Белорусский государственный университет

УТВЕРЖДАЮ
Проректор по учебной работе и образовательным инновациям
_____ О.Н. Здрок
«__» _____ 20__г.

Регистрационный № УД-_____/уч.

ПРИКЛАДНОЙ АНАЛИЗ ДАННЫХ

Учебная программа учреждения высшего образования по учебной дисциплине для специальности:

1-31 80 07 Радиофизика

Профилизации:
Квантовая радиофизика
Компьютерная безопасность
Радиофизика и электроника

2020 г.

Учебная программа составлена на основе ОСВО 1-31 80 07 – 2019 и учебных планов № G31-043/ уч.; G31-044/ уч.; G31-045/ уч. от 11.04.2019 г.

СОСТАВИТЕЛИ:

Николай Николаевич Яцков, доцент кафедры системного анализа и компьютерного моделирования, факультета радиофизики и компьютерных технологий Белорусского государственного университета, кандидат физико-математических наук, доцент

Дмитрий Васильевич Щегрикович, доцент кафедры интеллектуальных систем, факультета радиофизики и компьютерных технологий Белорусского государственного университета, кандидат физико-математических наук

РЕЦЕНЗЕНТЫ:

Владимир Васильевич Голенков, профессор кафедры интеллектуальных информационных технологий УО «Белорусский государственный университет информатики и радиоэлектроники», доктор технических наук, профессор

Владимир Владимирович Апанасович, первый проректор Института ИТ и бизнес-администрирования, доктор физико-математических наук, профессор

РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ:

Кафедрой системного анализа и компьютерного моделирования Белорусского государственного университета

(протокол № 6 от 24 декабря 2019 г.)

Кафедрой интеллектуальных систем Белорусского государственного университета

(протокол № 7 от 10 декабря 2019 г.)

Научно-методическим Советом БГУ

(протокол № 3 от 3 января 2020 г.)

Заведующий кафедрой системного анализа
и компьютерного моделирования,
доцент В.В. Скакун

Заведующий кафедрой
интеллектуальных систем,
доцент Е.И. Козлова

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Цели и задачи учебной дисциплины

Цель учебной дисциплины – изучение теоретических основ прикладного анализа данных, включая базовые элементы статистического программирования и прикладного анализа больших наборов данных с использованием языка R.

Задачи учебной дисциплины:

1. сформировать у магистрантов комплексное представление об обобщенном математическом описании и принципах построения алгоритмов интеллектуального анализа данных;
2. научить производить расчеты с применением технологий анализа больших данных;
3. решать широкий спектр прикладных задач обработки больших наборов данных в среде статистического программирования R;
4. научить выбирать и адаптировать под конкретные данные общие алгоритмы машинного обучения;
5. сформировать представление о способах контроля качества моделей машинного обучения при переобучении.

Место учебной дисциплины в системе подготовки специалиста с высшим образованием (магистра).

Учебная дисциплина относится к модулю «Анализ данных» государственного компонента.

Учебная программа составлена с учетом межпредметных связей и программ по дисциплинам «Теория вероятностей и математическая статистика», «Программирование», «Интеллектуальный анализ данных», «Искусственный интеллект и методы машинного обучения».

Требования к компетенциям

Освоение учебной дисциплины «Прикладной анализ данных» должно обеспечить формирование следующих академических, социально-личностных и профессиональных компетенций:

углубленные профессиональные компетенции:

УПК-6. Владеть методами интеллектуального анализа данных для решения научных и практических задач.

В результате освоения учебной дисциплины студент должен:

знать:

- базовые понятия и принципы прикладного анализа данных;
- основные алгоритмы анализа больших данных и подходы к их созданию;
- задачи прикладного анализа больших наборов данных;
- способы выбора моделей машинного обучения в зависимости от входных данных;
- способы контроля качества моделей при обучении моделей машинного времени.

уметь:

– производить расчеты с применением алгоритмов прикладного анализа данных;
– применять методы прикладного анализа данных в среде статистического программирования R для решения практических задач управления и обработки больших объемов информации;
– творчески и эффективно использовать полученные знания в профессиональной деятельности.

владеть:

– навыками работы на многоядерных вычислительных системах;
– инструментами разработки программных средств с использованием ресурсов Интернет-проектов статистического программирования R-project, RStudio и среды R;
– технологиями прикладного анализа больших данных с использованием среды статистического программирования R;
– навыками контроля качества моделей машинного обучения при решении реальных прикладных задач.

Структура учебной дисциплины

Дисциплина изучается в 2 и 3 семестрах. Всего на изучение учебной дисциплины «Прикладной анализ данных» отведено для очной формы получения высшего образования – 306 часов, в том числе 112 аудиторных часов, из них:

– 2 семестр. Лекции – 20 часов. Лабораторные занятия – 44 часа (из них управляемая самостоятельная работа – 20 часов, в дистанционной форме – 12 часов). Трудоемкость учебной дисциплины – 6 зачетных единиц.
– 3 семестр. Лекции – 20 часов, лабораторные занятия – 28 часов. Трудоемкость учебной дисциплины – 3 зачетных единицы.

Трудоемкость учебной дисциплины составляет 9 зачетных единиц.

Форма текущей аттестации во втором семестре – зачет.

Форма текущей аттестации в третьем семестре – экзамен.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Раздел 1. Введение в прикладной анализ данных

Тема 1.1 Основные понятия дисциплины.
Данные. Большие данные. Анализ больших данных. Основные стратегии анализа больших данных. Структура подхода к анализу больших данных. Информационные ресурсы. Задачи, методы и модели анализа больших данных. Классификация. Кластерный анализ. Регрессия. Ассоциация. Визуализация. Анализ текстовой информации.

Раздел 2. Статистическое программирование на языке R

Тема 2.1. История и основные принципы организации среды R.
Программные ресурсы для анализа данных. Среда статистического программирования R. Типы данных. Константы. Атрибуты объектов. Базовые операторы и функции. Векторы. Скрытая коэрсия. Матрицы. Списки. Факторы. Пропущенные значения. Таблицы. Имена.
Тема 2.2. Работа с данными в R.
Управление наборами данных. Доступ к элементам структур данных, матриц и списков. Частичное совпадение. Векторные операции. Чтение и запись данных. Функции read.table, readLines, source, dget, load, write.table, writeLines, dump, dput.
Тема 2.3. Организация вычислений в R.
Управляющие структуры if, else, for, while, repeat, break, next, return. Функции. Сопоставление аргументов функций. Правила следования аргументов функций. Ленивое вычисление (Lazy Evaluation).
Тема 2.4. Обработка даты и времени, векторизированные вычисления.
Дата и время. Классы Date, POSIXct и POSIXlt. Функции для работы с объектами классов Date, POSIXct и POSIXlt. Векторизированные функции. Функции lapply, sapply, apply, tapply, mapply, split.
Тема 2.5. Правила видимости свободных переменных, отладка программного кода и имитационное моделирование.
Понятие видимого пространства памяти (environment). Правила видимости свободных переменных в функциях. Поиск и исправление ошибок в программном коде. Сообщения message, warning и error. Функции отладки программного кода traceback, debug, browser, trace, recover, print. Программная оболочка для среды программирования R – RStudio. Имитационное моделирование. Базовые функции R для моделирования дискретных и непрерывных случайных величин.
Тема 2.6. Анализ больших данных с использованием ресурсов среды программирования R.
Импорт данных из СУБД. Пакеты DBI и RODBC. Организация высокопроизводительного доступа к данным. Контроль оперативной памяти. Функции memory.limit, memory.size, object.size, gc, memory.profile, Rprofmemory. Оптимизация чтения больших файлов данных. Пакеты readr и

LaF. Оптимизация структуры объектов данных в ходе вычислений. Пакеты data.table, bigmemory и ff. Ускорение программного кода. Пакеты compiler, Rcpp, rJava, inline, Rinside, reticulate. Примеры программирования с использованием пакета Rcpp. Подключение дополнительных программных ресурсов. Revolution R, Microsoft R Open, платформа H2O. Распараллеливание вычислений в R. Пакеты. Rmpi, snow, snowfall, parallel, foreach. Система Hadoop и R. Вычислительная модель Map/Reduce. Распределенная файловая система HDFS. Пакеты RHadoop, RHIPE и Hadoop streaming.

Раздел 3. Интеллектуальный анализ больших данных в среде R

Тема 3.1. Предварительный анализ данных.
Описательная статистика. Характеристики центральной тенденции. Характеристики вариации. Графическое представление данных. Двумерный график. Гистограмма. Изоповерхности и контурные линии. Коробчатая диаграмма. Столбиковые диаграммы. Диаграмма рассеяния. Диаграмма рассеяния в 3D. Поверхность функции. Очистка данных. Нормировка и стандартизация данных. Анализ выбросов и аномальных значений.

Тема 3.2. Корреляционный и регрессионный анализ.
Корреляционный анализ. Коэффициент корреляции Пирсона. Определение значимости коэффициента корреляции. Ранговая корреляция. Критерий Спирмена. Критерий Кендэла. Частная корреляция. Регрессионный анализ. Общие модели линейной и нелинейной регрессий. Проверка точности регрессионной модели.

Тема 3.3. Дисперсионный анализ и методы снижения размерности данных.
Дисперсионный анализ. Однофакторный и двухфакторный дисперсионный анализ. Методы снижения размерности данных. Метод главных компонент. Метод главных координат. Факторный анализ.

Тема 3.4. Иерархические методы кластерного анализа.
Кластерный анализ. Основные элементы кластерного анализа. Этапы кластерного анализа. Типы кластеров. Расстояния между объектами данных. Математические характеристики кластера. Критерии качества кластеризации. Классификация алгоритмов кластерного анализа. Иерархические методы кластерного анализа. Дендрограмма. Формула Ланса-Уильямса. Меры сходства кластеров данных. Иерархический агломеративный, дивизимный и гибридный кластерный анализ. Оценка значимости кластеров. Пакет pvclust.

Тема 3.5. Неиерархические методы кластерного анализа.
Алгоритмы на основе k-средних – Fuzzy k-средних, k-медоидов, PAMk, CLARA. Алгоритмы DBSCAN и спектральный.

Тема 3.6. Методы классификации: k-ближайших соседей и байесовских сетей.
Методы классификации данных. Алгоритмы k-ближайших соседей. Методы V-кратного перекрестного контроля и bootsrap. Байесовская классификация.

Тема 3.7. Методы классификации: деревьев решений, нейронных сетей, опорных векторов.

Деревья решений. Анализ данных с использованием деревьев решений. Методика «разделяй и властвуй». Критерии и функции качества разбиения узлов дерева. Индекс Джини. Энтропия. Ошибка классификации. Остановка построения дерева. Сокращение дерева. Обработка пропущенных значений. Извлечение правил из деревьев. Алгоритмы построения деревьев решений. Алгоритмы Conditional Inference Tree, CART, Random Forests. Нейронные сети. Нейрон. Нейронная сеть. Обучение нейронной сети. Нейронные сети Кохонена. Метод опорных векторов.

Тема 3.8. Методы поиска ассоциативных правил и визуализация многомерных данных.

Методы поиска ассоциативных правил. Ассоциативные правила. Алгоритм Apriori. Визуальное представление данных. Ресурсы среды R для визуального представления данных.

Раздел 4. Описание поведения пользователей продуктов с применением методов машинного обучения

Тема 4.1. Сегментация пользователей по поведению.

Выбор базиса описания поведения. Выбор метода кластеризации. Переход от методов кластеризации к методам классификации.

Тема 4.2. Сегментация пользователей по платежам.

Выбор базиса описания платежного поведения. RFM-анализ. Сегментация на основании квартилей.

Тема 4.3. Контроль качества сегментации.

Оценка штрафа переобучения модели сегментации. Оценка смещения центров кластеров. Оценка перераспределения наблюдений между кластерами.

Тема 4.4. Особенности сегментации пользователей в B2C и B2B сегментах рынка.

Сегментация аккаунтов и сегментация пользователей внутри аккаунтов. Выделение ролей и персон на основе кластеризации данных.

Раздел 5. Предсказание поведения пользователей продуктов с применением методов машинного обучения

Тема 5.1. Предсказание ухода пользователя.

Выбор факторов для описания. Выбор алгоритма. Определение метрики качества.

Тема 5.2. Предсказание прекращения подписки пользователя.

Выбор базиса метрик для описания пользователя. Различие понятия ухода пользователя и прекращения подписки. Контроль качества моделей машинного обучения.

Тема 5.3. Предсказание следующей покупки пользователя.

Подготовка данных. Задача регрессии. Задача классификации. Выбор горизонта предсказания.

Тема 5.4. Предсказание LTV пользователя.

Понятие LTV. Методы регрессии для оценки LTV. Задача мульти-классификации.

Тема 5.5. Особенности сбора требований при постановке задач описания и предсказания поведения пользователей.

Определение целевой переменной. Определение расписания переобучения моделей машинного обучения. Визуализация результатов работы разработанных решений по обработке данных.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

Дневная форма получения образования с применением дистанционных образовательных технологий

Номер раздела, темы	Название раздела, темы	Количество аудиторных часов						Форма контроля знаний
		Лекции	Практические занятия	Семинарские занятия	Лабораторные занятия	Иное	Количество часов УСР	
1	2	3	4	5	6	7	8	9
1	**Введение в прикладной анализ данных**	**2**						
1.1	Основные понятия дисциплины	2						Опрос
2	**Статистическое программирование на языке R**	**12**			**8**			
2.1.	История и основные принципы организации среды R							Опрос
2.2.	Работа с данными в R	2			2			Опрос. Отчет по лабораторной работе
2.3.	Организация вычислений в R	2			2			Опрос. Отчет по лабораторной работе
2.4.	Обработка даты и времени, векторизированные вычисления	2						Опрос
2.5.	Правила видимости свободных переменных, отладка программного	2						Опрос
2.6.	Анализ больших данных с использованием ресурсов среды	4			4			Опрос. Отчет по лабораторной работе

	программирования R							
3	**Интеллектуальный анализ больших данных в среде R**	**6**			**16**		**20**	
3.1.	Предварительный анализ данных	2			4			Опрос. Метод учебной дискуссии. Отчет по лабораторной работе
3.2.	Корреляционный и регрессионный анализ							Отчет по лабораторной работе
3.3	Дисперсионный анализ и методы снижения размерности данных				4		4	Отчет по лабораторной работе
3.4	Иерархические методы кластерного анализа	2			4			Опрос. Отчет по лабораторной работе
3.5	Неиерархические методы кластерного анализа	2			4		4	Опрос. Метод учебной дискуссии. Отчет по лабораторной работе
3.6.	Методы классификации: k-ближайших соседей и байесовских сетей						6 (ДО)	Опрос. Отчет по лабораторной работе
3.7.	Методы классификации: деревьев решений, нейронных сетей, опорных векторов						6 (ДО)	Опрос. Метод учебной дискуссии. Отчет по лабораторной работе
3.8.	Методы поиска ассоциативных правил и визуализация многомерных данных							Отчет по лабораторной работе
4	**Описание поведения**	**10**			**10**			

	пользователей продуктов с применением методов машинного обучения							
4.1.	Сегментация пользователей по поведению	2			4			Отчет по лабораторной работе
4.2.	Сегментация пользователей по платежам	2						Опрос
4.3.	Контроль качества сегментации	3			2			Отчет по лабораторной работе
4.4.	Особенности сегментации пользователей в B2C и B2B сегментах рынка	3			4			Опрос. Метод учебной дискуссии. Отчет по лабораторной работе
5	**Предсказание поведения пользователей продуктов с применением методов машинного обучения**	**10**			**18**			
5.1.	Предсказание ухода пользователя	2			4			Отчет по лабораторной работе
5.2.	Предсказание прекращения подписки пользователя	2			4			Опрос. Метод учебной дискуссии. Отчет по лабораторной работе
5.3.	Предсказание следующей покупки пользователя	2						
5.4.	Предсказание LTV пользователя	2			4			Опрос. Отчет по лабораторной работе
5.5.	Особенности сбора требований	2			6			Опрос. Метод

	при постановке задач описания и предсказания поведения пользователей							учебной дискуссии. Отчет по лабораторной работе
	Итого	**40**			**52**		**20**	

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Перечень основной литературы

1. Machine Learning Models and Algorithms for Big Data Classification / S. Suthaharan – Integrated Series in Information Systems 36, Springer Science+Business Mrdia New York, 2016. – 359 pp.
2. Лесковец, Ю. Анализ больших наборов данных / Ю. Лесковец, А. Раджараман, Д. Д. Ульман – Москва: ДМК Пресс, 2016. – 498 с.
3. Статистический анализ и визуализация данных с помощью R / С.Э. Мастицкий, В.К. Шитиков – Хайдельберг – Лондон – Тольятти. 2014. – 401 с.
4. Bishop, C. M., Pattern recognition and machine learning / C. M. Bishop. –New York: Springer, 2016. – 738 p.
5. Ghahramani, Z. Unsupervised Learning / Z. Ghahramani. – Gatsby Computational Neuroscience Unit University College London, 2004. – 32 p.
6. Яцков, Н. Н. Интеллектуальный анализ данных : пособие / Н. Н. Яцков. – Минск : БГУ, 2014. – 151 с.
7. Анализ больших данных : методические указания к лабораторным работам / Н. Н. Яцков, Е. В. Лисица. – Минск: БГУ, 2019. – 50 с.

Перечень дополнительной литературы

1. Learning Data Mining with R / B. Makhabel – Packt Publishing. 2015. – 314 pp.
2. Bramer, M. Principles of Data Mining / M. Bramer. – Third edition. London : Springer-Verlag London Ltd. 2016. – 526 p.
3. Aggarwal, C.C. Data Mining: The Textbook / C.C. Aggarwal. – Springer International Publishing Switzerland. 2015. 734 P.
4. Rojas, R. Neural Networks. A systematic introduction. / R. Rojas, J. Feldman. – Berlin: Springer Science & Business Media, 1996. – 502 p.
5. Segaran, T. Programming Collective Intelligence: Building Smart Web 2.0 Applications / T. Segaran. – O'Reilly Media, 2008. – 368 p.
6. Grus, J. Data Science from Scratch: First Principles with Python / J. Grus. – O'Reilly Media, 2015. – 330 p.

Перечень рекомендуемых средств диагностики и методика формирования итоговой оценки

Для диагностики компетенций учащихся рекомендуется использовать следующие формы: устная и техническая.

Оценка за ответы на лекциях (опрос) может включать в себя полноту ответа, наличие аргументов, примеров из практики и т.д. Используется ***метод учебной дискуссии***.

При оценивании лабораторных работ принимается во внимание правильность полученных результатов, владение соответствующим теоретическим материалом, ответы на контрольные вопросы, способность учащегося теоретически обосновать и детально пояснить полученные результаты и практическую реализацию задания.

Техническая форма диагностики реализуется в виде электронных тестов, проводимых с использованием специализированных информационных систем, применяемых в БГУ.

Формой текущей аттестации по дисциплине «Прикладной анализ данных» учебным планом предусмотрены: во втором семестре – зачет, в третьем семестре – экзамен.

При формировании итоговой оценки используется рейтинговая оценка знаний студента, дающая возможность проследить и оценить динамику процесса достижения целей обучения. Рейтинговая оценка предусматривает использование весовых коэффициентов для текущего контроля знаний и текущей аттестации студентов по дисциплине.

Формирование рейтинговой оценки за текущую успеваемость:

- опрос по материалам лекций – 40 %;
- отчет лабораторных работ – 60 %.

Итоговая оценка по дисциплине рассчитывается на основе рейтинговой оценки текущей успеваемости и оценки на зачете/экзамене с учетом их весовых коэффициентов. Вес рейтинговой оценки по текущей успеваемости в итоговой оценке на зачете/экзамене составляет 50%.

Примерный перечень заданий для управляемой самостоятельной работы студентов

1. Метод главных координат (4 часа).
2. Спектральный метод кластерного анализа (4 часа).
3. Метод k-ближайших соседей (6 часов). ДО. Ознакомится с видеоматериалами по теме «Методы классификации: k-ближайших соседей и байесовских сетей», размещенными в системе управления обучением Moodle. Задание: разработать и реализовать на языке программирования R алгоритмы метода k-ближайших соседей для решения задачи классификации данных и V-кратного перекрестного контроля. Определить оптимальное число k ближайших соседей с использованием алгоритма V-кратного перекрестного контроля. Форма отчета: работающий программный модуль, реализующий

задание, тексты программ. Результаты классификации тестируемых данных методом k-ближайших соседей. Выводы по результатам обработки экспериментальных данных.

4. Метод опорных векторов (6 часов). ДО. Ознакомится с видеоматериалами по теме «Методы классификации: деревьев решений, нейронных сетей, опорных векторов», размещенными в системе управления обучением Moodle. Задание: решить задачу классификации данных с использованием метода опорных векторов. Форма отчета: работающий программный модуль, реализующий задание, тексты программ. Результаты классификации данных методом опорных векторов. Выводы по результатам обработки экспериментальных данных.

Примерная тематика лабораторных занятий

2-й семестр.

1. Основы работы в R (4 часа).
2. Предварительный анализ больших наборов данных (4 часа).
3. Распараллеливание вычислений (4 часа).
4. Метод главных компонент для сжатия больших данных (4 часа).
5. Иерархические методы кластерного анализа больших данных (4 часа).
6. Неиерархические методы кластерного анализа. Метод k-медоидов (4 часа).

3-й семестр.

1. Сегментация пользователей по поведению. Выбор метода кластеризации и переход к классификации (4 часа).
2. Контроль качества сегментации. Модель сегментации и оценка ее переобучения (2 часа).
3. Особенности сегментации пользователей в B2C и B2B сегментах рынка. Выделение ролей и персон на основе кластеризации данных (4 часа).
4. Предсказание ухода пользователя. Выбор факторов для описания и алгоритма. Определение метрики качества (4 часа).
5. Предсказание прекращения подписки пользователя. Выбор базиса метрик для описания пользователя. Контроль качества моделей машинного обучения (4 часа).
6. Задача мульти- классификации. Методы регрессии для оценки LTV (4 часа).
7. Сбор требований в задаче описания и предсказания поведения пользователей. Определение целевой переменной и расписания переобучения моделей машинного обучения. Визуализация результатов работы разработанных решений по обработке данных (6 часов).

Описание инновационных подходов и методов к преподаванию учебной дисциплины

При организации образовательного процесса используется ***метод учебной дискуссии,*** который предполагает участие студентов в целенаправленном обмене мнениями, идеями для предъявления и/или согласования существующих позиций по определенной проблеме.

Использование метода обеспечивает появление нового уровня понимания изучаемой темы, применение знаний (теорий, концепций) при решении проблем, определение способов их решения.

При организации образовательного процесса ***используется метод проектного обучения,*** который предполагает организацию учебной деятельности студентов, направленной на развитие навыков планирования, самоорганизации, сотрудничества и предполагающий создание собственного программного продукта.

Использование метода проектного обучения обеспечивает практическое применение полученных знаний и позволяет приобрести опыт коллективной работы, необходимый для молодых специалистов.

Методические рекомендации по организации самостоятельной работы обучающихся

При изучении учебной дисциплины рекомендуется использовать следующие формы самостоятельной работы:

– поиск (подбор) и обзор литературы и электронных источников по индивидуально заданной проблеме курса;

– решение задач и выполнение упражнений, выдаваемых на практических занятиях;

– подготовка презентаций в качестве отчетов по лабораторным работам;

– подготовка к лабораторным занятиям с использованием размещенных в сетевом доступе учебных и учебно-методических материалов (программа курса, электронные учебные материалы лекций, методические указания, задания и информационные ресурсы для выполнения лабораторных работ, список рекомендуемой литературы и др.).

Примерный перечень вопросов к экзамену/зачету

Примерный перечень вопросов к зачету (2-й семестр)

1. Основные понятия дисциплины. Данные. Большие данные.
2. Основные стратегии анализа больших данных.
3. Структурная схема подхода к анализу больших данных.
4. Задачи и методы анализа больших данных. Классификация. Кластерный анализ. Регрессия. Ассоциация. Визуализация. Анализ текстовой информации.
5. История и основные принципы организации среды R.

6. R. Типы данных. Константы. Атрибуты объектов. Базовые операторы и функции. Векторы. Скрытая коэрсия.
7. R. Матрицы. Списки. Факторы. Пропущенные значения. Таблицы. Имена.
8. R. Управление наборами данных. Доступ к элементам структур данных, матриц и списков.
9. R. Чтение и запись данных. Функции read.table, readLines, source, dget, load, write.table, writeLines, dump, dput.
10. R. Управляющие структуры if, else, for, while, repeat, break, next, return.
11. R. Функции. Сопоставление аргументов функций. Правила следования аргументов функций. Ленивое вычисление (Lazy Evaluation).
12. R. Обработка даты и времени.
13. R. Векторизированные функции lapply, sapply, apply, tapply, mapply, split.
14. R. Правила видимости свободных переменных в функциях.
15. R. Поиск и исправление ошибок в программном коде. Сообщения message, warning и error.
16. R. Поиск и исправление ошибок в программном коде. Функции отладки программного кода traceback, debug, browser, trace, recover, print.
17. R. Имитационное моделирование. Базовые функции R для моделирования дискретных и непрерывных случайных величин.
18. R. Анализ больших данных с использованием ресурсов среды программирования R. Импорт данных из СУБД.
19. R. Контроль оперативной памяти. Функции memory.limit, memory.size, object.size, gc, memory.profile, Rprofmemory.
20. R. Распараллеливание вычислений в R. Пакет foreach.
21. R. Система Hadoop и R. Вычислительная модель Map/Reduce. Распределенная файловая система HDFS.
22. Предварительный анализ данных. Описательная статистика
23. Предварительный анализ данных. Графическое представление данных.
24. Предварительный анализ данных. Очистка, нормировка и стандартизация данных.
25. Предварительный анализ данных. Анализ выбросов и аномальных значений.
26. Корреляционный анализ. Коэффициент корреляции Пирсона. Определение значимости коэффициента корреляции.
27. Ранговая корреляция. Критерий Спирмена.
28. Ранговая корреляция. Критерий Кендэла.
29. Регрессионный анализ. Модель линейной регрессии.
30. Дисперсионный анализ. Однофакторный и двухфакторный дисперсионный анализ.
31. Метод главных компонент.
32. Метод главных координат.
33. Факторный анализ.
34. Кластерный анализ. Типы кластеров. Математические характеристики кластера. Критерии качества кластеризации.

35. Иерархические методы кластерного анализа. Дендрограмма. Формула Ланса-Уильямся для вычисления меры сходства кластеров.
36. Иерархический агломеративный, дивизимный и гибридный кластерный анализ.
37. Неиерархические методы кластерного анализа. Алгоритм Fuzzy k-средних.
38. Неиерархические методы кластерного анализа. Алгоритм k-медоидов.
39. Неиерархические методы кластерного анализа. Алгоритмы PAMk и CLARA.
40. Неиерархические методы кластерного анализа. Алгоритм DBSCAN.
41. Неиерархические методы кластерного анализа. Алгоритм спектральный.
42. Алгоритмы k-ближайших соседей. Метод V-кратного перекрестного контроля.
43. Метод bootstrap.
44. Байесовская классификация.
45. Деревья решений. Анализ данных с использованием деревьев решений.
46. Алгоритмы построения деревьев решений. Алгоритмы Conditional Inference Tree, CART, Random Forests.
47. Нейрон. Нейронная сеть. Обучение нейронной сети.
48. Нейронные сети Кохонена.
49. Поиск ассоциативных правил. Алгоритм Apriori.
50. Визуальное представление данных. Ресурсы среды R для визуального представления данных.

Примерный перечень вопросов к экзамену (3-й семестр)

1. Описание поведения пользователей продуктов с применением методов машинного обучения.
2. Сегментация пользователей по поведению.
3. Выбор базиса описания поведения.
4. Выбор метода кластеризации для описания поведения пользователей.
5. Переход от методов кластеризации к методам классификации для описания поведения пользователей.
6. Сегментация пользователей по платежам.
7. Выбор базиса описания платежного поведения.
8. RFM-анализ.
9. Сегментация на основании квартилей в задаче сегментации пользователей по платежам.
10. Контроль качества сегментации.
11. Оценка штрафа переобучения модели сегментации.
12. Оценка смещения центров кластеров.
13. Оценка перераспределения наблюдений между кластерами.
14. Особенности сегментации пользователей в B2C и B2B сегментах рынка.
15. Сегментация аккаунтов и сегментация пользователей внутри аккаунтов.
16. Выделение ролей и персон на основе кластеризации данных.

17. Предсказание поведения пользователей продуктов с применением методов машинного обучения.
18. Предсказание ухода пользователя.
19. Выбор факторов для описания ухода пользователя.
20. Выбор алгоритма для описания ухода пользователя.
21. Определение метрики качества для описания ухода пользователя.
22. Предсказание прекращения подписки пользователя.
23. Выбор базиса метрик для описания пользователя в задаче предсказания отмены подписки.
24. Различие понятия ухода пользователя и прекращения подписки.
25. Контроль качества моделей машинного обучения в задаче предсказания отмены подписки.
26. Предсказание следующей покупки пользователя.
27. Подготовка данных в задаче предсказания следующей покупки.
28. Задача регрессии в проблематике предсказания следующей покупки.
29. Задача классификации в проблематике предсказания следующей покупки.
30. Выбор горизонта предсказания в проблематике предсказания следующей покупки.
31. Предсказание LTV пользователя.
32. Понятие LTV.
33. Методы регрессии для оценки LTV.
34.Задача мульти- классификации для оценки LTV.
35. Особенности сбора требований при постановке задач описания и предсказания поведения пользователей.
36. Определение целевой переменной.
37. Определение расписания переобучения моделей машинного обучения.
38. Визуализация результатов работы разработанных решений по обработке данных.

ПРОТОКОЛ СОГЛАСОВАНИЯ УЧЕБНОЙ ПРОГРАММЫ УВО

Название учебной дисциплины, с которой требуется согласование	Название кафедры	Предложения об изменениях в содержании учебной программы учреждения высшего образования по учебной дисциплине	Решение, принятое кафедрой, разработавшей учебную программу (с указанием даты и номера протокола)
Искусственный интеллект и методы машинного обучения	Кафедра интеллектуальных систем	Нет	Изменений не требуется. Протокол № 6 от 24 декабря 2019 г. Протокол № 7 от 10 декабря 2019 г.

ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ К УЧЕБНОЙ ПРОГРАММЕ ПО ИЗУЧАЕМОЙ УЧЕБНОЙ ДИСЦИПЛИНЕ

на _____/_____ учебный год

№ п/п	Дополнения и изменения	Основание

Учебная программа пересмотрена и одобрена на заседании кафедры системного анализа и компьютерного моделирования (протокол № ____ от _____ 202_ г.)

Учебная программа пересмотрена и одобрена на заседании кафедры интеллектуальных систем (протокол № ____ от _____ 202_ г.)

Заведующий кафедрой системного анализа
и компьютерного моделирования, доцент В.В. Скакун

Заведующий кафедрой
интеллектуальных систем, доцент Е.И. Козлова

УТВЕРЖДАЮ
Декан факультета
_____ _____ _____