Федеральное государственное автономное образовательное учреждение высшего образования
«Российский университет дружбы народов»

Факультет физико-математических и естественных наук

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Интеллектуальный анализ больших данных

Рекомендована МССН для направления подготовки/специальности:

09.04.03 Прикладная информатика

Освоение дисциплины ведется в рамках реализации основной профессиональной образовательной программы высшего образования (ОП ВО):

Магистерская программа «Искусственный интеллект и анализ данных»

2022 г.

1. Цель освоения дисциплины

Настоящая дисциплина ставит своей целью ознакомление обучающихся с задачами, возникающими в области интеллектуального анализа (Data Mining) больших данных (Big Data), и методами их решения, которые помогут выявлять, формализовывать и успешно решать практические задачи интеллектуального анализа данных, возникающие в процессе профессиональной деятельности.

В ходе изучения дисциплины перед обучающимися ставятся следующие задачи:

- изучение методов и моделей интеллектуального анализа данных;
- изучение методов и моделей больших данных;
- получение представления об алгоритмах построения деревьев решений;
- изучение алгоритмов классификации и регрессии;
- изучение алгоритмов поиска ассоциативных правил;
- изучение методов кластеризации.

2. Требования к результатам освоения дисциплины

Освоение дисциплины «Интеллектуальный анализ больших данных» направлено на формирование у обучающихся следующих компетенций:

Таблица № 2.1. Перечень компетенций, формируемых у обучающихся при освоении дисциплины (результаты освоения дисциплины)

Шифр	**Компетенция**	**Индикаторы достижения компетенции** (в рамках данной дисциплины)
УК-1	Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий	УК-1.1. Знает принципы сбора, отбора и обобщения информации
		УК-1.2. Умеет соотносить разнородные явления и систематизировать их в рамках избранных видов профессиональной деятельности
		УК-1.3. Имеет практический опыт работы с информационными источниками, опыт научного поиска, создания научных текстов
ОПК-1	Способность самостоятельно приобретать, развивать и применять математические, естественнонаучные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте	ОПК-1.1. Обладает фундаментальными знаниями в области математических и естественных наук, информатики и теории коммуникаций
		ОПК-1.2. Умеет осуществлять первичный сбор и анализ материала, интерпретировать различные математические и информационные объекты
		ОПК-1.3. Имеет практический опыт работы с решением математических и информационных задач и применяет его в профессиональной деятельности
ПК-1	Проведение работ по обработке и анализу научно-технической информации и результатов	ПК-1.1 Знает основы научно-исследовательской деятельности в области информационных технологий; владеет знанием основ философии и методологии науки; владеет методами научных

Шифр	Компетенция	Индикаторы достижения компетенции (в рамках данной дисциплины)
	исследований	исследований, умеет применять их на практике.
		ПК-1.2 Знает принципы построения научной работы, методы сбора и анализа полученного материала, способы аргументации; владеет навыками подготовки научных обзоров, публикаций, рефератов и библиографий по тематике проводимых исследований на русском и иностранном языке; способен готовить публикации в научно-технических тематических изданиях
		ПК-1.3 Умеет применять полученные знания в области фундаментальных научных основ математики и информатики, а также решать стандартные задачи собственной научно-исследовательской деятельности; умеет решать научные задачи с пониманием существующих подходов к верификации моделей по тематике исследований в соответствии с выбранной методикой

3. Место дисциплины в структуре ОП ВО

Дисциплина «Интеллектуальный анализ больших данных» относится к *вариативной* компоненте блока Б1 ОП ВО.

В рамках ОП ВО обучающиеся также осваивают другие дисциплины и/или практики, способствующие достижению запланированных результатов освоения дисциплины «Интеллектуальный анализ больших данных».

Таблица 3.1. Перечень компонентов ОП ВО, способствующих достижению запланированных результатов освоения дисциплины

Шифр	Наименование компетенции	Предшествующие дисциплины/модули, практики	Последующие дисциплины/модули, практики
УК-1	Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий	Глубокое обучение и обучение с подкреплением	Научно-исследовательская работа Производственно-технологическая практика Преддипломная практика
ОПК-1	Способность самостоятельно приобретать, развивать и применять математические, естественнонаучные, социально-	Глубокое обучение и обучение с подкреплением	Научно-исследовательская работа Производственно-технологическая практика Преддипломная

Шифр	Наименование компетенции	Предшествующие дисциплины/модули, практики	Последующие дисциплины/модули, практики
	экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте		практика
ПК-1	Проведение работ по обработке и анализу научно-технической информации и результатов исследований	Глубокое обучение и обучение с подкреплением	Научно-исследовательская работа Производственно-технологическая практика Преддипломная практика

4. Объем дисциплины и виды учебной работы

Общая трудоемкость дисциплины «Интеллектуальный анализ больших данных» составляет **4** зачетные единицы.

Таблица 4.1. Виды учебной работы по периодам освоения ОП ВО

Вид учебной работы	**Всего часов**	**Семестры**
		3
Контактная работа, ак. ч.	54	54
в том числе:		
Лекции (**ЛК**)	18	18
Лабораторные работы (**ЛР**)		
Практические/семинарские занятия (**СЗ**)	36	36
Самостоятельная работа обучающихся, ак. ч.	54	54
Контроль (экзамен/зачет с оценкой), ак. ч.		
Общая трудоемкость дисциплины, ак. ч.	108	108
Общая трудоемкость дисциплины, зач. ед.	3	3

5. Содержание дисциплины

Таблица 5.1. Содержание дисциплины по видам учебной работы

Наименование раздела дисциплины	**Содержание раздела (темы)**	**Вид учебной работы**
Раздел 1. Интеллектуальный анализ данных и	Тема 1.1. Интеллектуальный анализ данных	ЛК, СЗ
	Тема 1.2. Методы машинного обучения без учителя	ЛК, СЗ

Наименование раздела дисциплины	Содержание раздела (темы)	Вид учебной работы
большие данные	Тема 1.3. Метод опорных векторов	ЛК, СЗ
	Тема 1.4. Обучение дерева решений	ЛК, СЗ
Раздел 2. Глубокое обучение с большими данными	Тема 2.1. Искусственные нейронные сети	ЛК, СЗ
	Тема 2.2. Глубокое обучение с библиотекой TensorFlow	ЛК, СЗ
Раздел 3. Обработка больших данных в распределенных вычислительных средах	Тема 3.1. Распределенная вычислительная среда Hadoop	ЛК, СЗ
	Тема 3.2. Интеллектуальный анализ данных на платформе Spark	ЛК, СЗ

6. Материально-техническое обеспечение дисциплины

Таблица 6.1. Материально-техническое обеспечение дисциплины

Тип аудитории	Оснащение аудитории	Специализированное учебное/лабораторное оборудование, ПО и материалы для освоения дисциплины
Лекционная	Аудитория для проведения занятий лекционного типа, оснащенная комплектом специализированной мебели; доской (экраном) и техническими средствами мультимедиа презентаций.	Аудитория оснащена комплектом специализированной мебели. Рабочие места обучающихся, технические средства: интерактивная доска Samsung, рабочая станция Samsung; выход в интернет через ЛВС и Wi-Fi; Программное обеспечение: продукты Microsoft (ОС, пакет офисных приложений, в т.ч. MS Office/ Office 365, Teams, Skype)
Семинарская	Аудитория для проведения занятий семинарского типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, оснащенная комплектом специализированной мебели и техническими средствами мультимедиа презентаций.	–

Тип аудитории	Оснащение аудитории	Специализированное учебное/лабораторное оборудование, ПО и материалы для освоения дисциплины
Для самостоятельной работы обучающихся	Аудитория для самостоятельной работы обучающихся (может использоваться для проведения семинарских занятий и консультаций), оснащенная комплектом специализированной мебели и компьютерами с доступом в ЭИОС.	Дисплейный класс оснащен комплектом специализированной мебели. Рабочие места обучающихся, технические средства: экран Prostar 153*20, переносной проектор, рабочее место обучающегося (моноблок Lenovo) - 12; выход в интернет через ЛВС и Wi-Fi;. Программное обеспечение: продукты Microsoft (ОС, пакет офисных приложений, в т.ч. MS Office/ Office 365, Teams, Skype) Операционная система Linux (дистрибутив Gentoo): - офисный пакет LibreOffice (лицензия MPL-2.0)

7. Учебно-методическое и информационное обеспечение дисциплины

Основная литература:

1. Data mining // [Электронный ресурс] URL: https://www.intuit.ru/studies/courses/6/6/info, режим доступа: свободный.

Дополнительная литература:

1. Введение в аналитику больших массивов данных // [Электронный ресурс] URL: https://www.intuit.ru/studies/courses/12385/1181/info, режим доступа: свободный.

Ресурсы информационно-телекоммуникационной сети «Интернет»:

1. ЭБС РУДН и сторонние ЭБС, к которым студенты университета имеют доступ на основании заключенных договоров:

- ЭБС «Университетская библиотека онлайн» http://www.biblioclub.ru
- ЭБС Юрайт http://www.biblio-online.ru
- ЭБС «Консультант студента» www.studentlibrary.ru
- ЭБС «Лань» http://e.lanbook.com/
- ЭБС «Троицкий мост»

2. Базы данных и поисковые системы:

- электронный фонд правовой и нормативно-технической документации http://docs.cntd.ru/
- поисковая система Яндекс https://www.yandex.ru/
- поисковая система Google https://www.google.ru/

Учебно-методические материалы для самостоятельной работы обучающихся при освоении дисциплины:

1. Лабораторный практикум по дисциплине «Интеллектуальный анализ больших данных»

8. Оценочные материалы и бально-рейтинговая система оценивания уровня

сформированности компетенций по дисциплине

Оценочные материалы и балльно-рейтинговая система оценивания уровня сформированности компетенций по итогам освоения дисциплины «Интеллектуальный анализ больших данных» представлены в Приложении (Фонд оценочных средств) к настоящей Рабочей программе дисциплины.

Разработчик:
доцент кафедры
информационных технологий С.Г. Шорохов

Зав. кафедрой информационных
Технологий Ю.Н. Орлов

Руководитель программы
доцент кафедры
информационных технологий М.Б. Фомин

Федеральное государственное автономное образовательное учреждение высшего образования
«Российский университет дружбы народов»

Факультет физико-математических и естественных наук

Кафедра информационных технологий

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

ПО УЧЕБНОЙ ДИСЦИПЛИНЕ

Интеллектуальный анализ больших данных

Рекомендуется для направления подготовки

09.04.03 – «Прикладная информатика»
Профиль – «Искусственный интеллект и анализ данных»

Квалификация (степень) выпускника

Магистр

Паспорт фонда оценочных средств по дисциплине

Направление: 09.04.03 – «Прикладная информатика», профиль – «Искусственный интеллект и анализ данных»

Дисциплина: Интеллектуальный анализ больших данных

Код контролируемой компетенции или ее части	Контролируемый раздел дисциплины	Контролируемая тема дисциплины	ФОСы (формы контроля уровня освоения ООП)								Баллы темы	Баллы раздела
			Аудиторная работа				Самостоя-тельная работа			Экзамен/Зачет		
			Опрос	Тест	Контрольная работа	Выполнение ЛР	Выполнение ДЗ	Реферат				
УК-1 ОПК-1 ПК-1	Раздел 1. Интеллектуальный анализ данных и большие данные	Тема 1.1. Интеллектуальный анализ данных. Понятие больших данных. Способы масштабирования анализа больших данных. Наборы больших данных. Числовые и категориальные признаки. Основные этапы интеллектуального анализа больших данных. Потоковая передача данных из источников. Предварительная обработка данных. Очистка данных. Пропущенные значения. Зашумленные данные. Нормализация данных. Стохастическое обучение. Пакетный градиентный спуск. Стохастический градиентный спуск (SGD). Определение параметров алгоритма SGD.			10					2	12	48

		Тема 1.2. Методы машинного обучения без учителя. Снижение размерности данных при помощи алгоритма PCA. Кластеризация больших данных при помощи алгоритма K-средних. Допущения алгоритма. Подбор оптимальной величины K. Масштабирование алгоритма K-средних. Алгоритм LDA и его масштабирование.			10					2	12	
		Тема 1.3. Метод опорных векторов (SVM). Гиперплоскости. Разделяющая гиперплоскость. Маржа и опорные вектора. Кусочно-линейная функция потерь и ее варианты. Реализация SVM для больших данных на основе SGD. Отбор признаков посредством регуляризации. Добавление нелинейности в алгоритм SGD. Доводка гиперпараметров SGD.			10					2	12	
		Тема 1.4. Обучение дерева решений. Агрегация выборок. Случайный лес и экстремально рандомизированный лес. Экстремально рандомизированные деревья и большие наборы данных. Алгоритм CART и бустинг. Алгоритм XGBoost. Регрессия на основе XGBoost. Потоковая передача больших наборов данных посредством XGBoost. Стохастический градиентный бустинг и сеточный поиск.			10					2	12	
УК-1 ОПК-1 ПК-1	Раздел 2. Глубокое обучение с большими данными	Тема 2.1. Искусственные нейронные сети. Архитектура нейронной сети. Параллелизация в нейронных сетях. Регуляризация в нейронных сетях. Гиперпараметрическая оптимизация в нейронных сетях. Глубокое обучение с большими данными.			10					3	13	26
		Тема 2.2. Глубокое обучение с библиотекой TensorFlow. Операции TensorFlow. Инкрементное глубокое обучение с большими данными. Сверточные нейронные сети (CNN) в TensorFlow. Сверточный слой. Объединяющий слой. Полносвязный слой. Обучение сети CNN при помощи			10					3	13	

		инкрементной тренировки. Вычисления на GPU.										
УК-1 ОПК-1 ПК-1	Раздел 3. Обработка больших данных в распределенных вычислительных средах	Тема 3.1. Распределенная вычислительная среда Hadoop. Архитектура Hadoop. Распределенная файловая система HDFS. Вычислительная парадигма MapReduce.			10					3	13	26
		Тема 3.2. Интеллектуальный анализ данных на платформе Spark. Распространение переменных по узлам кластера. Предобработка данных в среде Spark. Машинное обучение с платформой Spark. Библиотека pySpark.			10					3	13	
		ИТОГО:			80					20	100	100

Процесс изучения дисциплины направлен на формирование следующих компетенций: УК-1, ОПК-1, ПК-1.

УК-1 – Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий;

ОПК-1 – Способен самостоятельно приобретать, развивать и применять математические, естественнонаучные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте;

ПК-1 – Проведение работ по обработке и анализу научно-технической информации и результатов исследований

Балльно-рейтинговая система оценки уровня знаний

Таблица соответствия баллов и оценок

Баллы БРС	Традиционные оценки РФ	Оценки ECTS
95 - 100	5	A
86 - 94		B
69 - 85	4	C
61 - 68	3	D
51 - 60		E
31 - 50	2	FX
0 - 30		F
51-100	Зачет	Passed

Правила применения БРС

1. Раздел (тема) учебной дисциплины считаются освоенными, если студент набрал более 50 % от возможного числа баллов по этому разделу (теме).
2. Студент не может быть аттестован по дисциплине, если он не освоил все темы и разделы дисциплины, указанные в сводной оценочной таблице дисциплины.
3. По решению преподавателя и с согласия студентов, не освоивших отдельные разделы (темы) изучаемой дисциплины, в течение учебного семестра могут быть повторно проведены мероприятия текущего контроля успеваемости или выданы дополнительные учебные задания по этим темам или разделам. При этом студентам за данную работу засчитывается минимально возможный положительный балл (51 % от максимального балла).
4. При выполнении студентом дополнительных учебных заданий или повторного прохождения мероприятий текущего контроля полученные им баллы засчитываются за конкретные темы. Итоговая сумма баллов не может превышать максимального количества баллов, установленного по данным темам (в соответствии с приказом Ректора № 564 от 20.06.2013). По решению преподавателя предыдущие баллы, полученные студентом по учебным заданиям, могут быть аннулированы.
5. График проведения мероприятий текущего контроля успеваемости формируется в соответствии с календарным планом курса. Студенты обязаны сдавать все задания в сроки, установленные преподавателем.
6. Время, которое отводится студенту на выполнение мероприятий текущего контроля успеваемости, устанавливается преподавателем. По завершение отведенного времени студент должен сдать работу преподавателю, вне зависимости от того, завершена она или нет.
7. Использование источников (в том числе конспектов лекций и лабораторных работ) во время выполнения контрольных мероприятий возможно только с разрешения преподавателя.

8. Отсрочка в прохождении мероприятий текущего контроля успеваемости считается уважительной только в случае болезни студента, что подтверждается наличием у него медицинской справки, заверенной круглой печатью в поликлинике № 25, предоставляемой преподавателю не позднее двух недель после выздоровления. В этом случае выполнение контрольных мероприятий осуществляется после выздоровления студента в срок, назначенный преподавателем. В противном случае, отсутствие студента на контрольном мероприятии признается не уважительным.
9. Студент допускается к итоговому контролю знаний с любым количеством баллов, набранных в семестре.
10. Итоговый контроль знаний оценивается из 20 баллов независимо от числа баллов за семестр.
11. Если в итоге за семестр студент получил менее 31 балла, то ему выставляется оценка F и студент должен повторить эту дисциплину в установленном порядке. Если же в итоге студент получил 31-50 баллов (т. е. FX), то студенту разрешается добор необходимого (до 51) количества баллов путем повторного одноразового выполнения предусмотренных контрольных мероприятий, при этом по усмотрению преподавателя аннулируются соответствующие предыдущие результаты. Ликвидация задолженностей проводится в период с 07.02 по 28.02 (с 07.09 по 28.09) по согласованию с деканатом.

Примерный перечень оценочных средств

п/п	Наименование оценочного средства	Краткая характеристика оценочного средства	Представление оценочного средства в фонде
		Аудиторная работа	
1	Лабораторная работа	Система практических заданий, направленных на формирование практических навыков у обучающихся	Фонд практических заданий
2	Тест *	Система стандартизированных заданий (вопросов), позволяющая автоматизировать процедуру измерения уровня знаний и умений обучающегося.	База тестовых заданий
3	Опрос *	Средство контроля, организованное как специальная беседа преподавателя с обучающимся на темы, связанные с изучаемой дисциплиной, и рассчитанное на выяснение объема знаний обучающегося по определенному разделу или теме.	Вопросы по темам/разделам дисциплины
4	Экзамен *	Оценка работы студента в течение семестра (года, всего срока обучения и др.) и призван выявить уровень, прочность и систематичность полученных им теоретических и практических знаний, приобретения навыков самостоятельной работы, развития творческого мышления, умение синтезировать полученные знания и применять их в решении практических задач.	Примеры заданий/вопросов, пример экзаменационного билета

Самостоятельная работа			
1	Подготовка отчетов по результатам выполнения лабораторных работ	Форма проверки качества выполнения студентами лабораторных работ в соответствии с утвержденной программой.	Фонд практических заданий

Учебным планом на изучение дисциплины отводится один семестр. В дисциплине предусмотрены лекции, лабораторный практикум, контрольные мероприятия по проверке отчётов по лабораторным работам. В конце семестра проводится итоговый контроль знаний.

Оценивание результатов освоения дисциплины производится в соответствии с балльно-рейтинговой системой. По дисциплине предусмотрен экзамен.

(*) Итоговый контроль знаний по дисциплине проводится в форме тестирования, но при необходимости экзамен может проводится в форме письменного ответа на вопросы из билетов или в форме опроса.

Критерии оценки по дисциплине

95-100 баллов:

- полное и своевременное выполнение на высоком уровне лабораторных работ с оформлением отчетов, успешное прохождение контрольных мероприятий, предусмотренных программой курса;
- систематизированное, глубокое и полное освоение навыков и компетенций по всем разделам программы дисциплины;
- использование научной терминологии, стилистически грамотное, логически правильное изложение ответов на вопросы, умение делать обоснованные выводы;
- безупречное владение программным обеспечением, умение эффективно использовать его в постановке и решении научных и профессиональных задач;
- выраженная способность самостоятельно и творчески решать поставленные задачи;
- полная самостоятельность и творческий подход при изложении материала по программе дисциплины;
- полное и глубокое усвоение основной и дополнительной литературы, рекомендованной программой дисциплины и преподавателем.

86- 94 балла:

- полное и своевременное выполнение на хорошем уровне лабораторных работ с оформлением отчетов, успешное прохождение контрольных мероприятий, предусмотренных программой курса;
- систематизированное, глубокое и полное освоение навыков и компетенций по всем разделам программы дисциплины;
- использование научной терминологии, стилистически грамотное, логически правильное изложение ответа на вопросы, умение делать обоснованные выводы;
- хорошее владение программным обеспечением, умение эффективно использовать его в постановке и решении научных и профессиональных задач;
- способность самостоятельно решать поставленные задачи в нестандартных производственных ситуациях;

- усвоение основной и дополнительной литературы, нормативных и законодательных актов, рекомендованных программой дисциплины и преподавателем.

69-85 баллов:

- своевременное выполнение на хорошем уровне лабораторных работ с оформлением отчетов, прохождение контрольных мероприятий, предусмотренных программой курса;
- хороший уровень культуры исполнения лабораторных работ;
- систематизированное и полное освоение навыков и компетенций по всем разделам программы дисциплины;
- владение программным обеспечением, умение использовать его в постановке и решении научных и профессиональных задач;
- способность самостоятельно решать проблемы в рамках программы дисциплины;
- усвоение основной литературы;

51-68 баллов:

- выполнение на удовлетворительном уровне лабораторных работ с оформлением отчетов, прохождение контрольных мероприятий, предусмотренных программой курса;
- систематизированное и полное освоение навыков и компетенций по всем разделам программы дисциплины;
- удовлетворительное владение программным обеспечением, умение использовать его в постановке и решении научных и профессиональных задач;
- способность решать проблемы в рамках программы дисциплины;
- удовлетворительное усвоение основной литературы;

31 - 50 баллов – НЕ ЗАЧТЕНО:

- не выполнение, несвоевременное выполнение или выполнение на неудовлетворительном уровне лабораторных работ, не прохождение контрольных мероприятий, предусмотренных программой курса;
- недостаточно полный объем навыков и компетенции в рамках программы дисциплины;
- неумение использовать в практической деятельности научной терминологии, изложение ответа на вопросы с существенными стилистическими и логическими ошибками;
- слабое владение программным обеспечением по разделам программы дисциплины, некомпетентность в решении стандартных (типовых) производственных задач;
- способность решать проблемы в рамках программы дисциплины;
- удовлетворительное усвоение основной литературы;

0-30 баллов, НЕ ЗАЧТЕНО:

- отсутствие умений, навыков, знаний и компетенции в рамках программы дисциплины;
- невыполнение лабораторных заданий, не прохождение контрольных мероприятий, предусмотренных программой курса; отказ от ответов по программе дисциплины;
- игнорирование занятий по дисциплине по неуважительной причине.

Экзаменационные вопросы

Дисциплина Интеллектуальный анализ больших данных

1. Предварительная обработка данных. Пропущенные значения. Зашумленные данные. Метод биннинга. Преобразование данных. Нормализация данных.
2. Снижение размерности данных. Проекция в подпространство. Вектор ошибки.
3. Метод главных компонент (PCA). Направление с максимальной дисперсией. Минимальная среднеквадратичная ошибка (MSE).
4. Метод главных компонент (PCA). Наилучшая r-мерная аппроксимация. Выбор размерности.
5. Алгоритм метода главных компонент (PCA). Геометрия метода главных компонент.
6. Задача поиска ассоциативных правил. Наборы предметов. Транзакции. Бинарная база данных. Горизонтальное и вертикальное представления.
7. Поддержка набора предметов. Популярные наборы предметов. Ассоциативное правило. Поддержка правила. Достоверность правила. Популярные и сильные правила.
8. Майнинг наборов предметов и правил. Алгоритм Brute Force. Пространство и дерево поиска наборов предметов.
9. Поиск наборов предметов: алгоритм Apriori. Пример.
10. Поиск наборов предметов: алгоритм Eclat. Пересечения наборов транзакций. Пример.
11. Поиск наборов предметов: алгоритм dEclat. Разности наборов транзакций. Пример.
12. Алгоритм построения ассоциативных правил. Пример.
13. Постановка задачи кластеризации. Метрики для ошибки кластеризации. Кластеризация методом полного перебора.
14. Кластеризация через представителей. Алгоритм k средних (k-means). Сходимость алгоритма. Пример.
15. Начальные центры кластеров. Алгоритм выбора начальных центров кластеров.
16. Иерархическая кластеризация. Вложенные разбиения. Дендрограмма кластеризации.
17. Аггломеративная иерархическая кластеризация. Алгоритм аггломеративной кластеризации.
18. Расстояние между кластерами. Кластеризация методом одиночной связи. Формула Ланса-Уильямса.
19. Меры качества кластеризации. Таблица сопряженности. Чистота кластера. Чистота кластеризации.
20. Байесовский классификатор. Теорема Байеса. Оценка априорной вероятности класса.
21. Байесовский классификатор. Параметрический подход для числовых признаков. Алгоритм байесовской классификации. Пример.
22. Байесовский классификатор. Классификация категориальных признаков. Пример.
23. Наивный байесовский классификатор. Алгоритм наивной байесовской классификации. Пример.
24. Наивный байесовский классификатор. Классификация категориальных признаков. Пример.
25. Метод K ближайших соседей. Пример.
26. Классификатор дерева решений. Рекурсивные разбиения. Гиперплоскости. Чистота области.
27. Алгоритм построения дерева принятия решений.
28. Оценка разбиения: энтропия, информационный выигрыш, индекс Джини.
29. Оценка разбиения для числовых признаков. Алгоритм оценки числовых признаков.

30. Оценка разбиения для категориальных признаков. Алгоритм оценки категориальных признаков.

31. Линейный дискриминантный анализ. Проекция на прямую. Оптимальный линейный дискриминант.

32. Линейный дискриминантный анализ. Линейный дискриминант Фишера. Алгоритм линейного дискриминанта.

33. Оценка классификации. Меры качества классификации: доля ошибок, точность.

34. Меры оценки качества классификации на основе таблицы сопряженности. Точность и полнота класса. F-мера.

35. Регрессионная модель. Линейная регрессия. Одномерный случай. Нелинейная регрессия.

36. Оценка качества регрессии. Коэффициент детерминации.

37. Метод опорных векторов. Разделяющая гиперплоскость. Зазор и опорные векторы. Каноническая гиперплоскость.

38. Метод опорных векторов. Линейный и разделимый случай. Классификатор метода опорных векторов.

39. Метод опорных векторов с мягким зазором. Прямая задача оптимизации. Оптимизация методом Ньютона.

40. Метод опорных векторов с мягким зазором. Двойственная задача оптимизации. Градиентный подъем и стохастический градиентный подъем.

Критерии оценки итогового тестирования

Итоговое тестирование оценивается в соответствии с БРС и паспортом ФОС. Проверяется правильность и полнота ответов на вопросы экзаменационного билета.

Комплект заданий лабораторного практикума

Лабораторная работа № 1. Расчет статистических показателей и визуализация заданного набора данных.

Задание:

- Дано математическое ожидание **a** и корреляционная матрица **R** двумерного гауссовского распределения.
- Постройте **n** значений случайных признаков **X** и **Y**, имеющих двумерное гауссовское распределение с математическим ожиданием **a** и корреляционной матрицей **R**.
- Визуализируйте построенный набор данных на плоскости в виде набора точек с координатами $\{(x_i, y_i)\}$, $i=1,\ldots,n$.
- Вычислите и выведите на экран для построенных данных математические ожидания, дисперсии, а также корреляцию между данными.
- Считайте статистические данные для двух признаков из заданного набора данных репозитария UCI. Если в записи значение какого-либо из признаков не определено (символ “?”), то следует пропустить данную запись.
- Изобразите считанные из набора данные в виде точек на плоскости.
- Вычислите и выведите на экран для двух признаков математические ожидания, дисперсии, а также корреляцию между признаками.
- Подготовьте отчет о выполнении заданий лабораторной работы.

Лабораторная работа № 2. Решение задачи кластеризации данных при помощи алгоритма К-средних.

Задание:

- Для двух классов объектов даны значения количеств объектов (точек) для каждого класса ($\mathbf{n_1}$ и $\mathbf{n_2}$), значения векторов математических ожиданий для каждого класса ($\mathbf{a_1}$ и $\mathbf{a_2}$) и корреляционные матрицы для каждого класса ($\mathbf{R_1}$ и $\mathbf{R_2}$) для моделируемой выборки из гауссовских случайных векторов.
- Для каждого из классов постройте значения случайных признаков **X** и **Y**, имеющие двумерное гауссовское распределение с математическим ожиданием $\mathbf{a_i}$ и корреляционной матрицей $\mathbf{R_i}$.
- Изобразите построенные данные на плоскости в виде точек с координатами $\mathbf{\{(x_i, y_i)\}}$, $\mathbf{i=1,\ldots,n}$ и раскрасьте их разными цветами (красным и синим) для разных классов.
- Проведите кластеризацию построенных объектов с помощью алгоритма **k** средних для случая, когда количество кластеров равно двум.
- Изобразите на плоскости кластеризованные объекты разными цветами для точек разных кластеров и центры кластеров.
- Найдите для построенной кластеризации таблицу сопряженности (contingency table) и вычислите показатель чистоты кластеризации.
- Считайте статистические данные для двух признаков и класса из заданного набора данных репозитария UCI. Если в записи значение какого-либо из признаков или класса не определено (символ “?”), то следует пропустить данную запись.
- Изобразите считанные из набора данные в виде точек на плоскости и раскрасьте их разными цветами для разных классов.
- Проведите кластеризацию построенных объектов с помощью алгоритма **k** средних для случая, когда количество кластеров равно количеству классов.
- Изобразите на плоскости кластеризованные объекты разными цветами для точек разных кластеров и центры кластеров.
- Найдите для построенной кластеризации таблицу сопряженности (contingency table) и вычислите показатель чистоты кластеризации.
- Подготовьте отчет о выполнении заданий лабораторной работы.

Лабораторная работа № 3. Решение задачи классификации данных при помощи метода опорных векторов.

Задание:

- Считайте данные (два признака и метки) из заданного набора данных репозитария UCI.
- Разбейте метки на два класса (положительный и отрицательный) и подготовьте набор данных для обучения бинарного классификатора.
- При необходимости масштабируйте значения признаков при помощи StandardScaler.
- Случайным образом разделите считанные из набора данные на обучающую выборку и контрольную выборку в соотношении 80% на 20%.
- Изобразите обучающую выборку на плоскости в виде точек с координатами $\mathbf{(x_i, y_i)}$ с использованием разных цветов для положительного и отрицательного класса.
- Обучите бинарный классификатор метода опорных векторов LinearSVC на обучающей выборке.
- Произведите классификацию объектов контрольной выборки с помощью обученного классификатора LinearSVC.
- Изобразите объекты контрольной выборки на плоскости разными цветами для положительного и отрицательного классов (на одном рисунке) и для произведенной классификации (на другом рисунке).
- Для проведенной бинарной классификации постройте ROC-кривую и выведите на рисунке показатель AUC ROC (площадь под кривой).
- Подготовьте отчет о выполнении заданий лабораторной работы.

Лабораторная работа № 4. Решение задачи классификации данных при помощи обучения дерева решений.

Задание:

- Считайте статистические данные для двух признаков и класса из заданного набора данных репозитария UCI. Если в какой-либо записи значение какого-либо из признаков или класса не определено (символ “?”), то следует пропустить данную запись.
- Случайным образом разделите считанные из набора данные на обучающую выборку и контрольную выборку в соотношении 70% на 30%.
- Изобразите обучающую выборку на плоскости в виде точек с координатами координатами $\mathbf{\{(x_i, y_i)\}}$, $\mathbf{i=1,\ldots,n}$ и раскрасьте их разными цветами для разных классов.
- Обучите классификатор DecisionTreeClassifier на обучающей выборке, ограничивая глубину дерева значением 5.
- Выполните визуализацию полученного дерева решений.
- Произведите классификацию объектов контрольной выборки с помощью обученного классификатора DecisionTreeClassifier.
- Изобразите объекты контрольной выборки на плоскости разными цветами для исходных классов (на одном рисунке) и для произведенной классификации (на другом рисунке).
- Вычислите и выведите на экран показатель точности классификации.
- Подготовьте отчет о выполнении заданий лабораторной работы.

Лабораторная работа № 5. Решение задачи классификации данных при помощи байесовской классификации и классификации по ближайшим соседям.

Задание:

- Для двух классов объектов даны значения количеств объектов (точек) для каждого класса ($\mathbf{n_1}$ и $\mathbf{n_2}$), значения векторов математических ожиданий для каждого класса ($\mathbf{a_1}$ и $\mathbf{a_2}$) и корреляционные матрицы для каждого класса ($\mathbf{R_1}$ и $\mathbf{R_2}$) для моделируемой выборки из гауссовских случайных векторов.
- Для каждого из классов постройте значения случайных признаков **X** и **Y**, имеющие двумерное гауссовское распределение с математическим ожиданием $\mathbf{a_i}$ и корреляционной матрицей $\mathbf{R_i}$. Объедините построенные данные для двух классов в единый набор.
- Случайным образом разделите полученные данные на обучающую выборку и контрольную выборку в соотношении 80% на 20%.
- Изобразите обучающую выборку в трёхмерном пространстве в виде точек с координатами $\mathbf{\{(x_i, y_i, z_i)\}}$, $\mathbf{i=1,\ldots,n}$ и раскрасьте их разными цветами (например, красным и синим) для разных классов.
- Произведите классификацию объектов контрольной выборки, используя данные о классах объектов из обучающей выборки, с помощью алгоритма наивной байесовской классификации.
- Изобразите объекты контрольной выборки в трёхмерном пространстве разными цветами и маркерами для исходных классов (на одном рисунке) и для произведенной классификации (на другом рисунке).
- Вычислите и выведите на экран показатель точности классификации.
- Считайте статистические данные для трех признаков и класса из заданного набора данных репозитария UCI. Если в какой-либо записи значение какого-либо из признаков или класса не определено (символ “?”), то следует пропустить данную запись.
- Случайным образом разделите считанные из набора данные на обучающую выборку и контрольную выборку в соотношении 75% на 25%.
- Изобразите обучающую выборку в трёхмерном пространстве в виде точек с координатами координатами $\mathbf{\{(x_i, y_i, z_i)\}}$, $\mathbf{i=1,\ldots,n}$ и раскрасьте их разными цветами для разных классов.

- Произведите классификацию объектов контрольной выборки, используя данные о классах объектов из обучающей выборки, с помощью алгоритма **K** ближайших соседей для **K**, равного удвоенному количеству классов в наборе, но не менее 10.
- Изобразите объекты контрольной выборки в трёхмерном пространстве разными цветами для исходных классов (на одном рисунке) и для произведенной классификации (на другом рисунке).
- Вычислите и выведите на экран показатель точности классификации.
- Подготовьте отчет о выполнении заданий лабораторной работы.

Лабораторная работа № 6. Решение задач классификации и прогнозирования данных при помощи регресионного анализа.

Задание:

- Даны количество объектов (точек) **n**, параметры **a** и **b** и дисперсия гауссовского белого шума σ^2.
- Смоделируйте точки $\{(x_i, y_i)\}$, $i=1,\ldots,n$ согласно модели $y=ax+b+\varepsilon$, где в качестве ε используется гауссовский белый шум (нормально распределенная случайная величина) с нулевым математическим ожиданием и заданной дисперсией σ^2. Значения x_i выбираются через равные промежутки на отрезке [0;1].
- Визуализируйте на одном графике точки (x_i, y_i) и прямую $y=ax+b$ при $x \in [0, 1]$.
- Случайным образом разделите полученные данные на обучающую выборку и контрольную выборку в соотношении 75% на 25%.
- Постройте модель линейной регрессии $y=a'x+b'+\varepsilon$ на обучающей выборке.
- Выведите на экран полученные значения a', b' и сравните их с первоначальными значениями **a**, **b**.
- Визуализируйте на одном графике точки (x_i, y_i) из контрольной выборки и прямую $y = a'x+b'$ при $x \in [0, 1]$.
- Вычислите и выведите на экран показатели MSE, MAE и коэффициент детерминации.
- Считайте данные для независимой переменной (предиктора) и зависимой переменной из заданного набора данных репозитария UCI.
- Масштабируйте зависимую переменную на диапазон от 0.001 до 0.999
- Используйте для построения модели три подхода:
    - линейную регрессию
    - полиномиальную регрессию (степень полинома degree=2)
    - преобразование зависимой переменной при помощи логистической функции и последующего применения линейной регрессии
- Случайным образом разделите считанные из набора данные на обучающую выборку и контрольную выборку в соотношении 70% на 30%.
- Изобразите обучающую выборку на плоскости в виде точек с координатами (x_i, y_i).
- Постройте на обучающей выборке различные модели прогнозирования значений зависимой переменной.
- Визуализируйте на одном графике разными цветами точки (x_i, y_i) из контрольной выборки, а также точки с прогнозируемыми значениями зависимой переменной для трех моделей, соединенные линиями (для улучшения картинки может потребоваться сортировка точек контрольной выборки по возрастанию независимой переменной).
- Вычислите и сравните значения показателей MSE, MAE и коэффициента детерминации для различных моделей. Определите лучшую модель по показателю коэффициента детерминации.
- Подготовьте отчет о выполнении заданий лабораторной работы.

Лабораторная работа № 7. Поиск ассоциативных правил при помощи алгоритмов Apriori, Eclat, Declat, FPGrowth.

Задание:

- Скачайте заданный набор данных из репозитария UCI. Считайте из набора данных для последующего анализа только категориальные признаки, исключая числовые признаки.
- Проанализируйте набор данных и оставьте в наборе 10 категориальных, исключая признаки с неопределенными значениями и принимающие одно и то же значение для всех записей наборе данных.
- Преобразуйте записи набора данных в записи транзакционной базы данных (список) следующим образом:
  - в качестве первого элемента добавьте идентификатор транзакции (порядковый номер записи в наборе)
  - в качестве второго элемента добавьте список, составленный из значений признаков записи набора с учетом указанных ниже преобразований и отсортированный по возрастанию значений
  - если признак принимает логические (булевые) значения (True/False), то подставьте вместо значения True название признака, вместо значения False ничего подставлять не нужно, например, если признак имеет название age_gt_60, подставьте вместо этих значения t (True) значение age_gt_60
  - если признак принимает несколько вариантов значения, то подставьте вместо текущего значения конкатенацию названия признака и его текущего значения, например, если признак имеет название ar_c и возможные значения признака normal, elevated, absent, то подставьте вместо этих значений значения ar_c_normal, ar_c_elevated, ar_c_absent
- В качестве минимального значения поддержки примите значение, равное 1/4 количества записей в наборе данных (относительная поддержка 0.25).
- Используя алгоритм, указанный в Вашем варианте, найдите популярные наборы данных для указанного выше значения минимальной поддержки.
- Подготовьте отчет о выполнении заданий лабораторной работы.

Лабораторная работа № 8. Использование метода главных компонент для снижения размерности данных.

Задание:

- Проанализируйте заданный набор данных из репозитария UCI. Считайте из набора данных для последующего анализа только числовые признаки.
- Если в данных значение какого-либо из признаков не определено (символ “?”), то пропустите данную запись.
- Найдите 5 признаков, имеющих наибольшую дисперсию. Если числовых признаков меньше, чем 5, то используйте все числовые признаки.
- Для набора данных, состоящего из пяти признаков с наибольшей дисперсией, найдите размерность метода главных компонент, для которой доля объясняемой дисперсии будет не менее 95%.
- Пользуясь методом главных компонент, снизьте размерность набора данных до двух признаков и изобразите полученный набор данных в виде точек на плоскости. Отображайте точки различных классов разными цветами.
- Подготовьте отчет о выполнении заданий лабораторной работы.

Критерии оценки выполнения лабораторных работ

Оценивается полнота выполнения работы, оформление результатов, наличие примеров использования, полнота ответов на контрольные вопросы, если это предусмотрено заданием.