

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования
"Казанский (Приволжский) федеральный университет"
Высшая школа информационных технологий и интеллектуальных систем



УТВЕРЖДАЮ
Проректор по образовательной деятельности КФУ
Проф. Д.А. Таюрский

» 20__ г.

подписано электронно-цифровой подписью

Программа дисциплины

Анализ данных. Дополнительные главы

Направление подготовки: 09.03.04 - Программная инженерия

Профиль подготовки: Технологии разработки информационных систем

Квалификация выпускника: бакалавр

Форма обучения: очное

Язык обучения: русский

Год начала обучения по образовательной программе: 2017

Содержание

1. Перечень планируемых результатов обучения по дисциплине (модулю), соотнесенных с планируемыми результатами освоения ОПОП ВО
2. Место дисциплины (модуля) в структуре ОПОП ВО
3. Объем дисциплины (модуля) в зачетных единицах с указанием количества часов, выделенных на контактную работу обучающихся с преподавателем (по видам учебных занятий) и на самостоятельную работу обучающихся
4. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий
 - 4.1. Структура и тематический план контактной и самостоятельной работы по дисциплине (модулю)
 - 4.2. Содержание дисциплины (модуля)
5. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине (модулю)
6. Фонд оценочных средств по дисциплине (модулю)
7. Перечень литературы, необходимой для освоения дисциплины (модуля)
8. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)
9. Методические указания для обучающихся по освоению дисциплины (модуля)
10. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень программного обеспечения и информационных справочных систем (при необходимости)
11. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)
12. Средства адаптации преподавания дисциплины (модуля) к потребностям обучающихся инвалидов и лиц с ограниченными возможностями здоровья
13. Приложение №1. Фонд оценочных средств
14. Приложение №2. Перечень литературы, необходимой для освоения дисциплины (модуля)
15. Приложение №3. Перечень информационных технологий, используемых для освоения дисциплины (модуля), включая перечень программного обеспечения и информационных справочных систем

Программу дисциплины разработал(а)(и) доцент, к.н. (доцент) Голицына И.Н. (Кафедра программной инженерии, Высшая школа информационных технологий и интеллектуальных систем), Irina.Golicyna@kpfu.ru Новиков П.А.

1. Перечень планируемых результатов обучения по дисциплине (модулю), соотнесенных с планируемыми результатами освоения ОПОП ВО

Обучающийся, освоивший дисциплину (модуль), должен обладать следующими компетенциями:

Шифр компетенции	Расшифровка приобретаемой компетенции
ПК-12	способность к формализации в своей предметной области с учетом ограничений используемых методов исследования

Обучающийся, освоивший дисциплину (модуль):

Должен демонстрировать способность и готовность:

Знать возможности и сферы применения анализа данных.

Знать теоретические основы методов анализа данных.

Уметь реализовывать изучаемые методы на алгоритмическом языке программирования.

Уметь обрабатывать реальные данные с помощью изучаемых методов.

Понимать сложности аналитической обработки больших массивов данных и знать подходы к их преодолению.

Уметь ориентироваться в профессиональной литературе по анализу данных и машинному обучению.

Уметь представлять результаты анализа данных в надлежащем виде.

2. Место дисциплины (модуля) в структуре ОПОП ВО

Данная дисциплина (модуль) включена в раздел "Б1.В.ДВ.2 Дисциплины (модули)" основной профессиональной образовательной программы 09.03.04 "Программная инженерия (Технологии разработки информационных систем)" и относится к дисциплинам по выбору.

Осваивается на 3 курсе в 6 семестре.

3. Объем дисциплины (модуля) в зачетных единицах с указанием количества часов, выделенных на контактную работу обучающихся с преподавателем (по видам учебных занятий) и на самостоятельную работу обучающихся

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

Контактная работа - 72 часа(ов), в том числе лекции - 36 часа(ов), практические занятия - 36 часа(ов), лабораторные работы - 0 часа(ов), контроль самостоятельной работы - 0 часа(ов).

Самостоятельная работа - 36 часа(ов).

Контроль (зачёт / экзамен) - 0 часа(ов).

Форма промежуточного контроля дисциплины: зачет в 6 семестре.

4. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

4.1 Структура и тематический план контактной и самостоятельной работы по дисциплине (модулю)

N	Разделы дисциплины / модуля	Семестр	Виды и часы контактной работы, их трудоемкость (в часах)			Самостоятельная работа
			Лекции	Практические занятия	Лабораторные работы	
1.	Тема 1. Кластеризация.	6	9	9	0	9
2.	Тема 2. Понижение размерности.	6	9	9	0	9
3.	Тема 3. Рекомендательные системы.	6	9	9	0	9
4.	Тема 4. Анализ графов социальных сетей.	6	9	9	0	9
	Итого		36	36	0	36

4.2 Содержание дисциплины (модуля)

Тема 1. Кластеризация.

Точки, пространства и расстояния. Стратегии кластеризации. "Проклятие размерности". Иерархическая кластеризация. Иерархическая кластеризация в евклидовом пространстве. Эффективность иерархической кластеризации. Альтернативные правила управления иерархической кластеризацией. Иерархическая кластеризация в неевклидовых пространствах. Алгоритм k средних. Основы алгоритма k средних. Инициализация кластеров в алгоритме k средних. Выбор правильного значения k. Алгоритм Брэдли-Файяда-Рейна (BFR). Обработка данных в алгоритме BFR. Алгоритм CURE. Этап инициализации в CURE. Завершение работы алгоритма CURE. Кластеризация в неевклидовых пространствах. Представление кластеров в алгоритме GRGPF. Инициализация дерева кластеров. Добавление точек в алгоритме GRGPF. Разделение и объединение кластеров. Кластеризация для потоков и параллелизм. Модель потоковых вычислений. Алгоритм кластеризации потока. Инициализация интервалов. Объединение кластеров. Кластеризация в параллельной среде.

Тема 2. Понижение размерности.

Собственные значения и собственные векторы. Вычисление собственных значений и собственных векторов. Нахождение собственных пары степенным методом. Матрица собственных векторов. Метод главных компонент. Использование собственных векторов для понижения размерности. Матрица расстояний. Сингулярное разложение. Определение сингулярного разложения. Интерпретация сингулярного разложения. Понижение размерности с помощью сингулярного разложения. Запросы с использованием концептов. Вычисление сингулярного разложения матрицы. CUR-декомпозиция. Определение CUR-декомпозиции. Правильный выбор строк и столбцов. Построение средней матрицы. Полная CUR-декомпозиция. Исключение дубликатов строк и столбцов.

Тема 3. Рекомендательные системы.

Модель рекомендательной системы. Матрица предпочтений. Длинный хвост. Применения рекомендательных систем. Заполнение матрицы предпочтений. Рекомендации на основе фильтрации содержимого. Профили объектов. Выявление признаков документа. Получение признаков объектов из меток. Представление профиля объекта. Профили пользователей. Рекомендованные объекты пользователям на основе содержимого. Алгоритм классификации. Коллаборативная фильтрация. Измерение сходства. Двойственность сходства. Кластеризация пользователей и объектов. Понижение размерности. UV-декомпозиция. Среднеквадратичная ошибка. Инкрементное вычисление UV-декомпозиции. Оптимизация произвольного элемента. Построение полного алгоритма UV-декомпозиции. Задача Netflix.

Тема 4. Анализ графов социальных сетей.

Социальные сети. Социальные сети как графы. Разновидности социальных сетей. Графы с вершинами нескольких типов. Кластеризация графа социальной сети. Метрики для графов социальных сетей. Применение стандартных методов кластеризации. Промежуточность. Алгоритм Гирвана-Ньюмана. Использование промежуточности для нахождения сообществ. Прямое нахождение сообществ. Нахождение клика. Полные двудольные графы. Нахождение полных двудольных подграфов. Существование полных двудольных графов. Разрезание графов. Нормализованные разрезы. Некоторые матрицы, описывающие графы. Собственные значения матрицы Лапласа. Нахождение пересекающихся сообществ. Природа сообществ. Оценка максимального правдоподобия. Модель графа принадлежности. Случайные блуждания в социальном графе. Случайное блуждание с перезапуском. Подсчет треугольников. Алгоритм нахождения треугольников. Оптимальность алгоритма нахождения треугольников. Нахождение треугольников с помощью MapReduce. Окрестности в графах. Ориентированные графы и окрестности. Диаметр графа. Транзитивное замыкание и достижимость. Вычисление транзитивного замыкания с помощью MapReduce. Интеллектуальное транзитивное замыкание. Транзитивное замыкание посредством сокращения графа. Аппроксимация размеров окрестностей.

5. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине (модулю)

Самостоятельная работа обучающихся выполняется по заданию и при методическом руководстве преподавателя, но без его непосредственного участия. Самостоятельная работа подразделяется на самостоятельную работу на аудиторных занятиях и на внеаудиторную самостоятельную работу. Самостоятельная работа обучающихся включает как полностью самостоятельное освоение отдельных тем (разделов) дисциплины, так и проработку тем (разделов), осваиваемых во время аудиторной работы. Во время самостоятельной работы обучающиеся читают и конспектируют учебную, научную и справочную литературу, выполняют задания, направленные на закрепление знаний и отработку умений и навыков, готовятся к текущему и промежуточному контролю по дисциплине.

Организация самостоятельной работы обучающихся регламентируется нормативными документами, учебно-методической литературой и электронными образовательными ресурсами, включая:

Порядок организации и осуществления образовательной деятельности по образовательным программам высшего образования - программам бакалавриата, программам специалитета, программам магистратуры (утвержден приказом Министерства образования и науки Российской Федерации от 5 апреля 2017 года №301)

Письмо Министерства образования Российской Федерации №14-55-996ин/15 от 27 ноября 2002 г. "Об активизации самостоятельной работы студентов высших учебных заведений"

Устав федерального государственного автономного образовательного учреждения "Казанский (Приволжский) федеральный университет"

Правила внутреннего распорядка федерального государственного автономного образовательного учреждения высшего профессионального образования "Казанский (Приволжский) федеральный университет"
Локальные нормативные акты Казанского (Приволжского) федерального университета

6. Фонд оценочных средств по дисциплине (модулю)

Фонд оценочных средств по дисциплине (модулю) включает оценочные материалы, направленные на проверку освоения компетенций, в том числе знаний, умений и навыков. Фонд оценочных средств включает оценочные средства текущего контроля и оценочные средства промежуточной аттестации.

В фонде оценочных средств содержится следующая информация:

- соответствие компетенций планируемым результатам обучения по дисциплине (модулю);
- критерии оценивания сформированности компетенций;
- механизм формирования оценки по дисциплине (модулю);
- описание порядка применения и процедуры оценивания для каждого оценочного средства;
- критерии оценивания для каждого оценочного средства;
- содержание оценочных средств, включая требования, предъявляемые к действиям обучающихся, демонстрируемым результатам, задания различных типов.

Фонд оценочных средств по дисциплине находится в Приложении 1 к программе дисциплины (модулю).

7. Перечень литературы, необходимой для освоения дисциплины (модуля)

Освоение дисциплины (модуля) предполагает изучение основной и дополнительной учебной литературы. Литература может быть доступна обучающимся в одном из двух вариантов (либо в обоих из них):

- в электронном виде - через электронные библиотечные системы на основании заключенных КФУ договоров с правообладателями;
- в печатном виде - в Научной библиотеке им. Н.И. Лобачевского. Обучающиеся получают учебную литературу на абонементе по читательским билетам в соответствии с правилами пользования Научной библиотекой.

Электронные издания доступны дистанционно из любой точки при введении обучающимся своего логина и пароля от личного кабинета в системе "Электронный университет". При использовании печатных изданий библиотечный фонд должен быть укомплектован ими из расчета не менее 0,5 экземпляра (для обучающихся по ФГОС 3++ - не менее 0,25 экземпляра) каждого из изданий основной литературы и не менее 0,25 экземпляра дополнительной литературы на каждого обучающегося из числа лиц, одновременно осваивающих данную дисциплину.

Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины (модуля), находится в Приложении 2 к рабочей программе дисциплины. Он подлежит обновлению при изменении условий договоров КФУ с правообладателями электронных изданий и при изменении комплектования фондов Научной библиотеки КФУ.

8. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)

MachineLearning.ru - <http://www.machinelearning.ru>

Mining of Massive Datasets - <http://mmds.org>

The R Project for Statistical Computing - <https://www.r-project.org>

9. Методические указания для обучающихся по освоению дисциплины (модуля)

На лекционных занятиях следует вести запись лекции в тетради, даже если преподаватель предоставит Вам доступ к лекционному материалу в электронном формате. Запись вручную необходима для лучшего освоения материала и лучшей ориентации в материале. Вопросы во время лекции приветствуются; чтобы задать вопрос, необходимо поднять руку и дождаться, когда преподаватель даст Вам слово. На практических занятиях следует выполнять задания, данные преподавателем, самостоятельно, а в случае возникновения сложностей консультироваться с преподавателем. Приветствуются групповые обсуждения и разбор сложных вопросов студентами на доске. При выполнении практических заданий не злоупотребляйте готовыми решениями, которые можно найти в сети Интернет, сначала следует решить задачу самостоятельно, а затем можно сравнивать с другими решениями. При подготовке к контрольной работе следует обращать внимание на математические аспекты изучаемого материала, поскольку именно они представляют наибольшую сложность. При неуверенном владении математическим аппаратом первоначально следует повторить материал из линейной алгебры и геометрии и дискретной математики, а именно, разделы, связанные с матрицами, собственными значениями и рангом матриц, расстояниями в евклидовом пространстве и теорией графов. Перед контрольной следует определиться заранее, на каком языке программирования Вам будет наиболее комфортно выполнять задания контрольной работы, связанные с программированием. При выполнении заданий контрольной работы, связанных с составлением программы, рекомендуется сначала выполнить первые несколько шагов программы вручную и лишь затем составлять собственно программу. В заданиях, связанных с программированием, не допускается использование специализированных библиотек, если не оговорено обратное. Если Вам не хватает времени на выполнение задания контрольной работы в полном объеме, то следует вкратце описать дальнейшую последовательность действий, которые необходимо сделать для завершения задания; таким образом Вы сможете получить частичные баллы. При этом следует отметить, что нехватка времени на выполнение задания, скорее всего, свидетельствует о том, что Вы выполняете задание неоптимальным способом. При подготовке к зачету следует сосредоточиться на выстраивании общей канвы, взаимосвязей и аналогий между различными частями изучаемого материала. Во время зачета будет оцениваться в первую очередь Ваше понимание общих идей, понимание взаимосвязей между разными изучаемыми методами и умение выводить конкретные результаты из более общих идей и методов. В связи с этим следует остерегаться линейного заучивания материала, которое будет весьма трудоемко и при этом всё равно не позволит подготовиться на высокий балл. При подготовке к зачету используйте, но не злоупотребляйте профессиональными форумами и блогами: подобные ресурсы зачастую позволяют найти ответ на конкретный вопрос, однако это зачастую затеняет общие подходы и не способствует формированию знаний в системе. На вопросы зачета следует давать развернутый ответ, избегайте односложных ответов -- таковые, как правило, свидетельствуют о недостаточной подготовке студента. При нечеткой формулировке вопроса следует рассмотреть возможные интерпретации смысла вопроса и изложить свои соображения преподавателю.

10. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень программного обеспечения и информационных справочных систем (при необходимости)

Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень программного обеспечения и информационных справочных систем, представлен в Приложении 3 к рабочей программе дисциплины (модуля).

11. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

Материально-техническое обеспечение образовательного процесса по дисциплине (модулю) включает в себя следующие компоненты:

Помещения для самостоятельной работы обучающихся, укомплектованные специализированной мебелью (столы и стулья) и оснащенные компьютерной техникой с возможностью подключения к сети "Интернет" и обеспечением доступа в электронную информационно-образовательную среду КФУ.

Учебные аудитории для контактной работы с преподавателем, укомплектованные специализированной мебелью (столы и стулья).

Компьютер и принтер для распечатки раздаточных материалов.

Мультимедийная аудитория.

Компьютерный класс.

12. Средства адаптации преподавания дисциплины к потребностям обучающихся инвалидов и лиц с ограниченными возможностями здоровья

При необходимости в образовательном процессе применяются следующие методы и технологии, облегчающие восприятие информации обучающимися инвалидами и лицами с ограниченными возможностями здоровья:

- создание текстовой версии любого нетекстового контента для его возможного преобразования в альтернативные формы, удобные для различных пользователей;
- создание контента, который можно представить в различных видах без потери данных или структуры, предусмотреть возможность масштабирования текста и изображений без потери качества, предусмотреть доступность управления контентом с клавиатуры;

- создание возможностей для обучающихся воспринимать одну и ту же информацию из разных источников - например, так, чтобы лица с нарушениями слуха получали информацию визуально, с нарушениями зрения - аудиально;
- применение программных средств, обеспечивающих возможность освоения навыков и умений, формируемых дисциплиной, за счёт альтернативных способов, в том числе виртуальных лабораторий и симуляционных технологий;
- применение дистанционных образовательных технологий для передачи информации, организации различных форм интерактивной контактной работы обучающегося с преподавателем, в том числе вебинаров, которые могут быть использованы для проведения виртуальных лекций с возможностью взаимодействия всех участников дистанционного обучения, проведения семинаров, выступления с докладами и защиты выполненных работ, проведения тренингов, организации коллективной работы;
- применение дистанционных образовательных технологий для организации форм текущего и промежуточного контроля;
- увеличение продолжительности сдачи обучающимся инвалидом или лицом с ограниченными возможностями здоровья форм промежуточной аттестации по отношению к установленной продолжительности их сдачи:
- продолжительности сдачи зачёта или экзамена, проводимого в письменной форме, - не более чем на 90 минут;
- продолжительности подготовки обучающегося к ответу на зачёте или экзамене, проводимом в устной форме, - не более чем на 20 минут;
- продолжительности выступления обучающегося при защите курсовой работы - не более чем на 15 минут.

Программа составлена в соответствии с требованиями ФГОС ВО и учебным планом по направлению 09.03.04 "Программная инженерия" и профилю подготовки "Технологии разработки информационных систем".

Приложение 2
к рабочей программе дисциплины (модуля)
Б1.В.ДВ.2 Анализ данных. Дополнительные главы

Перечень литературы, необходимой для освоения дисциплины (модуля)

Направление подготовки: 09.03.04 - Программная инженерия

Профиль подготовки: Технологии разработки информационных систем

Квалификация выпускника: бакалавр

Форма обучения: очное

Язык обучения: русский

Год начала обучения по образовательной программе: 2017

Основная литература:

1. Буховец, А. Г. Алгоритмы вычислительной статистики в системе R : учебное пособие / А. Г. Буховец, П. В. Москалев. - 2-е изд., перераб. и доп. - Санкт-Петербург : Лань, 2015. - 160 с. - ISBN 978-5-8114-1802-2. - Текст : электронный // Лань : электронно-библиотечная система. - URL: <https://e.lanbook.com/book/68459> (дата обращения: 26.02.2020). - Режим доступа: для авториз. пользователей.
2. Гинис Л.А., Статистические методы контроля и управления качеством. Прикладные программные средства : учебное пособие / Гинис Л. А. - Ростов-на : Изд-во ЮФУ, 2017. - 81 с. - ISBN 978-5-9275-2619-2 - Текст : электронный // ЭБС 'Консультант студента' : [сайт]. - URL : <https://www.studentlibrary.ru/book/ISBN9785927526192.html> (дата обращения: 26.02.2020). - Режим доступа : по подписке.
3. Статистика : учебник / В.В. Глинский, В.Г. Ионин, Л.К. Серга [и др.] ; под ред. В.Г. Ионина. - 4-е изд., перераб. и доп. - Москва : ИНФРА-М, 2017. - 355 с. - (Высшее образование: Бакалавриат). - www.dx.doi.org/10.12737/25127. - ISBN 978-5-16-105518-2. - Текст : электронный. - URL: <https://znanium.com/catalog/product/552459> (дата обращения: 26.02.2020). - Режим доступа : по подписке.

Дополнительная литература:

1. Теория вероятностей и математическая статистика: учебное пособие / Бирюкова Л.Г., Бобрик Г.И., Матвеев В.И., - 2-е изд. - Москва: НИЦ ИНФРА-М, 2017. - 289 с. (Высшее образование: Бакалавриат) ISBN 978-5-16-011793-5. - Текст : электронный. - URL: <https://znanium.com/catalog/product/370899> (дата обращения: 26.02.2020). - Режим доступа : по подписке.
2. Белько, И. В. Теория вероятностей, математическая статистика, математическое программирование: учебное пособие / Белько И.В., Морозова И.М., Криштапович Е.А. - Москва: НИЦ ИНФРА-М, Нов. знание, 2016. - 299 с. (Высшее образование: Бакалавриат) ISBN 978-5-16-011748-5. - Текст : электронный. - URL: <https://znanium.com/catalog/product/542521> (дата обращения: 26.02.2020). - Режим доступа : по подписке.
3. Колдаев, В. Д. Численные методы и программирование : учебное пособие / В.Д. Колдаев ; под ред. проф. Л.Г. Гагариной. - Москва : ИД 'ФОРУМ' : ИНФРА-М, 2017. - 336 с. : ил. - (Профессиональное образование). - ISBN 978-5-16-101025-9. - Текст : электронный. - URL: <https://znanium.com/catalog/product/672965> (дата обращения: 26.02.2020). - Режим доступа : по подписке.

Приложение 3
к рабочей программе дисциплины (модуля)
Б1.В.ДВ.2 Анализ данных. Дополнительные главы

Перечень информационных технологий, используемых для освоения дисциплины (модуля), включая перечень программного обеспечения и информационных справочных систем

Направление подготовки: 09.03.04 - Программная инженерия

Профиль подготовки: Технологии разработки информационных систем

Квалификация выпускника: бакалавр

Форма обучения: очное

Язык обучения: русский

Год начала обучения по образовательной программе: 2017

Освоение дисциплины (модуля) предполагает использование следующего программного обеспечения и информационно-справочных систем:

Операционная система Microsoft Windows 7 Профессиональная или Windows XP (Volume License)

Пакет офисного программного обеспечения Microsoft Office 365 или Microsoft Office Professional plus 2010

Браузер Mozilla Firefox

Браузер Google Chrome

Adobe Reader XI или Adobe Acrobat Reader DC

Kaspersky Endpoint Security для Windows

Учебно-методическая литература для данной дисциплины имеется в наличии в электронно-библиотечной системе Издательства "Лань", доступ к которой предоставлен обучающимся. ЭБС Издательства "Лань" включает в себя электронные версии книг издательства "Лань" и других ведущих издательств учебной литературы, а также электронные версии периодических изданий по естественным, техническим и гуманитарным наукам. ЭБС Издательства "Лань" обеспечивает доступ к научной, учебной литературе и научным периодическим изданиям по максимальному количеству профильных направлений с соблюдением всех авторских и смежных прав.