МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное учреждение
высшего профессионального образования
"Казанский (Приволжский) федеральный университет"
Институт экологии и природопользования

УТВЕРЖДАЮ
Проректор по образовательной деятельности КФУ
Проф. Таюрский Д.А.
_____ _____20___ г.

подписано электронно-цифровой подписью

Программа дисциплины

Методы машинного обучения Б1.В.ДВ.13

Направление подготовки: 05.03.06 - Экология и природопользование
Профиль подготовки:
Квалификация выпускника: бакалавр
Форма обучения: очное
Язык обучения: русский
Автор(ы):
Савельев А.А.
Рецензент(ы):
Зарипов Ш.Х.

СОГЛАСОВАНО:
Заведующий(ая) кафедрой: Зарипов Ш. Х.
Протокол заседания кафедры No ___ от "____" _____ 201__г
Учебно-методическая комиссия Института экологии и природопользования:
Протокол заседания УМК No ____ от "____" _____ 201__г

Регистрационный No 244517

Казань
2017

Содержание

Программу дисциплины разработал(а)(и) профессор, д.н. (профессор) Савельев А.А. кафедра моделирования экологических систем отделение экологии ,
Anatoly.Saveliev.aka.saa@gmail.com

1. Цели освоения дисциплины

Целями освоения дисциплины (модуля) Методы машинного обучения являются знакомство с методами машинного обучения и приобретение навыков их применения для решения практических задач.

2. Место дисциплины в структуре основной образовательной программы высшего профессионального образования

Данная учебная дисциплина включена в раздел " Б1.В.ДВ.13 Дисциплины (модули)" основной образовательной программы 05.03.06 Экология и природопользование и относится к дисциплинам по выбору. Осваивается на 4 курсе, 7 семестр.

Дисциплина относится к ФТД разделу ООП и развивает представление о статистических методах. Для ее освоения нужны знания по теории вероятностей и математической статистике, представление об информационных технологиях и начальные навыки программирования. Освоение данной дисциплины способствует лучшему пониманию методов, используемых в статистической обработке данных.

3. Компетенции обучающегося, формируемые в результате освоения дисциплины /модуля

В результате освоения дисциплины формируются следующие компетенции:

Шифр компетенции	Расшифровка приобретаемой компетенции
ОПК-2 (профессиональные компетенции)	владение базовыми знаниями фундаментальных разделов физики, химии и биологии в объеме, необходимом для освоения физических, химических и биологических основ в экологии и природопользования; владение методами химического анализа, владение знаниями о современных динамических процессах в природе и техносфере, о состоянии геосфер Земли, экологии и эволюции биосферы, глобальных экологических проблемах, а также методами отбора и анализа геологических и биологических проб; владение навыками идентификации и описания биологического разнообразия, его оценки современными методами количественной обработки информации
ОПК-3 (профессиональные компетенции)	владение профессионально профилированными знаниями и практическими навыками в общей геологии, теоретической и практической географии, общего почвоведения и использование их в области экологии и природопользования
ОПК-4 (профессиональные компетенции)	владение базовыми общепрофессиональными (общеэкологическими) представлениями о теоретических основах общей экологии, геоэкологии, экологии человека, социальной экологии, охраны окружающей среды
ОПК-8 (профессиональные компетенции)	владение знаниями о теоретических основах экологического мониторинга, нормирования и снижения загрязнения окружающей среды, техногенных систем и экологического риска, способность к использованию теоретических знаний в практической деятельности

Шифр компетенции	Расшифровка приобретаемой компетенции
ОПК-9 (профессиональные компетенции)	способность решать стандартные задачи профессиональной деятельности на основе информационной и библиографической культуры с применением информационно-коммуникационных технологий и с учетом основных требований информационной безопасности
ПК-10 (профессиональные компетенции)	способность осуществлять контрольно-ревизионную деятельность, экологический аудит, экологическое нормирование, разработку профилактических мероприятий по защите здоровья населения от негативных воздействий хозяйственной деятельности, проводить рекультивацию техногенных ландшафтов, знать принципы оптимизации среды обитания
ПК-14 (профессиональные компетенции)	владение знаниями об основах землеведения, климатологии, гидрологии, ландшафтоведения, социально-экономической географии и картографии
ПК-16 (профессиональные компетенции)	владение знаниями в области общего ресурсоведения, регионального природопользования, картографии
ПК-17 (профессиональные компетенции)	способность решать глобальные и региональные геологические проблемы
ПК-3 (профессиональные компетенции)	владение навыками эксплуатации очистных установок, очистных сооружений и полигонов и других производственных комплексов в области охраны окружающей среды и снижения уровня негативного воздействия хозяйственной деятельности
ПК-4 (профессиональные компетенции)	способность прогнозировать техногенные катастрофы и их последствия, планировать мероприятия по профилактике и ликвидации последствий экологических катастроф, принимать профилактические меры для снижения уровня опасностей различного вида и их последствий
ПК-5 (профессиональные компетенции)	способность реализовывать технологические процессы по переработке, утилизации и захоронению твердых и жидких отходов; организовывать производство работ по рекультивации нарушенных земель, по восстановлению нарушенных агросистем и созданию культурных ландшафтов
ПК-8 (профессиональные компетенции)	владение знаниями теоретических основ экологического мониторинга, экологической экспертизы, экологического менеджмента и аудита, нормирования и снижения загрязнения окружающей среды, основы техногенных систем и экологического риска

В результате освоения дисциплины студент:

1. должен знать:

В результате освоения дисциплины обучающийся должен знать :Основы методов машинного обучения.

2. должен уметь:

Применять методы машинного обучения для решения практических задач.

3. должен владеть:

Соответствующими приемами программирования в статистической системе R.

4. Структура и содержание дисциплины/ модуля

Общая трудоемкость дисциплины составляет 3 зачетных(ые) единиц(ы) 108 часа(ов).

Форма промежуточного контроля дисциплины зачет в 7 семестре.

Суммарно по дисциплине можно получить 100 баллов, из них текущая работа оценивается в 50 баллов, итоговая форма контроля - в 50 баллов. Минимальное количество для допуска к зачету 28 баллов.

86 баллов и более - "отлично" (отл.);

71-85 баллов - "хорошо" (хор.);

55-70 баллов - "удовлетворительно" (удов.);

54 балла и менее - "неудовлетворительно" (неуд.).

4.1 Структура и содержание аудиторной работы по дисциплине/ модулю

Тематический план дисциплины/модуля

N	Раздел Дисциплины/ Модуля	Семестр	Неделя семестра	Виды и часы аудиторной работы, их трудоемкость (в часах)			Текущие формы контроля
				Лекции	Практические занятия	Лабораторные работы	
1.	Тема 1. Задачи обучения по прецедентам (примерам)	7	1-2	2	6	0	Лабораторные работы Устный опрос
2.	Тема 2. Байесовские методы обучения.	7	3-5	2	8	0	Устный опрос Контрольная работа
3.	Тема 3. Метрические методы обучения	7	6-8	4	10	0	Контрольная работа Устный опрос
4.	Тема 4. Линейные методы обучения	7	9-11	4	10	0	Контрольная работа Устный опрос
5.	Тема 5. Методы восстановления зависимостей	7	12-16	6	12	0	Устный опрос Контрольная работа
.	Тема . Итоговая форма контроля	7		0	0	0	Зачет
	Итого			18	46	0	

4.2 Содержание дисциплины

Тема 1. Задачи обучения по прецедентам (примерам)

лекционное занятие (2 часа(ов)):

Повторение основных понятий статистики. Характеристики распределения, плотность вероятности и функция распределения. Квантили. Функция правдоподобия. Основные возможности статистической системы R, чтение и запись данных, графическое представление. Объекты и признаки, степень измеримости, кодирование (контрасты). Основные понятия и определения машинного обучения. Типы задач, модель алгоритмов и метод обучения, функционал качества. Вероятностная постановка задачи обучения, проблема переобучения и понятие обобщающей способности, примеры прикладных задач: задачи классификации, задачи восстановления зависимостей, задачи ранжирования, задачи кластеризации, задачи поиска ассоциаций. Методология тестирования обучаемых алгоритмов, приёмы генерации модельных данных

практическое занятие (6 часа(ов)):

Выполнение стандартных манипуляций с данными. Построение графиков. Вычисление выборочных статистик. Загрузка и подключение пакетов. Получение справок по функциям и пакетам. Формулы в языке R и их использование. Модельная матрица. Генерация выборок из заданного распределения. Проверка гипотез для полученных выборок. Сравнение распределений, квантильные графики. Сортировка, ранги, индексы.

Тема 2. Байесовские методы обучения.

лекционное занятие (2 часа(ов)):

Вероятностная постановка задачи классификации: функционал среднего риска, оптимальное байесовское решающее правило. Задача восстановления плотности распределения. Непараметрическая классификация, непараметрические оценки плотности, метод окна Парзена. Нормальный дискриминантный анализ: многомерное нормальное распределение, квадратичный дискриминант, линейный дискриминант Фишера. Разделение смеси распределений: EM-алгоритм, смеси многомерных нормальных распределений, сети радиальных базисных функций

практическое занятие (8 часа(ов)):

Плотность вероятности многомерного нормального распределения. Построение плотности смеси распределений. Решение задач на оптимальное байесовское решающее правило. Решение задачи восстановления плотности распределения. Классификация с использованием непараметрической оценки плотности, метод окна Парзена. Сравнение полученной модели с тем распределением, из которого получена модельная выборка. Применение нормальный дискриминантный анализа: вычисление ковариационных матриц, вычисление решающего линейного правила. Использование фнкций lda(), qda(). Вычисление линейного дискриминанта Фишера (robCompositions::daFisher), сравнение результатов с нормальным линейным дискриминантом. Программирование алгоритма разделения смеси нормальных распределений. Использование функций из пакета mixtools. Использование сети радиальных базисных функций из пакета RSNNS.

Тема 3. Метрические методы обучения

лекционное занятие (4 часа(ов)):

Метод ближайшего соседа и его обобщения: обобщённый метрический классификатор, метод ближайших соседей и его вариации. Построение регрессии на основе метода ближайшего соседа. Отбор эталонных объектов: понятие отступа объекта, алгоритм STOLP для отбора эталонных объектов Метода топографического отображения, методы самоорганизации (обучение без учителя). Нейронные сети Кохонена.

практическое занятие (10 часа(ов)):

Решение задач методом ближайшего соседа с использованием пакетов class, kknn. Программирование регрессии на основе метода ближайшего соседа. Программирование отбора эталонных объектов с использованием пакета kknn Классификация + ординация классов. Использование нейронных сетей Кохонена из пакета Kohonen для разведочного анализа данных.

Тема 4. Линейные методы обучения

лекционное занятие (4 часа(ов)):

Аппроксимация и регуляризация эмпирического риска. Линейная модель классификации. Метод стохастического градиента: классические частные случаи, эвристики для улучшения градиентных методов обучения. Логистическая регрессия: обоснование логистической регрессии, метод стохастического градиента для логистической регрессии, скоринг и оценивание апостериорных вероятностей. Метод опорных векторов: линейно разделимая выборка, линейно неразделимая выборка, ядра и спрямляющие пространства. ROC-кривая и оптимизация порога решающего правила.

практическое занятие (10 часа(ов)):

Программирование метода стохастического градиента для задачи наименьших квадратов. Использование логистической регрессии, функция glm(). Связь логистической регрессии и отношения шансов. Интерпретация результатов. Оценивание апостериорных вероятностей. Функция стоимости для рисков, и оптимизация порога решающего правила. Использование ROC-кривой для сравнения качества классификаторов.

Тема 5. Методы восстановления зависимостей

лекционное занятие (6 часа(ов)):

Метод наименьших квадратов. Непараметрическая регрессия: ядерное сглаживание: формула Надарая?Ватсона, выбор ядра и ширины окна. Проблема выбросов: робастная непараметрическая регрессия, проблема краевых эффектов. Неквадратичные функции потерь, итерационный взвешенный МНК. LASSO Тибширани Линейная регрессия для некорректных задач: проекционные методы, сингулярное разложение, проблема мультиколлинеарности, гребневая регрессия. Линейная монотонная регрессия. Методы снижения размерности (метод главных компонент). Методы проекции на латентные структуры. Нелинейные методы восстановления регрессии: нелинейная модель регрессии, нелинейные одномерные преобразования признаков, обобщённые линейные модели. Выбор оптимального уровня нелинейности ? обобщенная перекрестная проверка.

практическое занятие (12 часа(ов)):

Программирование метода наименьших квадратов с использованием оптимизатора общего вида optim(). Сравнение разных методов оптимизации, параметры управления. Программирование робастной непараметрической регрессии ? функции loess(), lowess(). Неквадратичные функции потерь, итерационный взвешенный МНК. LASSO Тибширани ? использование функций из пакетов car, MASS, stats, ElemStatLearn, chemometrics, Программирование метода регуляризации Тихонова для некорректных задач, выбор оптимального параметра регуляризации. Программирование гребневой регрессии с функцией lm.ridge(). Линейная монотонная регрессия, использование функции isoreg(). Методы снижения размерности (метод главных компонент, функции prcomp(), proincomp()). Методы проекции на латентные структуры, функция pls(). Нелинейные методы восстановления регрессии, пакеты gam, mgcv. Построение регрессии, интепретация результатов.

4.3 Структура и содержание самостоятельной работы дисциплины (модуля)

N	Раздел Дисциплины	Семестр	Неделя семестра	Виды самостоятельной работы студентов	Трудоемкость (в часах)	Формы контроля самостоятельной работы
1.	Тема 1. Задачи обучения по прецедентам (примерам)	7	1-2	подготовка к устному опросу	8	устный опрос
2.	Тема 2. Байесовские методы обучения.	7	3-5	подготовка к контрольной работе	10	контрольная работа
3.	Тема 3. Метрические методы обучения	7	6-8	подготовка к контрольной работе	10	контрольная работа

N	Раздел Дисциплины	Семестр	Неделя семестра	Виды самостоятельной работы студентов	Трудоемкость (в часах)	Формы контроля самостоятельной работы
4.	Тема 4. Линейные методы обучения	7	9-11	подготовка к контрольной работе	10	контрольная работа
5.	Тема 5. Методы восстановления зависимостей	7	12-16	подготовка к контрольной работе	6	контрольная работа
	Итого				44	

5. Образовательные технологии, включая интерактивные формы обучения

Проводятся лекции и лабораторные занятия с использованием компьютеров с применением специализированного программного обеспечения. Часть материала изучается самостоятельно.

6. Оценочные средства для текущего контроля успеваемости, промежуточной аттестации по итогам освоения дисциплины и учебно-методическое обеспечение самостоятельной работы студентов

Тема 1. Задачи обучения по прецедентам (примерам)

устный опрос , примерные вопросы:

Опрос по темам лекции: Основные понятия статистики, характеристики распределения, плотность вероятности и функция распределения. Объекты и признаки, степень измеримости, кодирование (контрасты).

Тема 2. Байесовские методы обучения.

контрольная работа , примерные вопросы:

Повторить элементы теории вероятностей, случайные величины, их распределение, теорема Байеса, вычисление правдоподобия.Прочитать документацию по языку R, установить статистическую систему R на домашнем компьютере, установить необходимые пакеты. Выполнить элементарные манипуляции с данными: 1. Считать набор данных 2.Вычислить выборочные статистики 3.Построить график

Тема 3. Метрические методы обучения

контрольная работа , примерные вопросы:

Прочитать документацию к пакетам программ, используемых для решения задач. Использовать методы вычисления расстояний между векторами в R. Пример варианта контрольной: Даны наборы данных: Выполните классификацию и ординацию, используя пакеты class, kknn.

Тема 4. Линейные методы обучения

контрольная работа , примерные вопросы:

Повторить описание методов наименьших квадратов и максимального правдоподобия. Прочитать документацию к пакетам программ, используемых для решения задач. Пример варианта контрольной: Даны наборы данных : необходимо выполнить программирование метода стохастического градиента для задачи наименьших квадратов. Построить логистическую регрессию (функция glm()). Оценить апостериорные вероятности.

Тема 5. Методы восстановления зависимостей

контрольная работа , примерные вопросы:

Прочитать документацию по пакетам, используемым для решения задач. Пример варианта контрольной: Для наборов данных выполнить программирование метода наименьших квадратов с использованием оптимизатора общего вида optim(). Сравнить разные методы оптимизации. Построить робастную непараметрическую регрессию - loess(), lowess(). Интепретировать результат.

Тема . Итоговая форма контроля

Примерные вопросы к зачету:

Вопросы на зачет:

1. Записать общую формулу байесовского классификатора.

2. Какие вы знаете три подхода к восстановлению плотности распределения по выборке?

3. Что такое наивный байесовский классификатор?

4. Что такое оценка плотности Парзена-Розенблатта.

5. На что влияет ширина окна, а на что вид ядра в методе парзеновского окна?

6. Многомерное нормальное распределение. Формула квадратичного дискриминанта. При каком условии он становится линейным?

7. На каких предположениях осован линейный дискриминант Фишера?

8. Что такое "проблема мультиколлинеарности", в каких задачах и при использовании каких алгоритмов она возникает? Какие есть подходы к её решению?

9. Что такое "смесь распределений"?

10. Что такое EM-алгоритм, какова его основная идея? Какая задача решается на E-шаге, на M-шаге? Каков вероятностный смысл скрытых переменных?

11. Что такое стохастический EM-алгоритм, какова основная идея? В чём его преимущество (какой недостаток стандартного EM-алгоритма он устраняет)?

12. Что такое сеть радиальных базисных функций?

13. Что такое "выбросы"? Как осуществляется фильтрация выбросов?

14. Что такое метод потенциальных функций? Идея алгоритма настройки. Сравните с методом радиальных базисных функций.

15. Зачем нужен отбор опорных объектов в метрических алгоритмах классификации?

16. Метод стохастического градиента.

17. Обоснование логистической регрессии, основные посылки и следствия. Как выражается апостериорная вероятность классов.

18. Две мотивации и постановка задачи метода опорных векторов. Постановка задачи SVM.

19. Что такое ядро в SVM? Зачем вводятся ядра? Любая ли функция может быть ядром?

20. Что такое ROC-кривая, как она определяется? Как она эффективно вычисляется?

21. В каких алгоритмах классификации можно узнать не только классовую принадлежность классифицируемого объекта, но и вероятность того, что данный объект принадлежит каждому из классов?

22. Каков вероятностный смысл регуляризации? Какие типы регуляризаторов Вы знаете?

23. Что есть общего между ядром в непараметрической регрессии и ядром SVM?

24. На что влияет ширина окна, а на что вид ядра в непараметрической регрессии?

25. Постановка задачи многомерной линейной регрессии.

26. Что такое сингулярное разложение? Как оно используется для решения задачи наименьших квадратов?

27. Что такое "проблема мультиколлинеарности" в задачах многомерной линейной регрессии? Какие есть три подхода к её устранению?

28. Сравнить гребневую регрессию и лассо. В каких задачах предпочтительнее использовать лассо?

29. Какую проблему решает метод главных компонент в многомерной линейной регрессии?

30. Метод настройки с возвращениями (backfitting): постановка задачи и основная идея метода.

31. Какие методы построения логистической регрессии Вы знаете?

7.1. Основная литература:

1.Нейронные сети: основы теории [Электронный ресурс] / Галушкин А.И. - М. : Горячая линия - Телеком, 2012. - http://www.studentlibrary.ru/book/ISBN9785991200820.html

2. Кохонен Т., Самоорганизующиеся карты [Электронный ресурс] : учеб. пособие ? Электрон. дан. ? Москва : Издательство 'Лаборатория знаний', 2017. ? 660 с. ? Режим доступа: https://e.lanbook.com/book/94143. ? Загл. с экрана.

7.2. Дополнительная литература:

1.Нейронные сети : полный курс : перевод с английского / С. Хайкин ; Пер. под ред. Н. Н. Куссуль; Пер. А. Ю. Шелестова .-Издание 2-е, исправленнное .- Москва ; Санкт-Петербург ; Киев : Вильямс, 2008 .- 1104 с .- ISBN 978-5-8459-0890-2.

2. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных [Электронный ресурс] / Флах П. - М. : ДМК Пресс, 2015. - http://www.studentlibrary.ru/book/ISBN9785970602737.html

7.3. Интернет-ресурсы:

Machine Learning for Developers - https://xyclade.github.io/MachineLearning/

Видео лекции по Машинному обучению - https://yandexdataschool.ru/edu-process/courses/machine-learning

Интересные публикации - https://habrahabr.ru/hub/machine_learning/

Машинное обучение с примерами - http://www.uic.unn.ru/~zny/ml/Lectures/

Сайт ?Машинное обучение? - http://www.machinelearning.ru/

8. Материально-техническое обеспечение дисциплины(модуля)

Освоение дисциплины "Методы машинного обучения" предполагает использование следующего материально-технического обеспечения:

Компьютерный класс, представляющий собой рабочее место преподавателя и не менее 15 рабочих мест студентов, включающих компьютерный стол, стул, персональный компьютер, лицензионное программное обеспечение. Каждый компьютер имеет широкополосный доступ в сеть Интернет. Все компьютеры подключены к корпоративной компьютерной сети КФУ и находятся в едином домене.

Компьютерный класс

Программа составлена в соответствии с требованиями ФГОС ВПО и учебным планом по направлению 05.03.06 "Экология и природопользование" .

Автор(ы):
Савельев А.А. _____
"__" _____ 201 __ г.

Рецензент(ы):
Зарипов Ш.Х. _____
"__" _____ 201 __ г.