МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования
"Казанский (Приволжский) федеральный университет"
Институт вычислительной математики и информационных технологий

подписано электронно-цифровой подписью

Программа дисциплины

Статистические методы анализа больших данных

Направление подготовки: 02.04.02 - Фундаментальная информатика и информационные технологии
Профиль подготовки: Машинное обучение и компьютерное зрение
Квалификация выпускника: магистр
Форма обучения: очное
Язык обучения: русский
Год начала обучения по образовательной программе: 2022

Содержание

Программу дисциплины разработал(а)(и): научный сотрудник, к.н. Заикин А.А. (НИЛ изучения состояния и эволюции подземных резервуаров, Научный центр мирового уровня Рациональное освоение запасов жидких углеводородов планеты (головной центр)), Kaskrin@gmail.com ; доцент, к.н. Салимов Р.Ф. (кафедра математической статистики, Институт математики и механики им. Н.И. Лобачевского), Rustem.Salimov@kpfu.ru

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
ОПК-1	Способен находить, формулировать и решать актуальные проблемы прикладной математики, фундаментальной информатики и информационных технологий
ОПК-3	Способен проводить анализ математических моделей, создавать инновационные методы решения прикладных задач профессиональной деятельности в области информатики и математического моделирования
УК-1	Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий

Обучающийся, освоивший дисциплину (модуль):

Должен знать:

- классы задач анализа больших данных.
- способы предварительной обработки выборочных данных;
- процедуры проведения множественного тестирования;

Должен уметь:

- обосновать применение того или иного алгоритма анализа больших данных для решения конкретной задачи.
- проверять гипотезы над данными;
- измерять по результатам эксперимента метрики качества;

Должен владеть:

- навыками применения программных инструментов для проведения анализа больших данных.
- навыками применения методов анализа больших выборочных данных.
- навыками проведения анализа метрик качества по результатам эксперимента.

Должен демонстрировать способность и готовность:

- использовать полученные навыки и знания для своей научно-исследовательской и профессиональной деятельности.

2. Место дисциплины (модуля) в структуре ОПОП ВО

Данная дисциплина (модуль) включена в раздел "Б1.О.06 Дисциплины (модули)" основной профессиональной образовательной программы 02.04.02 "Фундаментальная информатика и информационные технологии (Машинное обучение и компьютерное зрение)" и относится к обязательной части ОПОП ВО.

Осваивается на 1 курсе в 2 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 6 зачетных(ые) единиц(ы) на 216 часа(ов).

Контактная работа - 36 часа(ов), в том числе лекции - 18 часа(ов), практические занятия - 0 часа(ов), лабораторные работы - 18 часа(ов), контроль самостоятельной работы - 0 часа(ов).

Самостоятельная работа - 126 часа(ов).

Контроль (зачёт / экзамен) - 54 часа(ов).

Форма промежуточного контроля дисциплины: экзамен во 2 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, всего	Лекции в эл. форме	Практические занятия, всего	Практические в эл. форме	Лабораторные работы, всего	Лабораторные в эл. форме	
1.	Тема 1. Классификация многомерных измерений	2	3	0	0	0	3	0	25
2.	Тема 2. Кластеризация многомерных измерений	2	3	0	0	0	3	0	25
3.	Тема 3. Методы регуляризации алгоритмов оценивания	2	4	0	0	0	4	0	25
4.	Тема 4. Методы понижения размерности.	2	4	0	0	0	4	0	25
5.	Тема 5. Множественное тестирование.	2	4	0	0	0	4	0	26
	Итого		18	0	0	0	18	0	126

4.2 Содержание дисциплины (модуля)

Тема 1. Классификация многомерных измерений

Лекции:

Классификация многомерных измерений. Дискриминантные информанты и классификация. Оценка вероятностей ошибочных классификаций. Классификация на линейных дискриминантных формах.

Лабораторные работы:

Наивный байесовский подход. Линейный дискриминант Фишера. Логистическая регрессия.

Самостоятельная работа:

Проверка предположений использованных моделей классификации. Изучение вариантов таких методов.

Тема 2. Кластеризация многомерных измерений

Лекции:

Кластеризация. Выбор метрики. Метод k средних. Модель смеси распределений. Оценка смеси с помощью EM-алгоритма. Иерархическая кластеризация на основе дендрограммы.

Лабораторные работы:

Применение методов кластеризации на данных. Сравнение различных метрик.

Самостоятельная работа:

Изучение модели смеси с бесконечным количеством компонент на основе процесса Дирихле.

Тема 3. Методы регуляризации алгоритмов оценивания

Лекции:

Необходимость регуляризации в практических задачах. Регуляризация Тихонова в применении к методу максимального правдоподобия. Методы LASSO, LAR для регрессии. Робастные функции потерь для регрессии.

Лабораторные работы:

Применение методов регуляризации для данных с различными возмущениями и сравнение этих методов.

Самостоятельная работа:

Изучение методов устранения выбросов на основе расстояния Кука, метода Граббса и расстояния Махалонобиса.

Тема 4. Методы понижения размерности.

Лекции:

Задача понижения размерности. Метод главных компонент. Факторный анализ, некоторые модели факторного анализа. Понижение размерности с помощью методов LASSO и линейного дискриминантного анализа. Методы AIC и BIC для выбора модели.

Лабораторные работы:

Понижение размерности данных с помощью метода главных компонент и линейного дискриминантного анализа для численного и категориального отклика соответственно. Сравнение с результатами на основе LASSO.

Самостоятельная работа:

Изучение какого-либо нелинейного метода понижения размерности.

Тема 5. Множественное тестирование.

Лекции:

Задача множественного тестирования. Измерение частоты ошибок первого рода. Типы методов множественного тестирования: одношаговый, шаг-вниз, шаг-вверх. Методы на основе FWER: процедура Бонферонни, процедура Холма, процедура Хохберга. Процедура Беньямини-Хохберга контроля FDR.

Лабораторные работы:

Применение процедур множественного тестирования для контроля FWER и их сравнение.

Самостоятельная работы:

Изучение процедур minP и maxT, а также процедуры Саймса для контроля FWER.

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 науки и высшего образования Российской Федерации от 6 апреля 2021 года №245)

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)

Kaggle - https://www.kaggle.com

Портал образовательных ресурсов по ИТ - http://www.intuit.ru

Профессиональный интернет-ресурс по машинному обучению - http://www.machinelearning.ru/

Хабрахабр - https://habrahabr.ru/hub/machine_learning/

Школа анализа данных Яндекс - https://yandexdataschool.ru/edu-process/courses/machine-learning

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

Вид работ	Методические рекомендации
лекции	В ходе прохождения цикла занятий лекционного типа по дисциплине обучающемуся слушателю для лучшего и полноценного усвоения осваиваемого материала и теории необходимо проявлять повышенное внимание, постоянно анализировать полученную информацию, сопоставлять её с другими разделами и дисциплинами курса.
лабораторные работы	В ходе прохождения цикла занятий лабораторного типа по дисциплине обучающемуся слушателю для лучшего и полноценного усвоения осваиваемого материала и теории необходимо усердно и с инициативным рвением выполнять все задания для выполнения на лабораторных занятиях, анализировать соответствие выполненых работ с заданием и теорией.
самостоя-тельная работа	В ходе выполнения цикла самостоятельных работ по дисциплине обучающемуся слушателю курса рекомендуется с целью лучшего и более полного усвоения осваиваемого материала и теории выполнять все работы для домашнего исполнения, изучать дополнительную литературу, формулировать вопросы на не полностью освоенные части курса.
экзамен	В ходе подготовки к экзамену по дисциплине обучающемуся слушателю курса рекомендуется с целью повышения его возможностей по успешному прохождению экзамена повторить весь ранее изученный материал, как теоретического характера, так и практические и самостоятельные работы, определить возможные проблемные места усвоения материала и провести дополнительные образовательные действия для для разрешения выявленных ранее проблемных и неосвоенных участков курса.

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 02.04.02 "Фундаментальная информатика и информационные технологии" и магистерской программе "Машинное обучение и компьютерное зрение".

Приложение 2
к рабочей программе дисциплины (модуля)
Б1.О.06 Статистические методы анализа больших данных

Перечень литературы, необходимой для освоения дисциплины (модуля)

Направление подготовки: 02.04.02 - Фундаментальная информатика и информационные технологии
Профиль подготовки: Машинное обучение и компьютерное зрение
Квалификация выпускника: магистр
Форма обучения: очное
Язык обучения: русский
Год начала обучения по образовательной программе: 2022

Основная литература:

1. Статистические методы анализа данных : учебник / Л.И. Ниворожкина, С.В. Арженовский, А.А. Рудяга [и др.] ; под общ. ред. д-ра экон. наук, проф. Л.И. Ниворожкиной. - Москва : РИОР : ИНФРА-М, 2016. - 333 с. - (Высшее образование: Бакалавриат). - www.dx.doi.org/10.12737/21064. - ISBN 978-5-369-01612-1. - Текст : электронный. - URL: https://znanium.com/catalog/product/556760 (дата обращения: 21.01.2022). - Режим доступа: по подписке.

2. Мельниченко, А. С. Математическая статистика и анализ данных: учебное пособие / А. С. Мельниченко. - Москва: МИСИС, 2018. - 45 с. - ISBN 978-5-906953-62-9. - Текст : электронный // Лань : электронно-библиотечная система. - URL: https://e.lanbook.com/book/108035 (дата обращения: 21.01.2022). - Режим доступа: для авториз. пользователей.

3. Немирко, А. П. Математический анализ биомедицинских сигналов и данных / А. П. Немирко, Л. А. Манило, А. Н. Калиниченко. - Москва: ФИЗМАТЛИТ, 2017. - 248 с. - ISBN 978-5-9221-1720-3. - Текст: электронный // Лань : электронно-библиотечная система. - URL: https://e.lanbook.com/book/104986 (дата обращения: 21.01.2022). - Режим доступа: для авториз. пользователей.

4. Макшанов, А. В. Технологии интеллектуального анализа данных : учебное пособие / А. В. Макшанов, А. Е. Журавлев. - 2-е изд., стер. - Санкт-Петербург : Лань, 2022. - 212 с. - ISBN 978-5-8114-4493-9. - Текст : электронный // Лань : электронно-библиотечная система. - URL: https://e.lanbook.com/book/206711 (дата обращения: 21.01.2022). - Режим доступа: для авториз. пользователей.

Дополнительная литература:

1. Крянев, А. В. Метрический анализ и обработка данных / А. В. Крянев, Г. В. Лукин, Д. К. Удумян. - Москва : ФИЗМАТЛИТ, 2012. - 308 с. - ISBN 978-5-9221-1068-6. - Текст : электронный // Лань : электронно-библиотечная система. - URL: https://e.lanbook.com/book/59523 (дата обращения: 21.01.2022). - Режим доступа: для авториз. пользователей.

2. Туганбаев, А. А. Теория вероятностей и математическая статистика : учебное пособие / А. А. Туганбаев, В. Г. Крупин. - Санкт-Петербург : Лань, 2022. - 320 с. - ISBN 978-5-8114-1079-8. - Текст : электронный // Лань : электронно-библиотечная система. - URL: https://e.lanbook.com/book/210536 (дата обращения: 21.01.2022). - Режим доступа: для авториз. пользователей.

3. 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, 2022. - 160 с. - ISBN 978-5-8114-1802-2. - Текст : электронный // Лань : электронно-библиотечная система. - URL: https://e.lanbook.com/book/212195 (дата обращения: 21.01.2022). - Режим доступа: для авториз. пользователей.

Приложение 3
к рабочей программе дисциплины (модуля)
Б1.О.06 Статистические методы анализа больших данных

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

Направление подготовки: 02.04.02 - Фундаментальная информатика и информационные технологии

Профиль подготовки: Машинное обучение и компьютерное зрение

Квалификация выпускника: магистр

Форма обучения: очное

Язык обучения: русский

Год начала обучения по образовательной программе: 2022

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 "ZNANIUM.COM", доступ к которой предоставлен обучающимся. ЭБС "ZNANIUM.COM" содержит произведения крупнейших российских учёных, руководителей государственных органов, преподавателей ведущих вузов страны, высококвалифицированных специалистов в различных сферах бизнеса. Фонд библиотеки сформирован с учетом всех изменений образовательных стандартов и включает учебники, учебные пособия, учебно-методические комплексы, монографии, авторефераты, диссертации, энциклопедии, словари и справочники, законодательно-нормативные документы, специальные периодические издания и издания, выпускаемые издательствами вузов. В настоящее время ЭБС ZNANIUM.COM соответствует всем требованиям федеральных государственных образовательных стандартов высшего образования (ФГОС ВО) нового поколения.

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.