифференцируема.

Пусть имеется скалярная функция $\varphi(\mathbf{x})$ от $\mathbf{n}$ переменных, имеющая непрерывные производные в точке $\mathbf{p}$. Напомню, что поверхность уровня функции - подмножество об-

ласти определения функции, такая что функция в этом подмножестве принимает одно и то же значение. Проведем поверхность уровня функции $\varphi(\mathbf{x})$ в точке $\mathbf{p}$, определяемую уравнением

$$\varphi(\mathbf{x}) = \varphi(\mathbf{p})$$

Далее проведем к этой поверхности уровня касательную плоскость в точке $\mathbf{p}$. Поскольку значение функции на поверхности уровня всюду постоянно, то производная по любому направлению $\mathbf{s}$ в касательной плоскости к поверхности уровня будет равно нулю

$$\frac{\partial \varphi}{\partial \mathbf{s}} = 0$$

Если вспомнить формулу производной по направлению (2.1), то можно заметить, что такое возможно, только если градиент ортогонален направлению $\mathbf{s}$. Действительно, производная по направлению представляет собой скалярное произведение градиента на вектор направления.

$$\frac{\partial \varphi}{\partial \mathbf{s}} = \varphi \cdot \mathbf{s} = |\varphi|\,|\mathbf{s}|\,cos(\varphi, \mathbf{s})$$

Тогда ее значение будет равно нулю (при условии, что нормы перемножаемых векторов строго положительны) будет равносильно равенству нулю косинуса угла $cos(\varphi, \mathbf{s})$, что возможно, только если векторы ортогональны. Так как направление выбрано произвольно, то из этого следует, что градиент ортогонален и касательной к поверхности уровня.

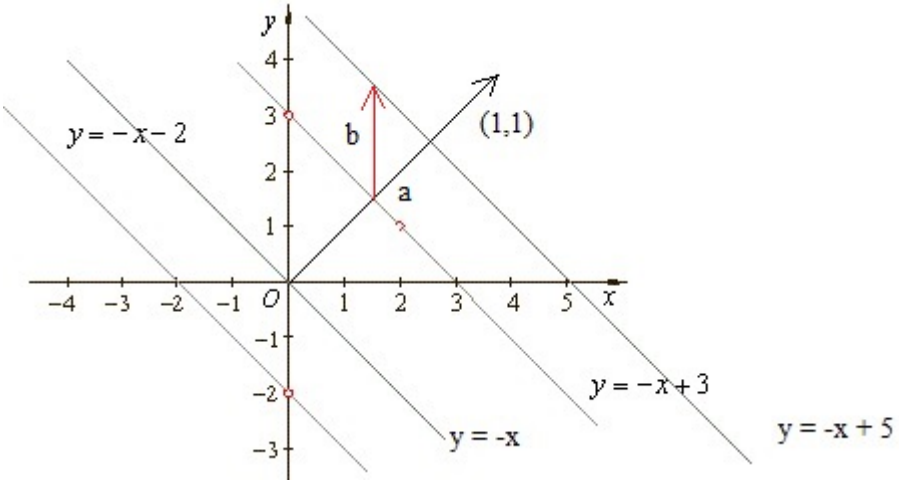
Замечание: в ходе доказательства использован факт, что если сумма попарных произведений векторов $\vec{a}$, $\vec{b}$ равна нулю, т.е $\sum_{i=1}^{n} a_i b_i = 0$, то они ортогональны. Это справедливо только в системах координат с ортогональным базисом, следовательно, доказательство получено только для таких систем. Но на практике обычно используют ортогональные координаты, поэтому этого в целом достаточно.

2.4 Примеры градиента функции

2.4.1 Параллельные прямые

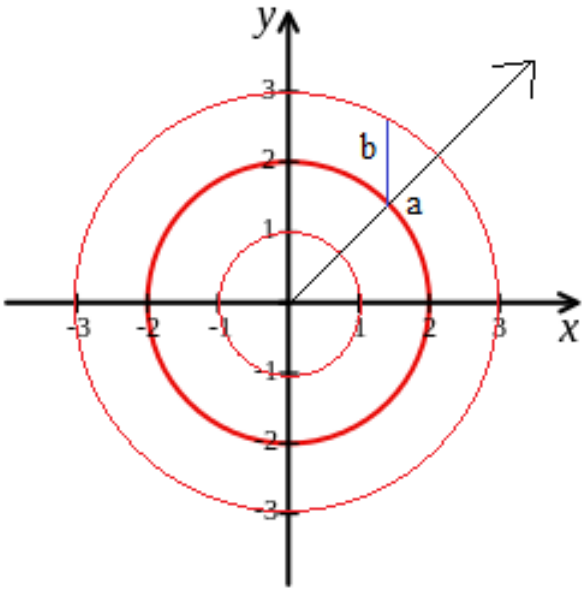
Рассмотрим функцию $z = x + y$. Заметим, что градиент этой функции равен постоянному вектору (1,1). Будем поочередно присваивать z значения $-2, 0, 3, 5...N$. Построим получившиеся графики с параметром t: $x+y = t$. Вернее, проекцию полученных графиков на ось Oxy для иллюстрации, т.к плоскости $x + y = t$ при различных t будут параллельны Oxy, как и сам вектор градиента. Как мы видим, направление роста функции z(x,y) совпадает с вектором (1,1). В любую другую сторону функция "растет медленнее"или даже убывает. На рисунке видно, что отрезок a, коллинеарный с градиентом короче, чем отрезок $\mathbf{b}$, в то время как $\mathbf{a}$ и $\mathbf{b}$ соединяют поверхности, на которых функция принимает постоянное значение. Иными словами, вдоль $\mathbf{a}$ функция растет быстрее. При этом градиент перпендикулярен поверхности уровня. В этом заключено свойство градиента как направления наискорейшего роста функции в точке. Область определения, в которой функция принимает одно и то же значение так же называют

эквипотенциальной поверхностью. Например, прямая $x + y = 1$ - эквипотенциальная поверхность функции $z = x + y$, при которой $z = 1$



2.4.2 Окружности с общим центром

Рассмотрим функцию $z = x^2 + y^2$. Пусть $z > 0$. Очевидно, ее эквипотенциальные поверхности - это окружности с центром в начале координат на вещественной плоскости. Градиент функции равен $(2x, 2y) = 2(x, y)$. Он совпадает с радиус-вектором, проведенным из центра окружности. Как известно из геометрии, он ортогонален любой касательной, проведенной окружности. Поскольку градиент параллелен Oxy, построим векторы градиента и поверхности уровня на этой плоскости.



Упражнение 1. Нарисуйте поверхности уровня и градиенты в различных точках для функции $z = x^2/4 + y^2/9$ на плоскости Oxy.

Решение: перебрать различные значения z и нарисовать соответствующие эллипсы. Это будут поверхности уровня. Найти аналитическое выражение вектора частных производных в произвольной точке, с его помощью построить нормали в различных

точках.

Упражнение 2. Нарисуйте поверхности уровня и градиенты в различных точках для функции $z = G\frac{m_1 m_2}{x^2+y^2}$ на плоскости Oxy. Здесь допустим, что произведение Gm_1m_2 равно 1.

Решение: аналогично предыдущей задаче.

2.4.3 Градиент в физике

Понятие градиента пришло из физики. Например, напряженность электрического поля равна градиенту его потенциала с обратным знаком: $\vec{E} = -\nabla\phi$. Сила потенциального поля выражается через его потенциальную энергию : $F = -\nabla U$. К примеру, сила гравитационного притяжения равна

$$\vec{F} = G\frac{m_1 m_2}{r^2}\vec{e} = -\nabla U = -grad(G\frac{m_1 m_2}{r}) \tag{2.12}$$

где U - потенциальная энергия тела в поле притяжения другого. $\vec{e}$ - здесь единичный вектор между телами. Символ ∇ приставленный к обозначению функции f означает градиент функции f, т.е $grad(f) = \nabla f$.

Если принять между телами притяжения расстояние r равным, например, 10, то получим поверхность в виде сферы радиусом 10, на котором потенциальная энергия тела в поле сил притяжения постоянна и равна $G\frac{m_1 m_2}{10}$, т.е эквипотенциальна.

2.4.4 Применение. Градиентный спуск

Не всегда имеется возможность найти глобальный/локальный максимум/минимум с помощью классических методов, таких как приравнивание к нулю производной. Например, если есть данные, в которых один параметр зависит от других по неизвестным закономерностям. Всё, что есть в данном случае - значения искомой величины Y и значения критериев $x_1, x_2, ...x_m$, в n по счёту наблюдении. Какой формуле подчиняется зависимость - неизвестно. Мы не будем подробно останавливаться на обсуждении данной проблемы. Отметим лишь, что для ее решения используется градиентный спуск (точнее, в настоящее время одна из его модификаций).

Без ущерба общности разберем алгоритм градиентного спуска нахождения минимума на примере функции Z = f(x,y). Здесь будет разобран самый примитивный градиентный спуск с постоянным шагом λ. В общем случае λ может меняться от шага к шагу: уменьшаться постепенно например. Или принимать такое оптимальное значение, что в следующей точке значение целевой функции будет как можно меньше в направлении антиградиента в пределах отрезка некоторой длины.

1. Выберем точку, с которой начнем алгоритм (x_0, y_0)
2. Выберем N - максимально возможное количество итераций, α - положительная константа, шаг градиента(скорость обучения)

3. Далее повторяем для $i = 0...N-1$

4. Найдем градиент функции в данной точке $\nabla f_i = (f_x(x_i, y_i), f_y(x_i, y_i))$

5. Новая точка, в которой вычисляется значение функции. Ищется в направлении антиградиента: $(x_{i+1}, y_{i+1}) = (x_i, y_i) - \alpha \frac{\nabla f_i}{|\nabla f_i|}$. Обратите внимание, что здесь вектор градиента мы делим на его норму, т.к нас интересует лишь направление, а скорость обучения устанавливается величиной α, которая в общем(не в этом) случае не постоянна.

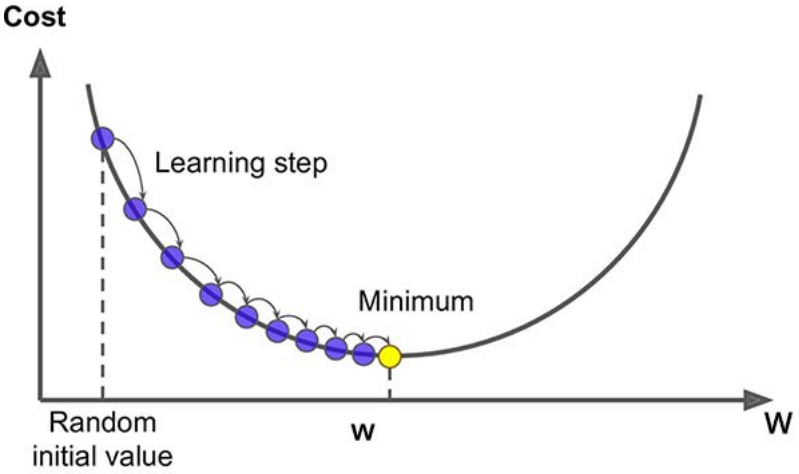
6. Новое значение функции: $f(x_{i+1}, y_{i+1})$

7. Далее повторяем следующий алгоритм (шаги 2-5), пока не исчерпаем максимальное количество итераций N либо не будет выполнено условие $|(f(x_{i+1}, y_{i+1}) - f(x_i, y_i)| < \epsilon$, где ϵ - константа, требуемая точность, которую мы устанавливаем заранее, либо $|(x_{i+1}, y_{i+1}) - (x_i, y_i)| < \epsilon$

Отмечу, что это лишь одна из простых возможных реализации, в которой заключена главная идея. Одна из проблем данного подхода в том, что если в некоторой точке все частные производные равны нулю, то данный алгоритм застрянет в ней.

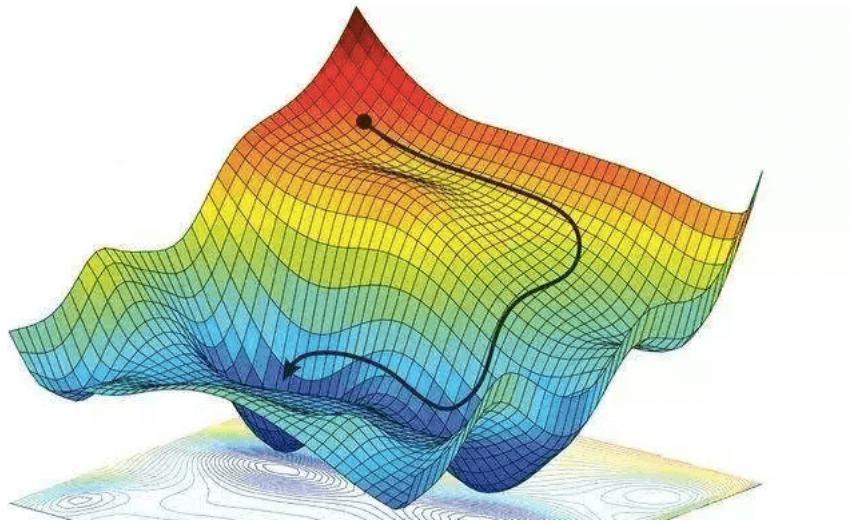
2.4.5 Графическая иллюстрация метода градиентного спуска

Градиентный спуск для функции одной переменной:



Приведенная выше функция выпуклая. Можно доказать, что для таких функций метод градиентного спуска всегда сходится при условии правильно выбранного шага. Для неудачно выбранного слишком большого шага метод, как можно догадаться, будет осциллировать.

Градиентный спуск функции двух переменных:



Здесь по оси **Z** представлены значения функции $f(x, y)$. Судя по поверхности функция имеет непростой вид. Если количество переменных не два, а тысячи, то очевидно, что нарисовать такую функцию не выйдет. Но этот график может дать представление о том, что происходит в многомерном пространстве, когда мы пытаемся минимизировать функцию.

Глава 3

Симплекс Метод

3.1 Задачи, приводящие к линейному программированию

3.1.1 Изготовление деталей на продажу

Завод изготавливает детали D_1 и D_2, каждая из которых продается на рынке по цене c_1 и c_2. Пусть завод изготовил x_1 и x_2 деталей D_1 и D_2 соответственно. Прибыль, которую он получит за эти детали равна $f(x_1, x_2) = c_1x_1 + c_2x_2$. Понятно, что стоит задача сделать эту прибыль как можно более крупной.

Для изготовления деталей нужно 3 вида ресурсов: S_1, S_2 и S_3. Количество этих ресурсов ограниченно значениями b_1, b_2, b_3 соответственно.

Для изготовления единицы детали D_1 нужно a_{11} ресурса S_1, a_{21} ресурса S_2 и a_{31} ресурса S_3. Таким образом, например, для изготовления x_1 деталей D_1 нужно $a_{11}x_1$ количества ресурса S_1.

Для изготовления единицы детали D_2 нужно a_{12} ресурса S_1, a_{22} ресурса S_2 и a_{32} ресурса S_3.

Очевидно, количество изготовленных деталей не может быть отрицательным. Записывая все вместе в виде системы, получаем:

$$\begin{aligned} & f(x_1, x_2) = c_1x_1 + c_2x_2 \to max \\ & a_{11}x_1 + a_{12}x_2 \leq b_1 \\ & a_{21}x_1 + a_{22}x_2 \leq b_2 \\ & a_{31}x_1 + a_{32}x_2 \leq b_3 \\ & x_i \geq 0 \end{aligned} \tag{3.1}$$

3.1.2 Задача о пищевом рационе для животных

Пусть ферме нужно накормить животных тремя видами продуктов P_1, P_2, P_3. Стоимость единицы каждого из них соответственно равна c_1, c_2, c_3. Тогда сумма изготовления x_1, x_2, x_3 единиц этих продуктов равна $f(x) = c_1x_1 + c_2x_2 + c_3x_3$. Из этих продуктов нужно составить рацион такой, чтобы в нем было не менее: белков b_1, углеводов b_2,

жиров b_3. При этом нужно сделать так, чтобы минимизировать цену составления этого рациона, т.е функцию $f(x)$. Ясно, что количество продуктов всех видов неотрицательно, т.е $x_i \geq 0$.

Известно, что в одной единице продукта P_1 содержится a_{11} белка, в P_2 - a_{12}, в P_3 - a_{13} соответственно. В единице продукта P_1 содержится a_{21} жиров, в P_2 - a_{22}, в P_3 - a_{23} соответственно. В единице продукта P_1 содержится a_{31} углеводов, в P_2 - a_{32}, в P_3 - a_{33} соответственно.

Складывая все вместе, получаем:

$$\begin{aligned} &f(x) = c_1x_1 + c_2x_2 + c_3x_3 \to min \\ &a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \geq b_1 \\ &a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \geq b_2 \\ &a_{31}x_1 + a_{32}x_2 + a_{33}x_3 \geq b_3 \\ &x_i \geq 0 \end{aligned} \tag{3.2}$$

3.2 Каноническая задача

Для начала рассмотрим каноническую задачу линейного программирования. Найти максимум функции:

$$f(x) = \sum_{j=1}^{n} c_j x_j \tag{3.3}$$

при ограничениях

$$\sum_{j=1}^{n} a_{ij} x_j = b_i, i = 1...m, m < n \tag{3.4}$$

$$x_j \geq 0, j = 1...n \tag{3.5}$$

Здесь n - количество переменных, данных в условии задачи. m - количество линейных уравнений - ограничений. Предполагается, что $b_i >= 0$. Даже если в условии задачи это не так, этого легко можно добиться, умножая уравнение, в котором $b_i < 0$ на -1. Для решения данных типов уравнений был разработан симплекс - метод, который будет рассмотрен далее.

Поскольку число вершин, определяемых (3.4) конечно, и где-то в них достигается экстремум, то вся суть алгоритма симплекс-метода заключается в том, чтобы перебрать эти вершины. В действительности, перебор не случайный, не в хаотичном порядке, а подчиняется простому правилу. Он осуществляется в направлении возрастания функции. Нужно смотреть на каждом шаге изменение какой из переменных даст наибольший прирост целевой функции. Это переменная, чей коэффициент в целевой функции положителен и максимален по модулю среди положительных. Если таких нет - то процесс поиска завершен, т.к расти больше некуда. Если ищется минимум целевой функции, то нужно смотреть, элементарно малый прирост какой из переменных даст наибольшую убыль целевой функции. Это переменная, чей коэффициент в целевой функции отрицателен и максимален по модулю среди отрицательных. Если таких нет, значит убывать

больше некуда. Эти правила похожи на правило роста функции в направлении градиента.

Поверхность, на которой ищется решение задачи (3.3) представляет собой политоп. Политоп - это конечное объединение симплексов. Симплекс - это обобщение треугольника на многомерные пространства. Например, в одномерном пространстве симплекс - это отрезок из 2 точек. В двумерном - треугольник из 3. В трехмерном - пирамида. Алгоритм симплекс метода сходится, поскольку число вершин политопа конечно, и в одной из них целевая функция принимает экстремальное значение.

Введем кое-какие понятия, необходимые нам в дальнейшем.

Переменные x_i, образующие единичную подматрицу размерностью $m \times m$ в системе ограничений (3.4) называются базисными. Все остальные - свободными. Поскольку базисные переменные образуют в системе линейных ограничений единичную матрицу, то каждая такая переменная входит с коэффициентом единица только в одно из уравнений (3.4). Благодаря этому, их можно выразить как функции остальных, так называемых свободных переменных. Это может внести путаницу с традиционным понятием базиса, где векторы выражаются как линейные комбинации базисных векторов. Нужно просто запомнить данную терминологию в контексте симплекс-метода.

Базисное решение - это такое решение, что базисные переменные равны $x_i = b_i$, все остальные равны нулю.

Решение является допустимым, если оно удовлетворяет условиям (3.5) и (3.4) .

3.3 Основная задача линейного программирования

Если в условиях (3.3) (3.4) (3.5), где f - целевая функция, которую надо максимизировать есть нестрогие неравенства в системе линейных ограничений, то задача линейного программирования формулируется в следующем виде:

$$f(x) = \sum_{j=1}^{n} c_j x_j \tag{3.6}$$

$$\begin{array}{l} \sum\limits_{j=1}^{n} a_{ij} x_j \geq b_i, i = 1...p \\ \sum\limits_{j=1}^{n} a_{ij} x_j \leq b_i, i = p+1...m \end{array} \tag{3.7}$$

$$x_j \geq 0, j = 1...n \tag{3.8}$$

Это так называемая основная задача линейного программирования. Она может быть приведена к каноническому виду (3.4) путем введения дополнительных переменных $x_j, k = n+1...n+m$. Действительно, пусть, например, i-е из неравенств задано в виде $\sum\limits_{j=1}^{n} a_{ij} x_j \leq b_i$. Значит, существует такое число x_{n+i}, которое обращает первое неравенство в равенство $\sum\limits_{j=1}^{n} a_{ij} x_j + x_{n+i} = b_i$. Если же $\sum\limits_{j=1}^{n} a_{ij} x_j \geq b_i$ то найдется такое

неотрицательное число x_{n+i}, что $\sum_{j=1}^{n} a_{ij}x_j - x_{n+i} = b_i$. Для другого неравенства найдется другое число x_{n+k}, приводящее неравенство к равенству и т.д. Это можно записать в след. виде:

$$\sum_{j=1}^{n} a_{ij}x_j + x_{n+i} = b_i, i = 1...m, m < n \tag{3.9}$$

Линейные ограничения в системе, заданной в виде (3.7) перед применением симплекс-метода должны быть избавлены от знаков неравенства, т.е задача перед решением должна быть приведена к каноническому виду.

Докажем несколько утверждений, связанных с основной задачей линейного программирования.

Для начала напомню, что такое линейная функция. Пусть у нас имеется векторное пространство $\mathbf{X}$ (например R^2) над полем $\mathbf{K}$ в этом поле (это могут быть к примеру вещественные либо комплексные числа, мы будем работать с вещественными). Тогда отображение $f : \mathbf{X} -> \mathbf{K}$ называется линейной функцией, если для любых $\mathbf{x}, \mathbf{y} \in \mathbf{X}$ и для любых $\alpha, \beta \in \mathbf{K}$ справедливы свойства:
1. $f(x + y) = f(x) + f(y)$
2. $f(\alpha x) = \alpha f(x)$
3. $f(\alpha x + \beta y) = \alpha f(x) + \beta f(y)$
Третье свойство является следствием первых двух. Теперь докажем пару важных свойств, на которых основан симплекс-метод.

Утверждение 1. Если функция f линейна в векторном пространстве R^n, $d\vec{x} = (dx_1, dx_2, dx_3, ...dx_n)$ - направление изменения функции (к примеру это может быть градиент) f в точке $\vec{x}$, $d\vec{x}, \vec{x} \in R^n$, $\alpha \in R$ то

$$f(\vec{x} + \alpha d\vec{x}) = f(\vec{x}) + \alpha f((dx_1, 0, ...0)) + \alpha f((0, dx_2, ...0)) + \alpha f((0, 0, ...dx_n))$$

Доказательство. Действительно, согласно третьему свойству $f(\vec{x} + \alpha d\vec{x}) = f(\vec{x}) + \alpha f(d\vec{x}) = f(\vec{x} + (x_1, 0...0)) + (0, x_2, ...0) + ... = f(\vec{x} + (x_1, x_2, 0...0)) + (0, 0, x_3, ...0) + ... = .. = f(\vec{x} + (x_1, x_2, x_3...x_n))$. Так как $d\vec{x} = (dx_1, 0...0) + (0, dx_2, ...0) + (0, 0, ...dx_n)$ то получаем требуемое свойство. Оно означает, что если линейная функция растет быстрее всего в точке $\vec{x}$ в направлении $d\vec{x}$, то необязательно "двигаться по всем координатам сразу, одновременно это можно сделать пошагово, за n шагов (n - размерность пространства), т.е каждый раз двигаясь в направлении лишь одной из координат значения вектора изменения, от 1й до n координаты (хотя порядок в тут в общем не важен). В случае, если имеются ограничения то это направление может не совпадать с градиентом целевой функции. И еще нужно= отметить, что на каждом шаге новое значение функции должно удовлетворять ограничениям, определяющим область определения.

В общем случае для канонической задачи доказано, что если решение есть, то оно достигается за конечное число шагов из некой начальной позиции x_0, если на каждом шаге изменять ту координату, которая дает наибольший прирост/убыль целевой функции не выходя за область определения. В целом это похоже на безусловную оптимизацию с помощью градиента. Читатели, знакомые с программированием могут так

же увидеть здесь аналогию с жадными алгоритмами, когда на каждом шаге нужно делать ход, доставляющий наибольший выигрыш на данном шаге.

В качестве примера рассмотрим функцию $f(x_1, x_2) = 2x_1 - 3x_2 + 7$. Ее градиент равен $(2, -3)$. Будем двигаться в направлении градиента, умноженного на 0.5. Несложно проверить, что здесь $f((x_1, x_2) + 0.5(2, -3)) = f(x_1, x_2) + f(0.5(2, -3)) = f((x_1 + 0.5 * 2, x_2 + 0)) + f(0.5(0, -3))$

Утверждение 2. Для основной задачи линейного программирования целевая функция (3.6) при ограничениях (3.7) может достигать экстремума только на границе области.

Что прежде всего можно сказать о целевой функции (3.6)? Она линейна. Из этого следует, что если коэффициенты c_j не равны нулю одновременно все, то среди частных производных целевой функции найдется как минимум одна, отличная от нуля. Предполагается, что хоть один коэффициент c_j не равен нулю, иначе задача не имеет смысла, т.к тогда целевая функция будет константой независимо от переменных. Для локального экстремума в точке необходимо, чтобы в ней все частные производные равнялись нулю. Поскольку коэффициенты c_j постоянны, то градиент этой функции тоже есть вектор-константа, постоянный в любой точке области определения. Следовательно, необходимое условие локального экстремума не выполняется нигде и функция постоянно растет в некотором направлении. Следовательно, если целевая функция и достигает экстремума, то происходит это где-то на границе области.

Действительно, возьмем любую внутреннюю точку P_1 области, ограниченной системой (3.7). Тогда в некоторой ее окрестности $\epsilon > 0$ найдется другая точка P_2, принадлежащая области определения, такая, что $f(P_2) - f(P_1) > 0$, если взять точку P_2 на малом расстоянии от точки P_1 в направлении возрастании градиента, а значит, и целевой функции. В таком случае решение может быть одно (одна точка), бесконечное число решений, представляющих собой одну из границ области целиком, если целевая функция на этой границе эквипотенциальна, т.е постоянна, либо решений может вообще не быть, если система (3.7) не совместна.

Задачу (3.6) при условиях (3.7) в двумерном и трехмерном случае, при которых можно визуально увидеть, какая из точек области находится "максимально далеко"в направлении градиента. Либо можно двигаться в направлении градиента очень малыми шагами, строя поверхности уровня и смотреть, где такие поверхности пересекаются с областью определения. Нужно двигаться так до тех пор, пока есть хоть одна точка пересечения. Очевидно, что вычислительно это очень дорого, особенно в многомерном случае, поэтому так поступают только для двумерной основной задачи, и то в качестве иллюстрации.

Небольшая историческая справка. Это задача называется задачей линейного программирования. Метод ее решения, так называемый симплекс-метод был разработан в 1947 году американским математиком Джорджем Бернардом Данцигом. Возможно, вы задались вопросом: причем здесь программирование? В целом это никакое не программирование. Данное название (линейное программирование) сформировалось исторически. В 20 веке, когда исследовались методы решения данной задачи все это назвали

программированием, чтобы денег на это дали побольше, поскольку "программирование"само по себе очень хорошо финансировалось, больше, чем теоретические изыскания в области математики. Возможно, правильнее было бы назвать эту задачу задачей поиска экстремума линейной функции при наличии системы линейных ограничений. Это чисто математическая задача. Есть еще задача нелинейного программирования. Там тоже нужно найти минимум/максимум целевой функции, но там ограничения и целевая функция (что-то одно из них либо и то, и другое) уже нелинейны относительно свободных переменных. Их решение существенно сложнее, чем линейная задача, и зачастую решается только численными методами. Там уже нет готовых рецептов, универсальных алгоритмов решения, как для задачи линейного программирования.

Существует специальный алгоритм симплекс-метода в виде таблицы. На первый взгляд он может показаться сборником длинных непонятных последовательностей инструкции. Чтобы лучше его понять, дадим его интуитивное представление на примере простых задач в виде линейных неравенств и уравнений, чтобы увидеть, что происходит "под капотом"симлекс-метода.

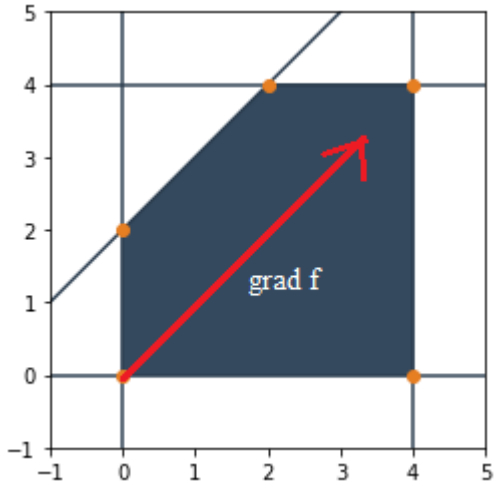
Подробно про алгоритм симплекс-метода вы можете почитать в книге [1] на с.317.

3.3.1 Пример 1

Даны условия

$$\begin{aligned} & f(x) = x_1 + x_2 \to max \\ & -x_1 + x_2 \leq 2 \\ & x_1 \leq 4 \\ & x_2 \leq 4 \\ & x_1 \geq 0 \\ & x_2 \geq 0 \end{aligned} \tag{3.10}$$

Графически их можно представить так:



Градиент целевой функции равен $\nabla f = (1, 1)$. Из рисунка видно, что точка, расположенная "дальше всего"в направлении градиента, или, говоря строже,точки, скалярное произведение радиус-вектора которой с градиентом максимально - это точка $(4, 4)$. В ней функция достигает максимума, равного 8.

Решим теперь эту задачу другим способом. Принципиально он ничем от графического не отличается. Такой же рост в направлении градиента по сути. Просто в силу того, что мы не можем представить себе многомерное пространство, возникает необходимость подходить к решению более формально, алгоритмически. Чтобы избавиться от знаков неравенства, ведем неотрицательные переменные x_3, x_4, x_5. Получим задачу в виде:

$$\begin{aligned} & f(x) = x_1 + x_2 \to max \\ & -x_1 + x_2 + x_3 = 2 \\ & x_1 + x_4 = 4 \\ & x_2 + x_5 = 4 \\ & x_i \geq 0 \end{aligned} \tag{3.11}$$

Базисными, как видно, являются x_3, x_4, x_5. Перепишем систему в виде

$$\begin{aligned} & f(x) = x_1 + x_2 \to max \\ & x_3 = 2 + x_1 - x_2 \\ & x_4 = 4 - x_1 \\ & x_5 = 4 - x_2 \\ & x_i \geq 0 \end{aligned} \tag{3.12}$$

Пусть $x_1, x_2 = 0$, тогда $x_3 = 2, x_4 = 4, x_5 = 4$. Тогда первое базисное решение: $(0, 0, 2, 4, 4)$.

Будем увеличивать переменную, прирост которой даст наибольший прирост функции. При этом не рассматриваем базисную переменную, так как манипуляции с ней приведут к той же системе уравнений, с которой мы начали. Смотрим, коэффициент какой из переменных x_1, x_2 максимален. Они равны. В таком случае, увеличим переменную с наибольшим индексом, x_2. Она входит в 1 и 3 уравнения. Если мы выразим x_2 в 3м уравнении через x_5 и увеличим с нуля до 4, уменьшая x_5 до нуля, то нарушим условие допустимости решения в 1м уравнений, так как x_3 станет равным -2. Поэтому выразим через нее переменные x_1, x_3 в 1м уравнении и подставим в целевую функцию. Получим:

$$\begin{aligned} & f(x) = 2 + 2x_1 - x_3 \to max \\ & x_2 = 2 + x_1 - x_3 \\ & x_4 = 4 - x_1 \\ & x_5 = 2 - x_1 + x_3 \\ & x_i \geq 0 \end{aligned} \tag{3.13}$$

Новое базисное решение: $(0, 2, 0, 4, 2)$. Коэффициент x_3 входит со знаком минус, поэтому увеличивать эту переменную не стоит, ее увеличение приведет к уменьшению целевой функции. Увеличим x_1, т.к коэффициент при данной переменной максимальный и положительный. Заметим, что если бы во все данные уравнения x_1 входил бы со знаком $+$, то решение задачи было бы неограниченно большим. Действительно, можно было бы увеличивать x_1 на 1, 2, 3.... и соответственно увеличивать значения переменных x_2, x_4, x_5 без нарушений условия допустимости решения. Выразим x_1 через x_5, x_3 в 3м

уравнении как в наиболее строгом относительно x_1, по аналогии с предыдущим пунктом. Получим систему:

$$\begin{aligned} &f(x) = 6 + x_3 - 2x_5 \to max \\ &x_2 = 4 - x_5 \\ &x_4 = 2 + x_5 - x_3 \\ &x_1 = 2 - x_5 + x_3 \\ &x_i \geq 0 \end{aligned} \tag{3.14}$$

Базисное решение : (2,4,0,2,0). Видим, что можно увеличит переменную x_3. Действуя аналогично, получаем:

$$\begin{aligned} &f(x) = 8 - x_4 - x_5 \to max \\ &x_2 = 4 - x_5 \\ &x_3 = 2 - x_4 - x_5 \\ &x_1 = 4 - x_4 \\ &x_i \geq 0 \end{aligned} \tag{3.15}$$

Как видно, увеличить переменные в целевой функции так, чтобы увеличилась целевая функция нельзя. Уменьшить тоже, т.к x_4, x_5 как свободные переменные равны нулю. Значит, мы пришли к решению. Итак, если при максимизации коэффициенты при переменных целевой функции неположительны, то значит, что расчет закончен, мы пришли к ответу. Базисное решение: (4,2,4,0,0). При данном базисном решении получаем ответ: 8.

Аналогично можно проделать с помощью классической симплекс таблицы. Ее алгоритм такой же, как в вышеприведенных рассуждениях, но формализованный. Вышеприведенные рассуждения удобны для людей, но не подходят для реализации на компьютере, которому нужен алгоритм. Поэтому для написания программы для решения задачи линейного программирования удобнее применять симплекс-таблицу.

3.3.2 Пример 2

Раннее я не упомянул, что не всегда возможно выделить единичную подматрицу в системе (3.4) так, чтобы правая часть системы из свободных коэффициентов b_i было неотрицательной, и, следовательно, все переменные были тоже неотрицательны в базисном решении. В этом случае используют прием введения искусственных переменных (slack variables). При этом в целевую функцию добавляется их сумма, умноженная на некоторое очень большое положительное число M. Если задача на максимум, то M вводится со знаком минус, если на минимум - со знаком $+$. Поясню на примере. Дана система:

$$\begin{aligned} & f(x) = x_1 + x_2 \to max \\ & x_1 \geq 4 \\ & x_2 \geq 4 \\ & x_1 \leq 6 \\ & x_2 \leq 6 \\ & x_i \geq 0 \end{aligned} \tag{3.16}$$

Для избавления от знаков неравенства введем дополнительные переменные. Получим

$$\begin{aligned} & f(x) = x_1 + x_2 \to max \\ & x_1 - x_3 = 4 \\ & x_2 - x_5 = 4 \\ & x_1 + x_7 = 6 \\ & x_2 + x_8 = 6 \\ & x_i \geq 0 \end{aligned} \tag{3.17}$$

Для возможности выделения единичной матрицы введем искусственные переменные x_4, x_6. Получим систему:

$$\begin{aligned} & f(x) = x_1 + x_2 - M(x_4 + x_6) \to max \\ & x_1 - x_3 + x_4 = 4 \\ & x_2 - x_5 + x_6 = 4 \\ & x_1 + x_7 = 6 \\ & x_2 + x_8 = 6 \\ & x_i \geq 0 \end{aligned} \tag{3.18}$$

Отметим, что для того, чтобы система была совместна, в итоговом решении искусственные переменные должны быть равны нулю. Действительно, пусть у нас есть равенство $3 = 3$. Если бы добавим в правую часть, например, число 2, то $3 + 2! = 3$. Так же и с данным типом переменных в линейных равенствах. Далее решаем как обычно, пользуясь теми же соображениями, что и в предыдущем примере. получим:

$$\begin{aligned} & f(x) = x_1 + x_2 - M(x_4 + x_6) \to max \\ & x_4 = 4 - x_1 + x_3 \\ & x_6 = 4 - x_2 + x_5 \\ & x_7 = 6 - x_1 \\ & x_8 = 6 - x_2 \\ & x_i \geq 0 \end{aligned} \tag{3.19}$$

Базисной решение: $(0, 0, 0, 4, 0, 4, 6, 6)$. Выберем переменную x_2 для возрастания в направлении градиента. Выразим ее 2 втором уравнении как наиболее строгом относительно нее.

$$\begin{aligned} f(x) &= 4 - x_6 + x_5 + x_1 - M(x_4 + x_6) \to max \\ x_2 &= 4 - x_6 + x_5 \\ x_4 &= 4 - x_1 + x_3 \\ x_7 &= 6 - x_1 \\ x_8 &= 2 + x_6 - x_5 \\ x_i &\geq 0 \end{aligned} \tag{3.20}$$

Базисной решение: $(0, 4, 0, 4, 0, 0, 6, 2)$. Выберем переменную x_5

$$\begin{aligned} f(x) &= 6 - x_8 + x_1 - M(x_4 + x_6) \to max \\ x_5 &= 2 + x_6 - x_8 \\ x_4 &= 4 - x_1 + x_3 \\ x_7 &= 6 - x_1 \\ x_2 &= 6 - x_8 \\ x_i &\geq 0 \end{aligned} \tag{3.21}$$

Базисной решение: $(0, 6, 0, 4, 2, 0, 6, 0)$. Выберем переменную x_1

$$\begin{aligned} f(x) &= 10 - x_8 + x_3 - x_4 - M(x_4 + x_6) \to max \\ x_1 &= 4 - x_4 + x_3 \\ x_2 &= 6 - x_8 \\ x_5 &= 2 + x_6 - x_8 \\ x_7 &= 2 + x_4 - x_3 \\ x_i &\geq 0 \end{aligned} \tag{3.22}$$

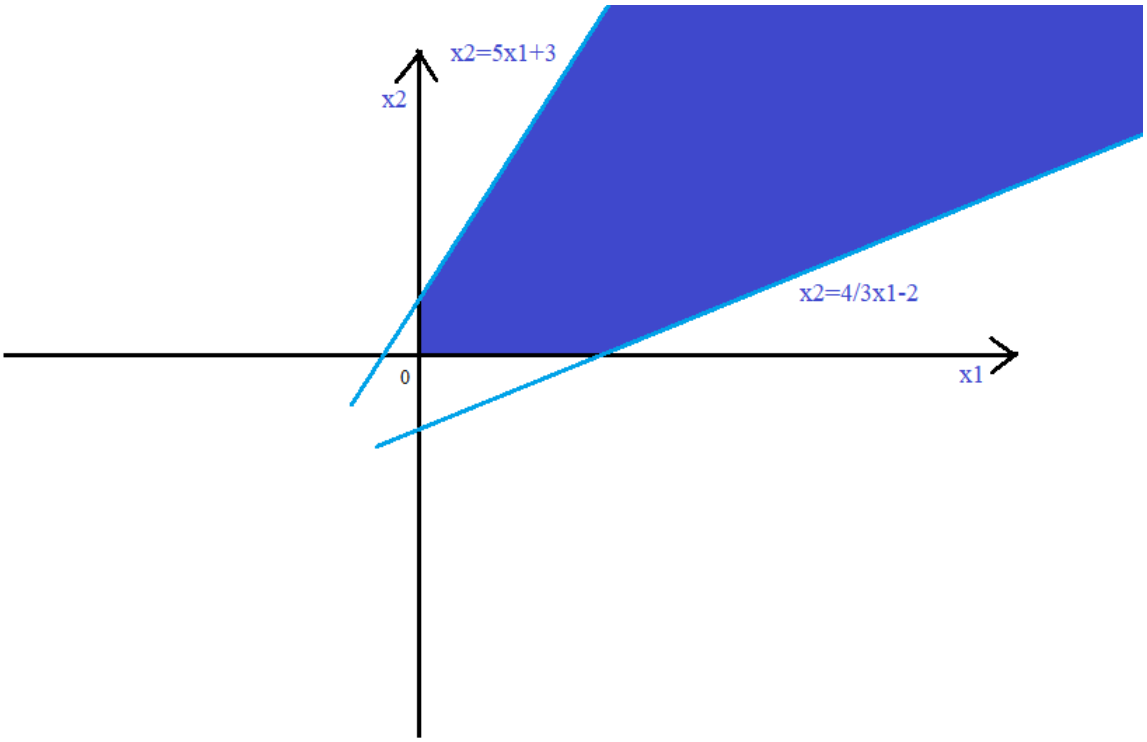
Базисной решение: $(4, 6, 0, 0, 2, 0, 2, 0)$. Выберем переменную x_3. Заметим, что если бы в правой части все x_3 были бы со знаком плюс, то эту переменную можно было бы бесконечно увеличивать, при этом базисные переменные x_1, x_7 тоже для выполнения равенства возрастали бы. Это бы привело к тому, что целевая функция бесконечно возрастала, и в ответе мы бы получили, что целевая функция неограниченна. Графически это значило бы, что область определения, при которой $x_i \geq 0$ так же не ограничена.

$$\begin{aligned} f(x) &= 12 - x_8 - x_7 - M(x_4 + x_6) \to max \\ x_3 &= 2 + x_4 - x_7 \\ x_1 &= 6 - x_7 \\ x_2 &= 6 - x_8 \\ x_5 &= 2 + x_6 - x_8 \\ x_i &\geq 0 \end{aligned} \tag{3.23}$$

Базисной решение: $(6, 6, 2, 0, 2, 0, 0, 0)$. В целевой функции все переменные имеют отрицательный знак, значит, увеличивать дальше некуда, расчет окончен. Видим, что искусственные переменные x_4, x_6 в конечном решении равны нулю, значит, решение есть, ответ : 12.

3.3.3 Пример 3

Теперь рассмотрим простой пример, когда решение неограниченно большое. Приведенный ниже график области определения не претендует на абсолютную точность, он схематичен.



$$\begin{aligned} &f(x) = 2x_1 + 3x_2 \to max \\ &-5x_1 + x_2 \leq 3 \\ &-4x_1 + 3x_2 \geq -2 \\ &x_i \geq 0 \end{aligned} \tag{3.24}$$

Очевидно, что в данной неограниченной области определения при $x_1 > 0, x_2 > 0$ данная целевая функция неограниченно большая. Посмотрим, как симплекс метод даст нам об этом знать. Умножим второе уравнение на -1, получим систему

$$\begin{aligned} &f(x) = 2x_1 + 3x_2 \to max \\ &-5x_1 + x_2 \leq 3 \\ &4x_1 - 3x_2 \leq 2 \\ &x_i \geq 0 \end{aligned} \tag{3.25}$$

Добавим переменные x_3, x_4 чтобы получить каноническую систему

$$\begin{aligned} &f(x) = 2x_1 + 3x_2 \to max \\ &-5x_1 + x_2 + x_3 = 3 \\ &4x_1 - 3x_2 + x_4 = 2 \\ &x_i \geq 0 \end{aligned} \tag{3.26}$$

Как видно, x_3, x_4 образуют единичную подматрицу, значит, их можно выбрать как базисные.

$$\begin{aligned} & f(x) = 2x_1 + 3x_2 \to max \\ & x_3 = 3 + 5x_1 - x_2 \\ & x_4 = 2 - 4x_1 + 3x_2 \\ & x_i \geq 0 \end{aligned} \tag{3.27}$$

Базисное решение - $(0, 0, 3, 2)$.

В целевой функции имеются переменные с положительными коэффициентами, значит, возможно, удастся увеличить некоторые из них и тем самым целевую функцию.

Как видно, коэффициент при x_2 положителен и максимален, значит, прирост x_2 даст наибольший прирост целевой функции.

Первое уравнение допускает введение x_2 как базисной без нарушения неотрицательности переменных, введем ее. Получим

$$\begin{aligned} & f(x) = 9 - 3x_3 + 17x_1 \to max \\ & x_2 = 3 + 5x_1 - x_3 \\ & x_4 = 11 + 11x_1 - 3x_3 \\ & x_i \geq 0 \end{aligned} \tag{3.28}$$

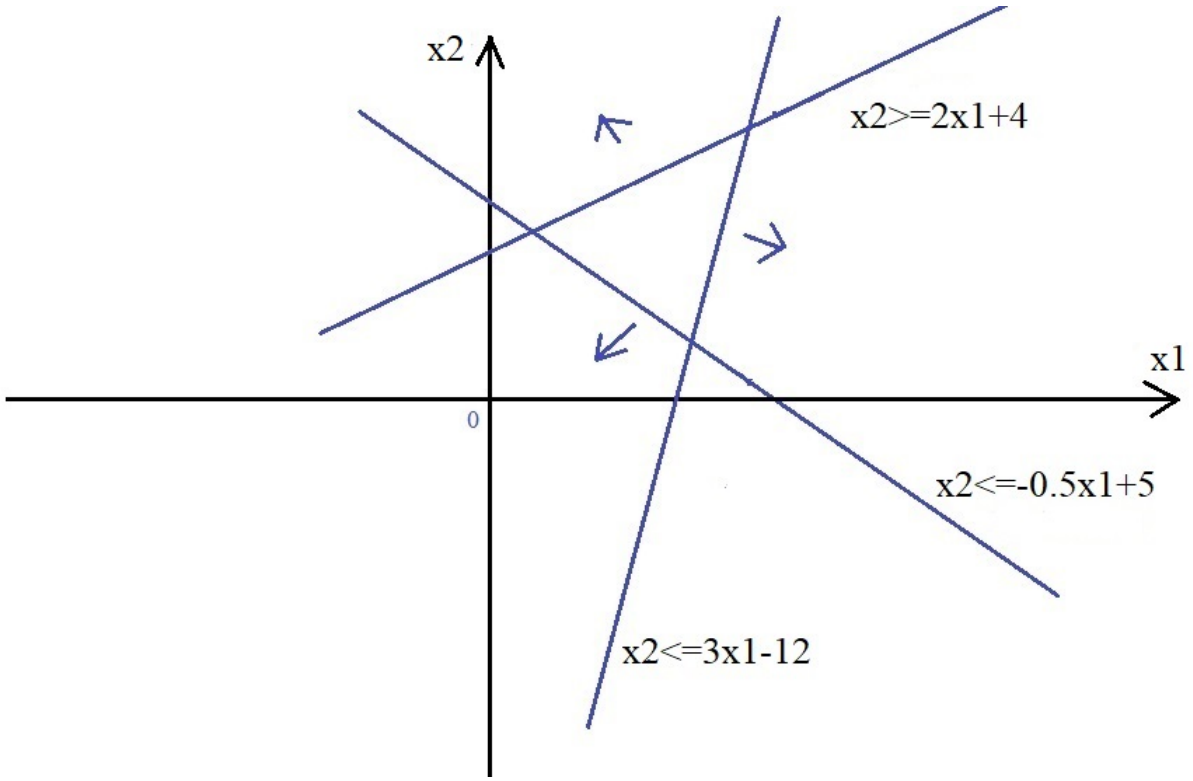
Базисное решение - $(0, 3, 0, 11)$.

Наибольший прирост даст x_1. Прирост x_3 лишь уменьшит целевую функцию.

Но, как видно, x_1 входит в правую часть всех уравнений со знаком $+$. Если мы будем увеличивать x_1, то возможно без нарушения линейных ограничений также бесконечно много увеличивать переменные x_2, x_4. Это значит, что целевая функция может бесконечно возрастать вместе с x_1, решение неограниченно большое.

3.3.4 Пример 4

Теперь рассмотрим пример, когда решения нет, т.к система линейных ограничений несовместна, область определения - пустое множество. Ниже схематичный рисунок



$$
\begin{aligned}
&f(x) = 3x_1 + 2x_2 \to max \\
&x_2 - 2x_1 \geq 4 \\
&x_2 + 0.5x_1 \leq 5 \\
&x_2 - 3x_1 \leq -12 \\
&x_i \geq 0
\end{aligned}
\tag{3.29}
$$

Умножим последнюю строку на -1, чтобы все свободные слагаемые в правой части стали неотрицательны

$$
\begin{aligned}
&f(x) = 3x_1 + 2x_2 \to max \\
&x_2 - 2x_1 \geq 4 \\
&x_2 + 0.5x_1 \leq 5 \\
&-x_2 + 3x_1 \geq 12 \\
&x_i \geq 0
\end{aligned}
\tag{3.30}
$$

Чтобы избавиться от неравенств, введем переменные x_3, x_4, x_5, после чего система примет вид

$$
\begin{aligned}
&f(x) = 3x_1 + 2x_2 \to max \\
&x_2 - 2x_1 - x_3 = 4 \\
&x_2 + 0.5x_1 + x_4 = 5 \\
&-x_2 + 3x_1 - x_5 = 12 \\
&x_i \geq 0
\end{aligned}
\tag{3.31}
$$

Видно, что единичную подматрицу, представляющую собой базисную функцию, выделить не получится, поэтому вводим фиктивные переменные x_6, x_7. В целевую функцию их вводим с коэффициентом -M, где M - очень большое число, заведомо большее

любых чисел, которые мы можем получить в ходе работы алгоритма, чтобы не было основания их увеличивать. Получаем

$$\begin{aligned} & f(x) = 3x_1 + 2x_2 - M(x_6 + x_7) \to max \\ & x_2 - 2x_1 - x_3 + x_6 = 4 \\ & x_2 + 0.5x_1 + x_4 = 5 \\ & -x_2 + 3x_1 - x_5 + x_7 = 12 \\ & x_i \geq 0 \end{aligned} \tag{3.32}$$

Отметим, что в случае поиска минимума фиктивные переменные вводятся с положительным знаком +M. Вводим в базис переменные x_4, x_6, x_7

$$\begin{aligned} & f(x) = 3x_1 + 2x_2 - M(x_6 + x_7) \to max \\ & x_4 = 5 - 0.5x_1 - x_2 \\ & x_6 = 4 - x_2 + 2x_1 + x_3 \\ & x_7 = 12 + x_5 - 3x_1 + x_2 \\ & x_i \geq 0 \end{aligned} \tag{3.33}$$

Базисное решение: $(0, 0, 0, 5, 0, 4, 12)$. Как видно из целевой функции, наибольший прирост даст x_1. Третье условие наиболее строго относительно данной переменной. Используя данное условие, введем x_1 в базис

$$\begin{aligned} & f(x) = 12 + x_5 + 3x_2 - M(x_6 + x_7) \to max \\ & x_1 = (12 + x_5 + x_2 - x_7)/3 \\ & x_6 = 12 - x_2/3 + 7x_5/3 + x_3 - 2x_7/3 \\ & x_4 = 3 - x_5/6 - 7x_2/6 + x_7/6 \\ & x_i \geq 0 \end{aligned} \tag{3.34}$$

Базисное решение: $(4, 0, 0, 3, 0, 12, 0)$. Увеличим переменную x_2. Используем 3е условие, как наиболее строгое. Получим

$$\begin{aligned} & f(x) = (138 + 4x_5 - 3x_4)/7 - M(x_6 + x_7) \to max \\ & x_2 = (18 - x_5 - x_4 + x_7)/3 \\ & x_6 = (234 + 48x_5 - 13x_7 - x_4 + 21x_3)/21 \\ & x_1 = (6x_5 - 6x_7 - x_4 + 102)/21 \\ & x_i \geq 0 \end{aligned} \tag{3.35}$$

Базисное решение: $(102/21, 6, 0, 0, 0, 234/21, 0)$. Единственное, что можно увеличить - это переменная x_5. Выражаем ее через первое условие, как единственно возможно (в других случаях она перенесется в левую часть со знаком минус, что недопустимо). Получим

$$\begin{aligned} & f(x) = (210 - 7x_4 - 12x_2)/7 - M(x_6 + x_7) \to max \\ & x_5 = 18 - x_4 + x_7 - 3x_2 \\ & x_6 = (1098 - 49x_4 + 48x_7 - 144x_2 + 21x_3)/21 \\ & x_1 = (210 - 7x_1 - 18x_2)/21 \\ & x_i \geq 0 \end{aligned} \tag{3.36}$$

Базисное решение: $(210/21, 0, 0, 0, 18, 1098/21, 0)$. В целевой функции не осталось переменных с положительным коэффициентом, это значит, что увеличивать больше нечего, расчет окончен. При этом фиктивная переменная x_6 не равна нулю. Это значит, что условия задачи не совместимы, решения нет. Ответ: система не совместна, нет решений.

3.3.5 Пример 5

В данном примере рассмотрим алгоритм нахождения минимума. Как вам известно, для нахождения минимума нужно двигаться в направлении антиградиента. Пусть целевая функция задана в виде $f = -x_1 + 2x_2 - x_3 - x_4$. Ее градиент - $(-1, 2, -1, -1)$ - направление роста. Тогда антиградиент будет равен $(1, -2, 1, 1)$. Антиградиент - это направление наискорейшего убывания. Т.к двигаться в направлении убывания переменных мы не можем (в данном случае это переменная x_2), то двигаемся в направлении возрастания переменных, дающих убыль целевой функции. Алгоритм в целом ничем не отличается от поиска максимума, за исключением того, что нужно продолжать поиск, пока в целевой функции не останутся только положительные элементы. На каждом шаге выбираем для роста переменную с отрицательным, максимальным по модулю коэффициентом. Продемонстрируем на примере:

$$\begin{aligned} &f(x) = -x_1 + 2x_2 - x_3 - x_4 \to min \\ &-x_1 + x_2 + x_3 = 2 \\ &x_1 + x_2 + x_4 = 4 \\ &x_i \geq 0 \end{aligned} \tag{3.37}$$

Выразим базисные переменные. Получим

$$\begin{aligned} &f(x) = -x_1 + 2x_2 - x_3 - x_4 \to min \\ &x_3 = 2 + x_1 - x_2 \\ &x_4 = 4 - x_1 - x_2 \\ &x_i \geq 0 \end{aligned} \tag{3.38}$$

Опорное решение - $(0, 0, 2, 4)$.

Пусть нам надо найти минимум функции f. Чтобы ее уменьшить, можно увеличить любую небазисную переменную с отрицательным коэффициентом. Это x_1. Важно: небазисную. Ибо если базисную подставить в целевую функцию, то ясно, что ее коэффициент там равен нулю. Да и если попытаться ее увеличить, то это приведет к цепной реакции, требующей уменьшения/увеличения других переменных, что приведет к каше. Выражаем ее с помощью 2го уравнения. Получим

$$\begin{aligned} &f(x) = -10 + x_4 + 5x_2 \to min \\ &x_3 = 6 - 2x_2 - x_4 \\ &x_1 = 4 - x_4 - x_2 \\ &x_i \geq 0 \end{aligned} \tag{3.39}$$

Все переменные в целевой функции входят с положительным знаком. Увеличить какую-либо переменную так, чтобы уменьшить целевую функцию не выйдет. Значит, расчет окончен. Опорное решение - $(4, 0, 6, 0)$. Ответ : -10.

3.3.6 Пример 6. Симплекс-таблица

Теперь найдем максимум той же задачи, параллельно показывая, что при этом происходит в симплекс таблице. Тем самым, попытаемся связать уже имеющиеся знания о том, что происходит при нахождении экстремума с помощью симплекс-метода с формальным алгоритмом в виде таблицы.

Будем решать задачу параллельно с помощью симплекс-таблицы и системы линейных уравнений, как было выше.

$$\begin{aligned} &f(x) = -x_1 + 2x_2 - x_3 - x_4 \to max \\ &x_3 - x_1 + x_2 = 2 \\ &x_4 + x_1 + x_2 = 4 \\ &x_i \geq 0 \end{aligned} \tag{3.40}$$

Для начала кое-какие пояснения. Далее речь идет о некой n-й итерации алгоритма.

Здесь Bvar (Basis Variable) - переменная, которая в данный момент в базисе, Bval (Basis Value) - её значение на данной итерации.

c_j - коэффициенты переменных, фигурирующих в данной задаче в целевой функции, заданной в условии задачи (они неизменны), c_{jB} - коэффициенты при базисных переменных в целевой функции (берутся из c_j).

Всё, что левее коэффициентов c_{jB} - это a_{ij} - коэффициенты при переменных, заданных в системе условий, определяющих задачу. От итерации к итерации значения a_{ij}. Это связано с вводом в базис других переменных.

$\triangle_j$ - значение коэффициента переменной x_j в целевой функции на данной итерации при условии, что в целевой функции нет базисных переменных, чего можно добиться путем подстановки вместо базисных переменных их представлений через свободные переменные в целевую функцию. Легко проверить, что для базисных переменных $\triangle_j$ равняется нулю. $\triangle_j$ меняется, поскольку на каждой итерации меняются базисные функции, подставляемые в целевую функцию $f(x)$.

Значения $\triangle_j$ задаются формулой:

$$\triangle_j = c_j - z_j = c_j - \sum_{i=1}^{m} c_{iB} a_{ij} \tag{3.41}$$

Здесь z_j - это фактически сумма коэффициентов при переменной x_j, полученная после подстановки базисной переменной в целевую функцию, не считая c_j.

Отношение $\frac{BVal}{a_{ir}}$ - это частное значения текущей базисной переменной i и a_{ir} - коэффициента из системы условий, где r - индекс переменной, выбранной для ввода в базис (как дающей наибольший прирост). С помощью данного отношения определяется, через какое уравнение выразить новую базисную переменную и ввести в базис.

Выбирается так, чтобы выбор был наиболее строгим к увеличению переменной r, иначе говоря, давал наименьший прирост среди все остальных уравнений.

Подробно описывать алгоритм симплекс метода я не буду, это вы можете найти в книге [1] на стр 321-323. Здесь я обращу внимания на параллели между симплекс-таблицей и линейными уравнениями, решая одну задачу одновременно двумя способами.

ШАГ 1

Занесем данные в таблицу. Получим

			-1	2	-1	-1	c_j
c_{iB}	BVar	BVal	x_1	x_2	x_3	x_4	$\frac{BVal}{a_{ir}}$
-1	x_3	2	-1	[1]	1	0	$\frac{2}{1}=2$
-1	x_4	4	1	1	0	1	$\frac{4}{1}=4$
			0	-2	-1	-1	z_j
			-1	4	0	0	$\triangle_j$

Аналогичная система уравнений

$$\begin{aligned} &f(x) = -x_1 + 2x_2 - x_3 - x_4 \to max \\ &x_3 = 2 + x_1 - x_2 \\ &x_4 = 4 - x_1 - x_2 \\ &x_i \geq 0 \end{aligned} \tag{3.42}$$

или, если в правой части оставить только свободные переменные

$$\begin{aligned} &f(x) = -x_1 + 2x_2 - x_3 - x_4 \to max \\ &x_3 - x_1 + x_2 = 2 \\ &x_4 + x_1 + x_2 = 4 \\ &x_i \geq 0 \end{aligned} \tag{3.43}$$

Сразу подставим базисные переменные в целевую функцию. Получим

$$\begin{aligned} &f(x) = -x_1 + 4x_2 - 6 \to max \\ &x_3 - x_1 + x_2 = 2 \\ &x_4 + x_1 + x_2 = 4 \\ &x_i \geq 0 \end{aligned} \tag{3.44}$$

Опорное решение - $(0, 0, 2, 4)$. Можно увеличить переменную x_2. Первое уравнение позволяет увеличить x_2 до $2/1$, второе до $4/1$. Первое более строго относительно x_2, поэтому используем его для введение x_2 в базис.

ШАГ 2

			-1	2	-1	-1	c_j
c_{iB}	BVar	BVal	x_1	x_2	x_3	x_4	$\frac{BVal}{a_{ir}}$
2	x_2	2	-1	1	1	0	$--$
-1	x_4	2	[2]	0	-1	1	$\frac{2}{2}=1$
			-4	2	3	-1	z_j
			3	0	-4	0	$\triangle_j$

$$\begin{aligned} & f(x) = 3x_1 - 4x_3 + 2 \to max \\ & x_2 - x_1 + x_3 = 2 \\ & x_4 + 2x_1 - x_3 = 2 \\ & x_i \geq 0 \end{aligned} \tag{3.45}$$

$$\begin{aligned} & f(x) = 3x_1 - 4x_3 + 2 \to max \\ & x_2 = 2 + x_1 - x_3 \\ & x_4 = 2 - 2x_1 + x_3 \\ & x_i \geq 0 \end{aligned} \tag{3.46}$$

Опорное решение - $(0, 2, 0, 2)$. Увеличим x_1.

Отметим, как меняются коэффициенты a_{ij} в симплекс-таблице. Пусть на s-строке мы выбрали r-переменную x_r для ввода в базис. a_{sr} называется в таком случае разрешающим элементом. Тогда элементы s строки в системе ограничений пересчитываются по формуле

$$a_{sj} = \frac{a_{sj}}{a_{sr}} \tag{3.47}$$

а все остальные a_{ij}

$$a_{ij} = a_{ij} - a_{sj}a_{ir} = a_{ij} - \frac{a_{sj}}{a_{sr}}a_{ir} \tag{3.48}$$

Для этих уравнений существует так же мнемоническое правило прямоугольника. Они могут показаться запутанными, но в сущности это всего лишь закон того, как изменяются коэффициенты системы уравнений, если в одном из равенств выразить одну переменную через другие в этом же равенстве и подставить эту переменную во все остальные уравнения системы.

ШАГ 3

			-1	2	-1	-1	c_j
c_{iB}	BVar	BVal	x_1	x_2	x_3	x_4	$\frac{BVal}{a_{ir}}$
2	x_2	3	0	1	0.5	0.5	$--$
-1	x_1	1	1	0	-0.5	0.5	$\frac{4}{1} = 4$
			-1	2	1.5	0.5	z_j
			0	0	-2.5	-1.5	$\triangle_j$

$$\begin{aligned} & f(x) = 5 - 2.5x_3 - 1.5x_4 \to max \\ & x_2 + 0.5x_3 + 0.5x_4 = 3 \\ & x_1 + 0.5x_4 - 0.5x_3 = 1 \\ & x_i \geq 0 \end{aligned} \tag{3.49}$$

$$\begin{aligned} & f(x) = 5 - 2.5x_3 - 1.5x_4 \to max \\ & x_2 = 3 - 0.5x_3 - 0.5x_4 \\ & x_1 = 1 - 0.5x_4 + 0.5x_3 \\ & x_i \geq 0 \end{aligned} \tag{3.50}$$

Положительных коэффициентом в целевой функции не осталось, расти дальше некуда. Опорное решение - $(1, 3, 0, 0)$. Следовательно, ответ - 5.

Аналогичное решение этой же задачи с помощью графического метода и табличного симплекс-метода можете посмотреть в книге [1] на стр. 327.

3.4 Двойственная задача линейного программирования (dual problem)

В этой короткой главе мы рассмотрим так называемую двойственную задачу линейного программирования. В случае линейного программирования это по сути та же самая задача, но записанная по-другому. Рассмотрим на примере.

Пусть дана следующая задача

$$\begin{aligned} & f(x) = x_1 - x_2 \to max \\ & -x_1 + 2x_2 \leq 4 \\ & 3x_1 + 2x_2 \leq 14 \\ & x_i \geq 0 \end{aligned} \tag{3.51}$$

Ответ в данной задаче: $4\frac{2}{3}$. При желании можете самостоятельно это проверить.

Зададимся вопросом: а можно ли на основании системы ограничении (3.51) понять, каким числом ограничена целевая функция? Оказывается, да. Умножим систему ограничений на некоторые числа y_1, y_2, получим

$$\begin{aligned} & f(x) = x_1 - x_2 \to max \\ & y_1(-x_1 + 2x_2) \leq 4y_1 \\ & y_2(3x_1 + 2x_2) \leq 14y_2 \\ & x_i \geq 0 \end{aligned} \tag{3.52}$$

Далее сложим ограничения левых и правых частей неравенств (3.52) соответственно, и сгруппируем относительно x_1, x_2

$$x_1(-y_1 + 3y_2) + x_2(2y_1 + 2y_2) \leq 4y_1 + 14y_2 \tag{3.53}$$

Выберем y_1, y_2 так, чтобы целевая функция была меньше левой части вышеприведенного неравенства (3.53), но при этом так, чтобы выражение $4y_1 + 14y_2$ было как можно меньше, чтобы наложить на значение целевой функции как можно более сильное ограничение

$$f(x) = x_1 - x_2 \leq x_1(-y_1 + 3y_2) + x_2(2y_1 + 2y_2) \leq 4y_1 + 14y_2 \tag{3.54}$$

Из этих неравенств получим задачу, которая является двойственной к (3.51)

$$\begin{aligned} & f(x) = 4y_1 + 14y_2 \to min \\ & -y_1 + 3y_2 \geq 1 \\ & 2y_1 + 2y_2 \geq -1 \\ & x_i \geq 0 \end{aligned} \tag{3.55}$$

Замечания

1. Для того, чтобы привести основную задачу линейного программирования (при поиске максимума) к ее двойственной форме, не обязательно, чтобы в системе ограничений (3.51) все левые части были нестрого меньше правых. То есть, нет необходимости, чтобы везде было

$$\sum_{j=1}^{n} a_{ij}x_j \leq b_i, i = 1...m \tag{3.56}$$

Достаточно, чтобы в (3.56) левые части были не больше правых. Это значит, что допустимы так же знаки равенства $=$. Главное, чтобы не было отношений $\geq$ между левой и правой частью ограничений (3.56) в задаче поиска максимума целевой функции, иначе уже нельзя гарантировать выполнение условий (3.54) и (3.53).

Действительно, если у нас имеются, например, система условий $x_1 \leq 3, x_2 \leq 4$, то мы можем быть уверены, что $x_1 + x_2 \leq 7$. Но в случае, если $x_1 \geq 3, x_2 \leq 4$, то обещать, что $x_1 + x_2 \leq 7$ уже нельзя, т.к x_1 может быть сколь угодно больше 3, может быть очень большим числом, а про x_2 мало что известно.

2. Вышеприведенные рассуждения подходят для произвольных задач на поиск максимума линейного программирования, удовлетворяющих условиям (3.56). Рассуждения, приводящие к двойственной задаче были приведены без ограничения общности для наглядности и простоты.

Теперь решим задачу (3.55). После формулировки исходной задачи в двойственной форме далее алгоритм решения такой же, как и решения основной задачи линейного программирования. Для начала приведем задачу к канонической форме. Вычтем из первого неравенства системы (3.55) переменную y_3 и умножим на (-1). Умножим второе неравенство на (-1) и прибавим переменную y_4. Получим систему

$$\begin{aligned} &f(x) = 4y_1 + 14y_2 \to min \\ &y_1 - 3y_2 + y_3 = -1 \\ &-2y_1 - 2y_2 + y_4 = 1 \\ &x_i \geq 0 \end{aligned} \tag{3.57}$$

Умножим первое уравнение на -1, и добавим фиктивную переменную y_5

$$\begin{aligned} &f(x) = 4y_1 + 14y_2 + My_5 \to min \\ &-y_1 + 3y_2 - y_3 + y_5 = 1 \\ &-2y_1 - 2y_2 + y_4 = 1 \\ &x_i \geq 0 \end{aligned} \tag{3.58}$$

Выразим как базисные переменные y_4, y_5 и подставим переменную y_5 в базис. Получим

$$\begin{aligned} &f(x) = 4y_1 + 14y_2 + M(1 + y_1 - 3y_2 + y_3) = \\ &= M + y_1(4 + M) + y_2(14 - 3M) + y_3M \to min \\ &y_5 = 1 + y_1 - 3y_2 + y_3 \\ &y_4 = 1 + 2y_1 + 2y_2 \\ &x_i \geq 0 \end{aligned} \tag{3.59}$$

Базисное решение - (0,0,0,1,1). Коэффициент при y_2 отрицательный (т.к M - это очень большое положительное число). Выразим y_2 как базисную переменную. Получим целевую функцию и систему в виде

$$\begin{aligned} f(x) &= 14/3 + 18y_1 + y_5(M - 14/3) + 14/3y_3 \\ y_2 &= 1/3 + 1/3y_1 - 1/3y_5 + 1/3y_3 \\ y_4 &= 5/3 + 8/3y_1 + 2/3y_3 - 2/3y_5 \\ x_i &\geq 0 \end{aligned} \tag{3.60}$$

Базисное решение - (0,1/3,0,5/3,0). В целевой функции больше нет отрицательных переменных, фиктивная переменная обнулилась, значит, мы получили ответ. Подставляя базисное решение в исходную целевую функцию (3.55) получаем, что ответ, внезапно, то же $4\frac{2}{3}$. На самом деле это закономерно, поскольку, как я уже подчеркивал раннее, двойственная задача линейного программирования - это по сути та же задача, переписанная по другому, а значит ответ должен быть тот же. Возможн, у вас возник вопрос: Зачем все это нужно было, не проще было бы сразу решить задачу на поиск максимума? Возможно и проще. Но в некоторых ситуациях решение двойственной задачи может быть проще, поэтому порой задачу линейного программирования решают с ее помощью.

Заметим, что можно так же сделать двойственную задачу для перехода от задачи на минимум к задаче на максимум, это работает и в обратную сторону.

Полезные ссылки:

- Пример решения задачи линейного программирования явно с помощью линейных уравнений как аналога симплекс-таблицы для простого уравнения: `https://geekrodion.medium.com/linear-programming-simplex-method-bc586b9aec10`
- Пример решения двойственной задачи линейного программирования: `https://geekrodion.medium.com/dual-simplex-method-b88250ecded1`
- Калькулятор онлайн для симплекс таблиц. Выводит подробные решения в виде симплекс-таблиц либо уравнений на выбор `https://math.semestr.ru/simplex/simplex.php`

Глава 4

Литература

[1] *A1* Пантелеев А.В., Летова Т.А. - Методы оптимизации в примерах и задачах. Издательство ”Лань”, 2015

[2] *A2* Вентцель Е.С - Исследование операций. Задачи, принципы, методология. Издательство - главная редакция физико-математической литературы издательства ”Наука”, 1980

[3] *A3* Э. Митчелл, Р. Уэйт - Метод конечных элементов для уравнений с частными производными. Издательство ”Мир”, 1981.

[4] *A4* Р. Курант - курс дифференциального и интегрального исчисления. Том 1. Издательство "Наука 1967.

[5] *A5* Р. Курант - курс дифференциального и интегрального исчисления. Том 2. Издательство "Наука 1967.

[6] *A6* К.А. Бохан, И.А. Егорова, К. В. Лащенов - Математический анализ. Том 1. Издательство ”Просвещение”, 1972.

[7] *A7* К.А. Бохан, И.А. Егорова, К. В. Лащенов - Математический анализ. Том 2. Издательство ”Просвещение”, 1972.

[8] *A8* Н.Е. Кочин - Векторное исчисление и начала тензорного исчисления. Девятое издание. Издание "Наука"Москва - 1965.