Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего образования «Ростовский государственный экономический университет (РИНХ)»

УТВЕРЖДАЮ
Директор Института магистратуры
Иванова Е.А.
«01» июня 2023 г.

Рабочая программа дисциплины
Методы машинного обучения

Направление 09.04.03 Прикладная информатика

магистерская программа
09.04.03.03 Машинное обучение и технологии больших данных

Для набора 2023 года

Квалификация
магистр

Кафедра Информационных систем и прикладной информатики

Составители рабочей программы:

доцент Хаймин Евгений Сергеевич

СОДЕРЖАНИЕ

I. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Цели освоения дисциплины:

– формирование у обучающихся способности совершенствовать, разрабатывать и внедрять новые методы, модели, алгоритмы машинного обучения.

Задачи освоения дисциплины:

– развитие у обучающихся умения выбирать и применять математические методы, алгоритмы машинного обучения, программные средства и технологии для решения задач интеллектуального анализа данных (регрессия, классификация, кластеризация, анализ ассоциаций и последовательностей, анализ связей);
– освоение методологии обнаружения в больших массивах «сырых» данных нетривиальных, ранее не известных, доступных интерпретации и практически полезных закономерностей (знаний), необходимых для принятия решений в профессиональной деятельности.

II. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ

Дисциплина относится к модулю обязательных профессиональных дисциплин обязательной части образовательной программы.

Данная дисциплина опирается на базовые знания, умения и навыки, формируемые при получении предшествующего уровня образования.

Знания, умения и навыки, формируемые данной дисциплиной, потребуются при освоении следующих элементов образовательной программы:

– Математические методы анализа больших данных.
– Технологии анализа больших данных.
– Экспертные системы и базы знаний.
– Программирование на языке Python.
– Математические методы и модели поддержки принятия решений.
– Информационный поиск и обработка естественного языка.
– Нейронные сети и глубокое обучение.

III. ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Освоение дисциплины направлено на формирование следующих компетенций в соответствии с образовательной программой:

Перечень планируемых результатов обучения по дисциплине, соотнесённых с индикаторами достижения компетенций

Компетенция	Индикаторы достижения компетенции	Результаты обучения
ПК-1. Способен адаптировать и применять методы и алгоритмы машинного обучения для решения прикладных задач в различных предметных областях	ПК-1.1. Ставит задачи по адаптации или совершенствованию методов и алгоритмов для решения комплекса задач предметной области	*Знания:* – Знает классы методов и алгоритмов машинного обучения. *Умения:* – Умеет ставить задачи и адаптировать методы и алгоритмы машинного обучения.
ПК-2. Способен руководить проектами по созданию систем искусственного интеллекта с применением новых методов и алгоритмов машинного обучения со стороны заказчика	ПК-2.1. Руководит разработкой архитектуры комплексных систем искусственного интеллекта со стороны заказчика	*Знания:* – Знает возможности современных инструментальных средств и систем программирования для решения задач машинного обучения. *Умения:* – Умеет проводить сравнительный анализ и осуществлять выбор инструментальных средств для решения задач машинного обучения.
	ПК-2.2. Осуществляет руководство созданием комплексных систем искусственного интеллекта с применением новых методов и алгоритмов машинного обучения	*Знания:* – Знает функциональность современных инструментальных средств и систем программирования в области создания моделей и методов машинного обучения. *Умения:* – Умеет применять современные инструментальные средства и системы программирования для разработки новых методов и моделей машинного обучения.

.

IV. СОДЕРЖАНИЕ И СТРУКТУРА ДИСЦИПЛИНЫ

Трудоёмкость дисциплины составляет 5 зачётных единиц, 180 часов.

Форма промежуточной аттестации: дифференцированный зачёт

4.1. Содержание дисциплины, структурированное по темам

№ п/п	Темы дисциплины	Семестр	Виды учебной работы и их трудоёмкость, часы (в том числе с использованием онлайн-курсов)				Наименования оценочных средств
			Контактная работа			Самостоя-тельная работа	
			Лекции	Практические занятия	Лабораторные занятия		
Модуль 1. Математическое и алгоритмическое обеспечение аналитики больших данных							
1	Введение в аналитику больших данных (Big Data). Обзор задач интеллектуального анализа данных: регрессия, классификация, кластеризация, анализ ассоциаций и последовательностей, поиск аномалий, анализ связей.	1	2	4	-	14	Тест № 1
2	Методы классификации. Алгоритмы машинного обучения: деревья решений, опорные вектора, байесовские классификаторы	1	2	4	-	14	Практическая работа №1 Тест № 1
3	Оценка эффективности и сравнительный анализ моделей обучения	1	2	4	-	14	Практическая работа №2 Тест № 1
4	Ансамблирование классификаторов	1	2	4	-	14	Практическая работа №2 Тест № 1
5	Методы кластерного анализа. Алгоритмы машинного обучения: метод k-средних, EM, Cobweb	1	2	4	-	14	Практическая работа №3 Тест № 1
6	Методы анализа ассоциаций и последовательностей	1	2	4	-	14	Практическая работа №4 Тест № 1
7	Введение в методы сетевого анализа (Social Network Analysis)	1	2	4	-	14	Практическая работа №5 Тест № 1
Модуль 2. Прикладной анализ данных							

№ п/п	Темы дисциплины	Семестр	Виды учебной работы и их трудоёмкость, часы (в том числе с использованием онлайн-курсов)				Наименования оценочных средств
			Контактная работа			Самостоя-тельная работа	
			Лекции	Практические занятия	Лабораторные занятия		
8	Машинное обучение и большие данные для решения прикладных задач	1	2	8	-	14	Тест № 2
9	Персонализация методами машинного обучения	1	2	-	-	14	Тест № 2
Итого часов			**18**	**36**	**-**	**126**	–

4.2. План внеаудиторной самостоятельной работы

№ п/п	Темы дисциплины	Семестр	Вид самостоятельной работы	Сроки выполнения (нед.)	Затраты времени (часы)	Учебно-методическое обеспечение
Модуль 1. Математическое и алгоритмическое обеспечение аналитики больших данных						
1	Введение в аналитику больших данных (Big Data). Обзор задач интеллектуального анализа данных: регрессия, классификация, кластеризация, анализ ассоциаций и последовательностей, поиск аномалий, анализ связей.	1	– проработка и повторение материала лекционных занятий; – подготовка к практическим занятиям	1–2	14	[1]-[4]
2	Методы классификации. Алгоритмы машинного обучения: деревья решений, опорные вектора, байесовские классификаторы	1	– проработка и повторение материала лекционных занятий; – подготовка к практическим занятиям	3-4	14	[1]-[4]
3	Оценка эффективности и сравнительный анализ моделей обучения	1	– проработка и повторение материала лекционных занятий;	5-6	14	[1]-[4]
4	Ансамблирование классификаторов	1	– проработка и повторение материала лекционных занятий; – подготовка к практическим занятиям	7-8	14	[1]-[4]
5	Методы кластерного анализа. Алгоритмы машинного обучения: метод k-средних, EM, Cobweb	1	– проработка и повторение материала лекционных занятий; – подготовка к практическим занятиям	9-10	14	[1]-[4]
6	Методы анализа ассоциаций и последовательностей	1	– проработка и повторение материала лекционных занятий;	11-12	14	[1]-[4]

№ п/п	Темы дисциплины	Семестр	Вид самостоятельной работы	Сроки выполнения (нед.)	Затраты времени (часы)	Учебно-методическое обеспечение
7	Введение в методы сетевого анализа (Social Network Analysis)	1	– проработка и повторение материала лекционных занятий; – подготовка к практическим занятиям	13-14	14	[1]-[4]
Модуль 2. Прикладной анализ данных						
8	Машинное обучение и большие данные для решения прикладных задач	1	– проработка и повторение материала лекционных занятий; – подготовка к практическим занятиям	15-16	14	[1]-[4]
9	Персонализация методами машинного обучения	1	– проработка и повторение материала лекционных занятий; – подготовка к практическим занятиям	17-18	14	[1]-[3]
Общая трудоёмкость самостоятельной работы по дисциплине					**126**	**–**

4.3. Содержание учебного материала

Модуль 1. Математическое и алгоритмическое обеспечение аналитики больших данных

Введение в аналитику больших данных (Big Data). Обзор задач интеллектуального анализа данных: регрессия, классификация, кластеризация, анализ ассоциаций и последовательностей, поиск аномалий, анализ связей. Понятие Data Mining. OLAP и Data Mining. Данные, информация и знания. Единая методология обнаружения знаний. Задача регрессии. Обучение с учителем. Задача классификации. Обучение без учителя. Задача кластеризации. Задача анализа ассоциаций и последовательностей. Программное обеспечение для интеллектуального анализа данных: Weka, R, SAS Enterprise Miner, SPSS Modeler. Большие данные. Масштабируемые алгоритмы. Hadoop, MapReduce. Microsoft Azure Machine Learning, RapidMiner.

Методы классификации. Алгоритмы машинного обучения: деревья решений, опорные вектора, байесовские классификаторы. Индукция деревьев решений. Информационный выигрыш. Индекс Gini. «Обрезка» деревьев: предредукция и постредукция. Решающие правила. Алгоритмы ID3, CART, C4.5. Алгоритм «случайный лес». Алгоритмы ограниченного перебора. Метод опорных векторов, линейная и нелинейная разделимость. Байесовская и наивная байесовская классификация.

Оценка эффективности и сравнительный анализ моделей обучения. Подготовка данных. Выбор значимых признаков. Наборы данных. Типы данных. Шкалы измерений. Форматы хранения данных. Качество данных. Очистка данных. Работа с дубликатами и пропущенными значениями. Снижение размерности данных. Интеграция данных. BI и визуализация данных. Методы отбора значимых признаков. Фильтры. Оболочки. Встроенные методы. Матрица несоответствий. Метрики качества: правильность, полнота, точность, F-мера, чувствительность, специфичность. Обучающее множество. Независимое тестовое множество. Подтверждающее множество. Проблема переобучения. Метод удержания. Метод перекрестной проверки. Метод «без одного». Стратификация данных. Метод самонастройки (бутстреп). Матрица стоимостей ошибок. Диаграмма выигрыша. Диаграмма роста. Кривая ошибок (ROC-кривая). AUC. Изолинии точности.

Ансамблирование классификаторов. Ансамбли (комитеты) моделей. Бэггинг. Бэггинг с рандомизацией. Последовательно обучающиеся классификаторы. Бустинг (усиление) ансамбля классификаторов. Алгоритм AdaBoost. Стэкинг.

Методы кластерного анализа. Алгоритмы машинного обучения: метод k-средних, EM, Cobweb. Типологический и таксономический анализ. Статистические методы кластеризации. EM-алгоритм. Метод k-средних. Меры расстояний. Иерархические методы кластеризации. Визуализация кластеров. Дендрограммы. Диаграммы рассеивания. Самоорганизующиеся карты Кохонена. Концептуальная кластеризация. Алгоритм Cobweb. Графовые методы кластеризации. Выделение связных компонент. Нечеткая кластеризация. FCM-алгоритм.

Методы анализа ассоциаций и последовательностей. Поиск часто встречающихся наборов элементов. Меры интересности: поддержка, достоверность, лифт. уверенность. Алгоритм Apriori. Задача анализа рыночных корзин. Количественные и нечеткие ассоциативные правила. Поиск последовательных шаблонов.

Введение в методы сетевого анализа (Social Network Analysis). Анализ связей. Понятия социальной сети и социального графа. Теория «малого мира». Диаметр, радиус, коэффициент кластеризации. Центральность: по степени, по близости, на основе собственных векторов. Позиционный и ролевой анализ в социальной сети.

Модуль 2. Прикладной анализ данных

Машинное обучение и большие данные для решения прикладных задач. Анализ рыночных корзин, «умные» рекламные объявления, совместная фильтрация, анализ тональности, чат-боты и др.

Персонализация методами машинного обучения. Массовая кастомизация, индивидуальный маркетинг. Методы и техники персонализации в интернете. Полуавтоматическая и автоматическая персонализация. Понятие адаптивного веб-ресурса. Цели, методы и приемы адаптации. Архитектура адаптивного веб-ресурса. Оверлейная и стереотипная модели пользователя. Формальная постановка задачи персонализации. Оценка эффективности рекомендаций. Рекомендации на основе композиционного правила нечеткого вывода. Методы совместной фильтрации. Теоретико-графовый метод хортинга.

Перечень тем практических занятий

№ п/п	Тема практического занятия	Количество часов
Модуль 1. Математическое и алгоритмическое обеспечение аналитики больших данных		
1	Введение в аналитику больших данных (Big Data). Обзор задач интеллектуального анализа данных: регрессия, классификация, кластеризация, анализ ассоциаций и последовательностей, поиск аномалий, анализ связей.	4
2	Методы классификации. Алгоритмы машинного обучения: деревья решений, опорные вектора, байесовские классификаторы	4
3	Оценка эффективности и сравнительный анализ моделей обучения	4
4	Ансамблирование классификаторов	4
5	Методы кластерного анализа. Алгоритмы машинного обучения: метод k-средних, EM, Cobweb	4
6	Методы анализа ассоциаций и последовательностей	4
7	Введение в методы сетевого анализа (Social Network Analysis)	4
Модуль 2. Прикладной анализ данных		
8	Машинное обучение и большие данные для решения прикладных задач	4
9	Персонализация методами машинного обучения	4
Всего часов		**36**

V. ОБРАЗОВАТЕЛЬНЫЕ ТЕХНОЛОГИИ

Наряду с традиционными образовательными технологиями, для реализации дисциплины будут использоваться технологии электронного обучения и дистанционные образовательные технологий в электронной информационно-образовательной среде университета.

Активные формы обучения, применяемые на практических занятиях, способствуют разнообразному (индивидуальному, групповому, коллективному) изучению (усвоению) учебных вопросов (проблем), активному взаимодействию обучающихся и преподавателя, живому обмену мнениями между ними, нацеленному на выработку правильного понимания содержания изучаемой темы и способов ее практического использования.

Аудиторные занятия и другие формы контактной работы обучающихся с преподавателем могут проводиться с использованием платформ Microsoft Teams, Moodle (BigBlueButton) и др., что позволяет обеспечить онлайн и офлайн взаимодействие преподавателя с обучающимися в рамках дисциплины

Основными методами текущего контроля являются электронный учёт и контроль учебных достижений студентов (использование средств сервиса балльно-рейтинговой системы; ведение электронного журнала успеваемости, проведение электронного тестирования и применение других средств контроля с использованием системы электронного обучения).

VI. УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

6.1. Основная литература

1. Келлехер, Д. Наука о данных: базовый курс : [16+] / Д. Келлехер, Б. Тирни ; науч. ред. З. Мамедьяров ; пер. с англ. М. Белоголовского. – Москва : Альпина Паблишер, 2020. – 224 с. : схем., табл. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=598235

2. Маккинли, Уэс Python и анализ данных / Уэс Маккинли ; перевод А. Слинкина. — 2-е изд. — Саратов : Профобразование, 2019. — 482 c. — ISBN 978-5-4488-0046-7. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. — URL: https://www.iprbookshop.ru/88752.html

3. Ракитский, А. А. Методы машинного обучения : учебно-методическое пособие / А. А. Ракитский. — Новосибирск : Сибирский государственный университет телекоммуникаций и информатики, 2018. — 32 c. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. — URL: https://www.iprbookshop.ru/90591.html

4. Чубукова, И. А. Data Mining : учебное пособие / И. А. Чубукова. — 3-е изд. — Москва, Саратов : Интернет-Университет Информационных Технологий (ИНТУИТ), Ай Пи Ар Медиа, 2020. — 469 c. — ISBN 978-5-4497-0289-0. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. — URL: https://www.iprbookshop.ru/89404.html

6.2. Дополнительная литература

5. Крутиков В. Н. Анализ данных / В.Н. Крутиков; В.В. Мешечкин - Кемерово: Кемеровский государственный университет, 2014. - 138 с. http://biblioclub.ru/index.php?page=book&id=278426

6. Жуковский О. И. Информационные технологии и анализ данных: учебное пособие / О.И. Жуковский - Томск: Эль Контент, 2014. - 130 с. http://biblioclub.ru/index.php?page=book&id=480500

6.3. Периодические издания

- IEEE Spectrum https://spectrum.ieee.org/
- Научный журнал «Машинное обучение и анализ данных» http://jmlda.org/ru/journal

6.4. Перечень ресурсов сети Интернет

- ЭБС IPR Books http://www.iprbookshop.ru/
- ЭБС «Университетская библиотека онлайн» http://biblioclub.ru.
- Образовательная платформа Юрайт https://urait.ru/
- IBM Academic Initiative http://ictis.sfedu.ru/ibm_academic_initiative/ (учебные материалы)
- http://github.com/
- http://habr.com/
- http://www.kdnuggets.com/
- Python, Свободное ПО, https://www.python.org/
- https://www.jetbrains.com/pycharm/

VII. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Лекционная аудитория
Мультимедийный проектор, экран

Компьютерный класс
Интерактивная доска с проектором; персональные компьютеры, ПО Microsoft Windows, Microsoft Office (Microsoft Teams), актуальные версии браузеров Chrome, Firefox, Edge, Safari с поддержкой протокола WebRTC, WEKA

VIII. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ

Дисциплина включает в себя лекционные и практические занятия, а также самостоятельную работу обучающихся.

Организация образовательного процесса по дисциплине осуществляется с использованием системы электронного обучения.

Все лекционные занятия проводятся с визуализацией учебного материала в форме презентаций лекционного материала, которые доступны в системе электронного обучения.

Лекционная часть курса включает следующие компоненты системы знаний учебной дисциплины: понятийный аппарат (тезаурус курса), теоретические утверждения, разъяснения и комментарии; междисциплинарные точки зрения; описание рассматриваемых разделов; ретроспективный и перспективный взгляды на изучаемую проблематику.

Практические занятия по всем модулям дисциплины требуют предварительной теоретической подготовки по соответствующим темам: проработка лекционного материала, ознакомление и изучение отдельных источников основной и дополнительной литературы.

Лекционные и практические занятия могут проводиться с применением дистанционных образовательных технологий с использованием платформ Microsoft Teams, Cisco, Moodle (BigBlueButton) и др.

Проведение лекционных и практических занятий осуществляется с постановкой проблемных вопросов, допускающих возникновение дискуссий, что предполагает активное включение студентов в образовательный процесс.

В организации процесса обучения используются как традиционные, характерные лекционно-семинарской форме обучения, так и инновационные (интерактивные, имитационные, проектные) технологии.

Используемые технологии обеспечивают:

– формирование компетенций, осознанное усвоение знаний, качественное освоение умений их применять и формирование заинтересованного отношения к изучаемым объектам в единстве;

– продуктивность познавательной деятельности, научный поиск, создание субъективно и объективно новых знаний или других продуктов;

– ориентацию на студентов, стимулирование их активности, самостоятельности, инициативы и ответственности;

– контекстный характер обучения, то есть привязку к реальным профессиональным задачам;

– вовлеченность студентов в выполняемую деятельность, возможность проявить и развить свой интеллектуальный, творческий, личностный, деловой потенциал.

Самостоятельная работа направлена на повышение качества обучения, углубление и закрепление знаний студента, развитие аналитических навыков по проблематике учебной дисциплины, активизацию учебно-познавательной деятельности студентов и снижение аудиторной нагрузки.

Максимальное количество баллов по каждому виду контрольных мероприятий указано в учебной карте дисциплины.

IX. УЧЕБНАЯ КАРТА ДИСЦИПЛИНЫ

Курс 1, семестр 1, очная форма обучения

№ п/п	Виды контрольных мероприятий (наименования оценочных средств)	Количество баллов	
		Текущий контроль	Рубежный контроль
Модуль 1. Математическое и алгоритмическое обеспечение аналитики больших данных			
1	Практическая работа № 1	12	–
2	Практическая работа № 2	12	–
3	Практическая работа № 3	12	–
4	Практическая работа № 4	12	–
5	Практическая работа № 5	12	–
6	Тест № 1	–	20
Модуль 2. Прикладной анализ данных			
7	Тест № 2	–	20
Всего		**60**	**40**
Бонусные баллы		Бонусные баллы начисляются за победу и призовые места в соревновании классификаторов, достижения в конкурсах (Kaggle и др.), активное участие в конференциях и выставках – до 10 баллов.	
Промежуточная аттестация в форме дифференцированного зачёта		Оценка по дисциплине выставляется по сумме баллов за текущий контроль и рубежный контроль: – 85–100 баллов – оценка «отлично»; – 71–84 балла – оценка «хорошо»; – 60–70 баллов – оценка «удовлетворительно»; – менее 60 баллов – оценка «неудовлетворительно»	

X. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

10.1. Паспорт фонда оценочных средств

№ п/п	Индикатор достижения компетенции	Наименование оценочного средства
1	ПК-1.1. Ставит задачи по адаптации или совершенствованию методов и алгоритмов для решения комплекса задач предметной области	– практические работы №1-№2 – тест №1
2	ПК-2.1. Руководит разработкой архитектуры комплексных систем искусственного интеллекта со стороны заказчика	– практические работы №3-№5 – тест №2
3	ПК-2.2. Осуществляет руководство созданием комплексных систем искусственного интеллекта с применением новых методов и алгоритмов машинного обучения	– практические работы №1-№2, №3-№5 – тест №1 – тест №2

10.2. Практическая работа №1

Классификация с использованием деревьев решений

Классификация – это упорядоченное по некоторому принципу множество объектов, которые имеют сходные классификационные признаки (одно или несколько свойств), выбранные для определения сходства или различия между этими объектами.

Целью классификации является построение оптимальной модели классификации, которая использует прогнозирующие атрибуты (вектор признаков) в качестве входных параметров для получения значения зависимого атрибута. Критерием оптимальности в нашем случае является показатель *точности*, рассчитываемый как отношение числа верно классифицированных примеров к их общему числу в выборке. При этом модель должна с высокой точностью классифицировать новые, ранее не рассмотренные примеры.

Для проведения классификации набор исходных данных разбивают на два множества: обучающее и тестовое. Оба множества содержат входные и выходные (целевые) значения атрибутов. В обучающем множестве (*training set*) выходные значения зависимых атрибутов предназначены для обучения (конструирования) модели. В тестовом множестве (*test set*) выходные значения используются для проверки работоспособности модели.

Подходы к оценке эффективности модели: на обучающем множестве (*Use training set*), на тестовом множестве (*Percentage split*), метод k-перекрестной проверки (*Cross-validation*).

В задачах обучения с учителем важно предупреждать *проблему переобучения* – нежелательное побочное явление, при котором ошибка алгоритма на обучающей выборке снижается, а на тестовой выборке растет.

Алгоритмы классификации. Для построения модели классификации в среде Weka используйте следующие алгоритмы на основе деревьев решений (в скобках приводятся параметры алгоритмов, требующие настройки):

- zeroR
- oneR
- Id3
- J4.8/C4.5 (unpruned, minNumObj, reducedErrorPruning, numFolds)
- JRip

Задание.

1. Используйте файлы данных в формате arff согласно вариантам заданий (табл. 1).

Таблица 1

Варианты заданий

Вариант 1 (Первая буква фамилии А-Ж)	Вариант 2 (Первая буква фамилии З-М)	Вариант 3 (Первая буква фамилии Н-С)	Вариант 4 (Первая буква фамилии Т-Я)
Contact-lenses.arff	Iris.arff	Labor.arff	Zoo.arff

2. Изучите файлы данных. Для каждого набора данных поочередно примените все доступные алгоритмы классификации с параметрами по умолчанию. Оцените точность классификаторов, используя обучающее множество, тестовое множество (2:1) и метод 10-перекрестной проверки. Объясните различие в результатах.
3. На основе результатов 10-перекрестной проверки сравните классификаторы между собой и отранжируйте по убыванию точности.
4. Настройте алгоритм J4.8 с предредукцией. Для этого установите значение параметра unpruned True. Далее произвольно изменяйте значения параметра minNumObj (минимальное число вершин в листьях) в диапазоне от 1 до 15. Оцените точность классификаторов. Сравните с результатами алгоритма J4.8, полученными на предыдущем этапе.
5. Используйте алгоритм J4.8 с постредукцией. Для этого измените значение параметра unpruned на False, а значение параметра reducedErrorPruning – на True. Далее произвольно изменяйте значение параметра numFolds в диапазоне от 2 до числа примеров в базе данных. Тем самым, вы определяете объем данных, используемых для постредукции: одна часть используется для обрезки, остальные – для выращивания дерева решений. Оцените точность классификаторов. Сравните с результатами J4.8, полученными на предыдущих этапах.
6. На результатах п.4-5 проиллюстрируйте наглядным примером проблему переобучения.
7. По итогам работы определите оптимальные классификаторы для каждого набора данных.

Критерии оценки:

11-12 баллов выставляется студенту, если он своевременно выполнил все задачи, предусмотренные в практической работе, подготовил отчет в соответствии с требованиями преподавателя и в процессе защиты продемонстрировал наличие теоретических знаний в объеме содержания учебной дисциплины, относящейся к практической работе. Сумел ответить на дополнительные вопросы, связанные не только с процессом выполнения практической работы, но и с пониманием совершенных действий и решенных задач.

9-10 баллов выставляется студенту, если он выполнил от 70% задач, предусмотренных в практической работе, подготовил отчет в соответствии с требованиями преподавателя и в процессе защиты продемонстрировал наличие теоретических знаний в объеме содержания учебной дисциплины, относящейся к практической работе. Сумел ответить на вопросы, связанные с процессом выполнения практической работы.

7-8 баллов выставляется студенту, если он более чем на 50% выполнил поставленные в практической работе задачи, способен ответить на вопросы, касающиеся теоретической составляющей в объеме содержания учебной дисциплины, относящейся к практической работе.

0 баллов выставляется студенту, если он более чем на 50% не выполнил поставленные в практической работе задачи, не способен ответить на вопросы, касающиеся теоретической составляющей в объеме содержания учебной дисциплины, относящейся к практической работе.

10.3. Практическая работа №2

Классификация текстов

1. **Введение**

Интеллектуальный анализ текстов (*text mining*) связан с задачей выделения релевантной информации из текстов на естественном языке и поиском в ней закономерностей. Классификация текстов – это пример сложной проблемы интеллектуального анализа данных, в которой мы имеем дело с данными большой размерности и неоднородными шаблонами искомых данных.

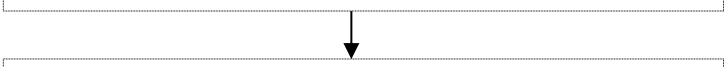
2. **Классификация текстов в среде Weka**

Для представления текстовой информации в Weka используется строковый атрибут типа STRING. Такой атрибут может содержать огромное число значений и поэтому напрямую классификаторы с ним не работают. В исходных данных всего два атрибута: атрибут класса и строковый атрибут. Мы преобразуем значения STRING в множество атрибутов, представляющих частоту встречаемости слов в строке. Такое преобразование можно реализовать с помощью фильтра StringToWordVector (рис. 1). На первой вкладке нажмите кнопку Choose, найдите указанный фильтр и примените его с помощью кнопки Apply.

```
@relation text

@attribute docs string
@attribute class {animals, not_animals}

@data
'cat mouse lion', animals
'tree flower', not_animals
```



```
@relation text-converted

@attribute class {animals, not_animals}
@attribute cat numeric
@attribute mouse numeric
@attribute lion numeric
@attribute tree numeric
@attribute flower numeric

@data
animals,1,1,1,0,0
not_animals,0,0,0,1,1
```

Рис. 1

Обратите внимание, что после преобразования атрибут класса обычно становится первым по счету. Однако в Weka по умолчанию атрибутом класса выбирается последний атрибут. Уточнить атрибут класса можно в выпадающем меню на вкладке Classify.

После преобразования можно начать обучение и сравнение классификаторов.
В этой лабораторной работе мы будем использовать четыре классификатора:
- алгоритм ближайшего соседа Nearest Neighbor (*IBk*),
- байесовский Naïve Bayes (*NB*),
- алгоритм на основе деревьев решений *J48*,

- алгоритм на основе решающих правил *JRip*.

3. Вавилонская башня

3.1 Описание

Дано:

- Коллекция из 189 предложений об интеллектуальном анализе данных из Wikipedia (не перевод!) на английском, французском, испанском и немецком языках.

Найти:

- Классификаторы, которые распознают язык текста.

Данные содержатся в файле dataMining.arff

3.2 Задание

Перейдите от строковой переменной к множеству атрибутов как описано выше.

Обучите классификаторы и сравните их между собой методом 10-перекрестной проверки. *Какой алгоритм лучший?* Выпишите статистики точности, изучите матрицы несоответствий и деревья решений классификаторов *J48* и *JRip*. Используйте эту информацию для ответа на следующие вопросы:

Являются ли языки, представленные в тексте, родственными? Как вы можете это подтвердить?

Классификатор IBk показывает невысокую точность для маленьких значений параметра k. Когда вы изменяете значение (например, увеличиваете значение k до 59), точность существенно растет. Как вы это объясните?

Ввиду высокой размерности данных (1902 числовых атрибута), обучение некоторых классификаторов (*J48, JRip*) занимает много времени. Для уменьшения времени работы алгоритма вы можете сократить число атрибутов, оставив только наиболее релевантные. Это можно сделать с помощью одного из методов выбора атрибутов, представленных на вкладке Select Attributes. Перейдите на указанную вкладку; убедитесь, что в выпадающем меню выбран алгоритм класса; запустите процедуру с параметрами по умолчанию. После вычисления оценок значимости атрибутов, вы можете удалить ненужные атрибуты из набора данных, используя фильтр Remove. Для этого нажмите Choose; выберите указанный фильтр; в настройки фильтра вставьте номера отобранных атрибутов и допишите 1 через запятую; измените false на true для инверсии, чтобы оставить только значимые атрибуты и атрибут класса; нажмите Apply. Повторите эксперименты и выпишите значения статистик.

Как вы видите, кроме сокращения времени работы, классификатор IBk повысил свою точность при малых значениях k. *Как вы это объясните?*

4. Музыканты шутят

4.1 Описание

Дано:

- Коллекция из 409 шуток о музыке на английском языке с сайта http://www.mit.edu/~jcb/jokes/.

Шутки помечены как смешные и не смешные.

Найти:

- Классификаторы, которые распознают, смешная шутка или не смешная.

Данные представлены в файле musicJokes.arff

4.2 Задание

Обучите и сравните между собой классификаторы. Выпишите статистики и изучите матрицы несоответствий классификаторов.

Применяя фильтр StringToWordVector, вы сначала получите ошибку: название переменной class совпадает с названием одного из атрибутов (слов в тексте). Нажмите на слово StringToWordVector и в параметрах фильтра выберите настройку stopwordsHandler – Choose – WordsFromFile. Нажмите на слово WordsFromHandler, затем на stopwords. Создайте текстовый файл stopwords.txt, в котором содержится одно слово class. Укажите путь к этому файлу. Теперь вы можете применить фильтр.

Используя визуализацию деревьев решений J48 и JRip, можете ли вы предположить, что делает шутку смешной?

Поскольку ни один из классификаторов не имеет высокой точности, попробуйте улучшить работу классификаторов, выбрав наиболее релевантные атрибуты, используя один из методов на вкладке Select Attributes. *Какой алгоритм лучший?*

5. Американский и китайский английский

5.1 Описание

Дано:

- Коллекция из 132 пар предложений на английском, написанных американцами и китайцами, изучающими английский (http://www.englishdaily626.com/c-mistakes.php).

Найти:

- Классификаторы, которые распознают, написано предложение на английском американцем или китайцем.

Данные представлены в файле americanChineseStyles.arff

5.2 Задание

Обучите и сравните между собой классификаторы. Осуществите отбор важных атрибутов.

Точность классификаторов не превышает 50%. *Можно ли, используя классификаторы с невысокой точностью, получить приемлемые результаты классификации? Каким классификатором пользоваться?*

Критерии оценки:

11-12 баллов выставляется студенту, если он своевременно выполнил все задачи, предусмотренные в практической работе, подготовил отчет в соответствии с требованиями преподавателя и в процессе защиты продемонстрировал наличие теоретических знаний в объеме содержания учебной дисциплины, относящейся к практической работе. Сумел ответить на дополнительные вопросы, связанные не только с процессом выполнения практической работы, но и с пониманием совершенных действий и решенных задач.

9-10 баллов выставляется студенту, если он выполнил от 70% задач, предусмотренных в практической работе, подготовил отчет в соответствии с требованиями преподавателя и в процессе защиты продемонстрировал наличие теоретических знаний в объеме содержания учебной дисциплины, относящейся к практической работе. Сумел ответить на вопросы, связанные с процессом выполнения практической работы.

7-8 баллов выставляется студенту, если он более чем на 50% выполнил поставленные в практической работе задачи, способен ответить на вопросы, касающиеся теоретической составляющей в объеме содержания учебной дисциплины, относящейся к практической работе.

0 баллов выставляется студенту, если он более чем на 50% не выполнил поставленные в практической работе задачи, не способен ответить на вопросы, касающиеся теоретической составляющей в объеме содержания учебной дисциплины, относящейся к практической работе.

10.4. Практическая работа №3

Кластеризация

В настоящей лабораторной работе вы проведете эксперименты с тремя алгоритмами кластеризации.

1. **Алгоритм k-средних**

Алгоритм k-средних – это простой, прямой алгоритм кластеризации. Каждый кластер определяется центроидом, и экземпляры принадлежат к тому кластеру, евклидово расстояние до центроида которого минимально. Затем для каждого кластера находят новый центроид как среднее арифметическое наблюдений в кластере, что может привести к перестановкам экземпляров между кластерами. Итеративный процесс завершается, когда центроиды перестают меняться.

Загрузите файл “marble.arff”. Набор данных содержит описание 15 шаров, использованных в качестве примера на лекции. После загрузки файла удалите атрибут «цвет» («color»), для чего выберите его и нажмите кнопку «Удалить» («Remove»). Обратите внимание, что атрибут «класс» («class») содержит естественные названия групп, обсуждавшихся на лекции. Проверим теперь, сможет ли алгоритм k-means найти эти группы.

Перейдите на вкладку «Кластеризация» («Cluster»). Выберите алгоритм SimpleKMeans. Поскольку известны четыре естественных кластера, используйте 4 или 5 в качестве значения параметра k. Убедитесь, что в качестве режима кластеризации выбран «Classes to clusters evaluation».

2.1 Поскольку алгоритм k-средних находит локальный минимум, обычно рекомендуется запускать алгоритм несколько раз и использовать наилучший из

результатов. Проведите эксперимент и изучите результаты. Насколько близко вы смогли приблизиться к «идеальному» разделению шаров на классы?

2. Алгоритм Cobweb

Данный алгоритм строит кластеры, пошагово добавляя элементы к дереву и включая их в существующий кластер, если это приводит к увеличению значения функции полезности, по сравнению с определением элемента в новый кластер. По необходимости существующий кластер может быть разбит на два новых кластера, если это положительно скажется на значении функции полезности. Результирующее множество кластеров называется «дендрограммой».

На вкладке «Preprocess» загрузите файл “marblespecific.arff” и удалите атрибут «цвет» («color»). Файл «marblespecific.arff» отличается от предыдущего набора данных в том плане, что атрибут «класс» («class») является уникальным для каждого шара, что позволяет распознавать шары в отчете Weka. Код для каждого шара состоит из четырех букв и цвета шара. В свою очередь, четыре буквы объясняют размер (Big / Small), расцветку (Monochrome / Polychrome), блеск (Shiny / Dull) и прозрачность (Transparent / Opaque).

Перейдите на вкладку «Кластеризация». Выберите алгоритм Cobweb. Установите флаг «Classes to clusters evaluation», чтобы алгоритм игнорировал значение атрибута «класс», но включил этот атрибут в финальный отчет для сравнения. Произведите кластеризацию.

3.1 Изучите результаты на основе визуализации дерева («Visualize tree»). Вы удовлетворены результатами кластеризации? Почему?

Увеличение значения параметра Cutoff алгоритма будет способствовать отнесению похожих шаров в один кластер.

3.2 Поэкспериментируйте со значением параметра Cutoff, пока вы не будете удовлетворены результатами кластеризации.

3. Алгоритм максимизации ожидания

Алгоритм EM – это вероятностный алгоритм кластеризации. Каждый кластер определяется вероятностями элементов обладать некоторыми конкретными значениями атрибутов и вероятностями принадлежности каждого элемента к кластеру. Для числовых значений это значение среднего и стандартное отклонение для каждого значения атрибута, для дискретных значений это вероятность для каждого значения атрибута.

Поскольку с дискретными значениями проще работать, а также для сравнения с результатами двух предыдущих алгоритмов, применим алгоритм EM к тому же набору данных. На вкладке «Preprocess» загрузите файл “marble.arff” и удалите атрибут «цвет» («color»).

Перейдите на вкладку «Кластеризация». Выберите алгоритм EM. Установите флаг «Classes to clusters evaluation». Произведите кластеризацию.

4.1 Изучите результаты. Сколько кластеров получилось? Почему? Можете ли вы получить другой результат с другим значением параметра seed?

Значение параметра numClusters по умолчанию -1, что позволяет алгоритму самостоятельно определять число кластеров. Если указать конкретное значение, алгоритм постарается обнаружить необходимое число кластеров.

4.2 Измените значение параметра на число кластеров, которое вы хотите получить (или немного больше) и повторите эксперимент. Проведите эксперименты с различными значениями seed. Как это влияет на результаты?

При изучении отчета вы увидите базовую вероятность кластера и значение вероятности для каждого атрибута, которую можно получить делением значения дискретной оценки (“discrete estimator”) на сумму оценок по каждому атрибуту («total»), например:

Атрибут: размер (size)
Discrete Estimator. Counts = 2.41 11.86 (Total = 14.26)

означает, что вероятность значения «большой» (big) для атрибута «размер» составит 2.41/14.26 = 0.169.

Порядок значений атрибутов следующий:

“size” {big, small}
“colouring” {monochrome, polychrome}
“shininess” {shiny, dull}
“transparency” {transparent, opaque}

Чтобы определить «лучший» кластер для шара, нужно перемножить базовую вероятность кластера и вероятности для каждого из значений атрибутов. Нормализованное число даст ожидаемую вероятность шара быть отнесенным к каждому из кластеров.

Для следующего эксперимента установите значение параметра seed равным 14, а число кластеров – 2. Произведите кластеризацию.

4.3 Пусть у вас имеется маленький, одноцветный, неяркий, непрозрачный шар. В каком кластере вы ожидаете его увидеть? (Для расчетов используйте стандартное приложение «Калькулятор».)

Заключение

Вы провели эксперименты с тремя алгоритмами на одном наборе данных.

5.1 Приведите (как минимум) одно преимущество и один недостаток каждого алгоритма.

5.2 Какой из трех алгоритмов кластеризации вы предпочитаете использовать для набора данных «Шары».

Критерии оценки:

11-12 баллов выставляется студенту, если он своевременно выполнил все задачи, предусмотренные в практической работе, подготовил отчет в соответствии с требованиями преподавателя и в процессе защиты продемонстрировал наличие теоретических знаний в объеме содержания учебной дисциплины, относящейся к практической работе. Сумел ответить на дополнительные вопросы, связанные не только с процессом выполнения практической работы, но и с пониманием совершенных действий и решенных задач.

9-10 баллов выставляется студенту, если он выполнил от 70% задач, предусмотренных в практической работе, подготовил отчет в соответствии с требованиями преподавателя и в процессе защиты продемонстрировал наличие теоретических знаний в объеме содержания учебной дисциплины, относящейся к практической работе. Сумел ответить на вопросы, связанные с процессом выполнения практической работы.

7-8 баллов выставляется студенту, если он более чем на 50% выполнил поставленные в практической работе задачи, способен ответить на вопросы, касающиеся теоретической составляющей в объеме содержания учебной дисциплины, относящейся к практической работе.

0 баллов выставляется студенту, если он более чем на 50% не выполнил поставленные в практической работе задачи, не способен ответить на вопросы, касающиеся теоретической составляющей в объеме содержания учебной дисциплины, относящейся к практической работе.

10.5. Практическая работа №4

Ассоциативные правила

Rules Wizard (разработчик BaseGroup Labs) – простое в использовании приложение, решающее задачи поиска ассоциативных связей (Association Rules). Программа позволяет выявлять ассоциативные правила, используя таблицу транзакций и представлять их разными способами - в виде форматированного текста, дерева, перекрестной и обычной таблицы.

Ассоциативные связи – это зависимости вида: если произошло событие A, то с определенной вероятностью произойдет событие B. Примером такого правила, служит утверждение, что покупатель, приобретающий Хлеб, приобретет и Молоко с вероятностью 75%.

Впервые задача поиска ассоциативных правил была предложена для нахождения типичных шаблонов покупок, совершаемых в супермаркетах, поэтому иногда ее еще называют анализом рыночной корзины (market basket analysis). Наибольшее использование ассоциативные правила находят в торговле, так как позволяют сказать, какой дополнительный товар может приобрести клиент, если он уже купил некоторые товары.

Задание: провести анализ рыночной корзины на основе набора данных о покупках в супермаркете (976 покупателей, 4852 покупки, 49 наименований товаров).

Объясните значение поддержки и достоверности для ассоциативного правила вида:

Если *Молоко* ***И*** *Печенье*
То *Чипсы*
Поддержка = 0,72%
Достоверность = 77,78%

1. Для БД транзакций из файла smarket.txt (в папке RulesWizard/Data) произведите анализ рыночной корзины со значениями поддержки s=0,3~100,0 и достоверности c=70,0~98,0. Изучите представления результатов в виде форматированного текста, дерева ассоциативных правил, перекрестной таблицы и таблицы правил. Зафиксируйте условия поиска, количество найденных правил и интервалы, на которых лежит значение поддержки и достоверности.

2. По дереву ассоциативных правил определите тройку товаров-бестселлеров. Перечислите все двухэлементные ассоциативные наборы-лидеры продаж (s>5%).

3. Для трех ассоциативных правил с низкой (<0,5%), средней (~1%) и высокой (>5%) степенью поддержки, соответственно, сделайте предположение, почему именно эти товары оказались в одной корзине. Составьте портрет-характеристику покупателя.

4. Используя представление в виде перекрестной таблицы, выясните, наличие каких товаров в корзине с большой вероятностью определит покупку ветчины, шоколада и кофе.

5. Подберите значение поддержки, при котором формулируется ~100 ассоциативных правил. Ознакомьтесь со статьей «Зачем купил - не знаю» (газета «Новые известия»). Предложите не менее 3 «интересных» ассоциативных наборов товаров, которые рекомендуется

размещать на одной полке или объединять в наборы со скидкой. Аргументируйте, почему эти ассоциативные правила можно считать «интересными».

Критерии оценки:

11-12 баллов выставляется студенту, если он своевременно выполнил все задачи, предусмотренные в практической работе, подготовил отчет в соответствии с требованиями преподавателя и в процессе защиты продемонстрировал наличие теоретических знаний в объеме содержания учебной дисциплины, относящейся к практической работе. Сумел ответить на дополнительные вопросы, связанные не только с процессом выполнения практической работы, но и с пониманием совершенных действий и решенных задач.

9-10 баллов выставляется студенту, если он выполнил от 70% задач, предусмотренных в практической работе, подготовил отчет в соответствии с требованиями преподавателя и в процессе защиты продемонстрировал наличие теоретических знаний в объеме содержания учебной дисциплины, относящейся к практической работе. Сумел ответить на вопросы, связанные с процессом выполнения практической работы.

7-8 баллов выставляется студенту, если он более чем на 50% выполнил поставленные в практической работе задачи, способен ответить на вопросы, касающиеся теоретической составляющей в объеме содержания учебной дисциплины, относящейся к практической работе.

0 баллов выставляется студенту, если он более чем на 50% не выполнил поставленные в практической работе задачи, не способен ответить на вопросы, касающиеся теоретической составляющей в объеме содержания учебной дисциплины, относящейся к практической работе.

10.6. Практическая работа №5

Анализ связей в социальных сетях

Целью лабораторного занятия является изучение базовых понятий и техники анализа социальных сетей на основе программы UCINET 6 компании Analytic Technologies, являющейся стандартом для специалистов в области анализа социальных сетей.

Теоретическая часть

Очень часто распределенные системы, в частности сети сотовой связи, компьютерные сети и Всемирная Паутина обладают сложной топологией и имеют в своей основе социальные процессы. Основным подходом к изучению таких структур является анализ социальных сетей.

Анализ социальных сетей – это методология и методы исследования взаимодействий между социальными объектами (акторами) и выявление условий возникновения этих взаимодействий.

Основными методами анализа социальных сетей являются методы теории графов.

Данные о связях акторов представляются в виде *социоматрицы* - квадратной или прямоугольной таблицы, элементы которой соответствуют показателям силы связи, исходящей от актора в *i*-й строке к актору в *j*-м столбце.

Всякой социоматрице может быть взаимнооднозначно сопоставлен граф. Связи в графе могут быть ненаправленными (ребра) и направленными (дуги). Граф с заданными на нем дугами называется ориентированным, или орграфом. Вершины, соединенные ребром, являются смежными. *Степенью вершины* называется число ребер, соединенные с ней. *Исходящей степенью вершины* называется число дуг, исходящих из вершины, *входящей степенью* - число дуг, входящих в вершину.

Последовательность вершин, соединенных ребрами, составляет *цепь*. В цепи направление связей между вершинами не имеет значения. В *простой цепи* ни одна из вершин и ни одно из ребер не повторяются. Число ребер цепи называется ее *длиной*. Длина самой короткой цепи, связывающей две вершины, называется *расстоянием* между вершинами (без учета направления связей).

Последовательность вершин, соединенных дугами, называется *путем* (направление связей существенно). В *простом пути* ни одна из вершин и ни одна из дуг не повторяются. Число дуг, составляющих путь, называется его *длиной*. Длину самого короткого пути, связывающего две вершины, называют *расстоянием* между ними (с учетом направления связей). Орграф, в котором из каждой вершины существует путь к любой другой вершине, называется *сильно связным* (путешествовать можно лишь по направлению дуг). Орграф, в котором существует цепь из каждой вершины к любой другой, называется *слабо связным* (можно путешествовать против направления дуг).

Плотность графа вычисляется как отношение числа существующих связей к потенциально возможному.

Показатели заметности. Идея центральности вершин в графе, их "важности" начала разрабатываться одной из первых в анализе социальных сетей. Источник этой идеи можно усмотреть в понятии "звезды" - самого популярного человека в группе. Будем называть меру заметности актора в сети (неориентированном графе) *центральностью*, для входящих связей в орграфе - *престижем*, для исходящих связей - *экспансивностью*.

Простой и интуитивно понятный подход к измерению центральности индивидов основывается на идее *степени*. Центральность на основе степени тем выше, чем больше число связей вершины с другими вершинами в графе. Индексы центральности по степени являются локальными характеристиками положения вершины в графе - они учитывают непосредственных соседей, ближайшую окрестность вершины и в этом смысле поверхностны. В отношении типа «дружба» престиж можно интерпретировать как популярность, а экспансивность как форму коммуникабельности.

Вторая группа показателей центральности основана на идее *близости* данной вершины ко всем остальным вершинам графа. Центральным является тот индивид, который быстро взаимодействует с другими либо непосредственно, либо через небольшое число посредников. Близость определяется как величина, обратная сумме длин самых коротких путей от данного индивида ко всем остальным. Доступная интерпретация близости - ожидаемое время движения ресурса (например, информации) от любого участника сети к данному индивиду. Центральность по близости является глобальной мерой сети. Недостаток показателя в том, что он не определен для изолированных вершин, поскольку при отсутствии связи между вершинами расстояние между ними бесконечно.

Взаимодействие двух несмежных индивидов может находиться под контролем возможных посредников. При поисках работы, например, важно не то, сколько знакомых у претендента, а сколько знакомых у этих знакомых. Метод оценки центральности по *посредничеству* для вершины заключается в нахождении доли самых коротких путей, соединяющих все пары вершин, которые проходят через данную вершину. Это сумма вероятностей того, что другие акторы в своих взаимодействиях будут прибегать к посредничеству данного актора.

Центральность на основе *собственных векторов* – более сложный показатель, учитывающий связи вершины с вершинами, имеющими большой вес. Примером такой меры центральности является параметр Google PageRank.

Показатели центральности, основанные на степени, информационно бедны. Центральность по посредничеству и центральность на основе собственных векторов предпочтительны в силу того, что они имеют большую изменчивость значений и более интересную интерпретацию.

Задание

1. На листке бумаге нарисуйте произвольный граф на 6 вершин и 8 неориентированных ребер (без изолированных вершин). Назовите вершины латинскими буквами или словами.
2. Представьте нарисованный граф в виде матрицы 6x6, в которой «1» соответствует ребру между вершинами, «0» - ребро отсутствует.
3. В Блокноте *на основе подготовленной матрицы* создайте исходный текстовый файл net1.txt в следующем формате (где n – число вершин):

dl n=**4** format=fullmatrix
data:
0 1 1 0
1 0 1 1
1 1 0 0
0 1 0 0

4. Запустите программу UCINET (Пуск > Программы > Analytic Technologies > Ucinet 6 for Windows).
5. Загрузите файл (Data > Import > DL). В первом поле (Input text file in DL format) укажите путь к исходному текстовому файлу и нажмите OK. Итоговый рабочий файл в формате UCINET будет иметь расширение *.##h. Откройте файл для просмотра (Data > Browse) и убедитесь, что импорт состоялся.
6. Запустите NetDraw (последняя иконка на панели инструментов UCINET) и откройте рабочий файл net1.##h (File > Open > Ucinet dataset > Network).
7. Перетаскивая вершины, приведите рисунок в соответствие с графом на бумаге. Сохраните рисунок в формате JPG (File > Save Diagram as > Jpeg) под именем net1.jpg.
8. Повторите шаги 3-7, но используйте теперь расширенный формат исходного файла net2.txt с названиями вершин:

dl n=**4**
labels:
A,B,C,D
data:
0 1 1 0
1 0 1 1
1 1 0 0
0 1 0 0

Файлы net1.txt, net1.##h, net1.jpg, net2.txt, net2.##h, net2.jpg служат отчетом по первой части работы.

Далее работа будет вестись с набором данных PADGETT (\Program Files\Analytic Technologies\Ucinet 6\DataFiles\PADGETT.##h) о социальных связях между 16 семействами Флоренции эпохи Возрождения. В файле содержатся две сети – супружеские и деловые связи. Мы будем работать с первой сетью. Для этого необходимо выделить сеть из набора, используя команду Data > Extract. В первом поле (Input dataset) укажите путь к исходному файлу, в поле Which matrices введите «1» и нажмите OK. В дальнейшем работайте с новым файлом PADGETT-Ext.##h

Отчет по второй части работы оформляется в виде файла net3.doc в формате Microsoft Word.

1. Изучите матрицу сети (Data > Display) и включите ее в отчет. (Используйте фиксированный шрифт, например Courier.)
2. Запустите NetDraw, загрузите сеть, сохраните рисунок в формате JPEG и включите его в отчет.
3. Вычислите плотность графа (NETWORK > COHESION > DENSITY). Проверьте правильность, рассчитав плотность вручную на основе матрицы из шага 1. Включите значение плотности и ход ручного расчета в отчет.
4. Вычислите расстояния между семействами (NETWORK > COHESION > DISTANCE). Проверьте правильность, рассчитав вручную расстояния между несколькими семействами на основе рисунка из шага 2. Включите матрицу Geodesic Distances и ход ручного расчета в отчет.
5. Вычислите и включите в отчет (только основные результаты – без статистик) следующие показатели центральности семейств:
    a. Центральность на основе степени (NETWORK > CENTRALITY > DEGREE).
    b. Центральность по посредничеству (NETWORK > CENTRALITY > FREEMAN BETWEENNESS > NODE BETWEENNESS).

    c. Центральность на основе собственных векторов (NETWORK > CENTRALITY > EIGENVECTOR).
6. Подготовьте для отчета таблицу, в которой 16 семейств будут отсортированы по убыванию показателя центральности на основе степени. Таблица – на 16 строк и 4 колонки (семейство, центральность на основе степени, центральность по посредничеству, центральность на основе собственных векторов). (Используйте Excel или инструмент Таблица в Word.)
7. Выберите произвольную пару семейств с одинаковой центральностью на основе степени, равной «4», но с совершенно разными оценками центральности по посредничеству. Используя социограмму из шага 2, предложите объяснение столько кардинальному различию в центральности по посредничеству. Сделайте то же самое для пары семейств с одинаковой центральностью на основе степени, равной «3». Отразите свой выбор и объяснение в отчете.
8. Выберите произвольную пару семейств с одинаковой центральностью на основе степени, равной «4», но с совершенно разными оценками центральности на основе собственных векторов. Используя социограмму из шага 2, предложите объяснение столь кардинальному различию в центральности на основе собственных векторов. Сделайте то же самое для пары семейств с одинаковой центральностью на основе степени, равной «3». Отразите свой выбор и объяснение в отчете.
9. Самоорганизация социальных сетей проявляется в виде сообществ – групп вершин с высокой плотностью ребер внутри группы и не высокой плотностью ребер между группами. Выделите сообщества в сети, используя следующий алгоритм [Girman, Newman, 2002]:
    a. Вычислите показатель посредничества для ребер в сети (NETWORK > CENTRALITY > FREEMAN BETWEENNESS > EDGE (LINE) BETWEENNESS).
    b. Удалите ребро с максимальным значением посредничества (DATA > BROWSE, обнуляем удаляемое ребро, FILE > SAVE)
    c. Изучите новый рисунок сети в NetDraw.

    Повторяйте шаги a)-c), пока граф не начнет распадаться на несвязные компоненты, постепенно обнаруживая сообщества. Включите в отчет 2-3 рисунка, наглядно иллюстрирующие процесс выделения сообществ.

Критерии оценки:

11-12 баллов выставляется студенту, если он своевременно выполнил все задачи, предусмотренные в практической работе, подготовил отчет в соответствии с требованиями преподавателя и в процессе защиты продемонстрировал наличие теоретических знаний в объеме содержания учебной дисциплины, относящейся к практической работе. Сумел ответить на дополнительные вопросы, связанные не только с процессом выполнения практической работы, но и с пониманием совершенных действий и решенных задач.

9-10 баллов выставляется студенту, если он выполнил от 70% задач, предусмотренных в практической работе, подготовил отчет в соответствии с требованиями преподавателя и в процессе защиты продемонстрировал наличие теоретических знаний в объеме содержания учебной дисциплины, относящейся к практической работе. Сумел ответить на вопросы, связанные с процессом выполнения практической работы.

7-8 баллов выставляется студенту, если он более чем на 50% выполнил поставленные в практической работе задачи, способен ответить на вопросы, касающиеся теоретической составляющей в объеме содержания учебной дисциплины, относящейся к практической работе.

0 баллов выставляется студенту, если он более чем на 50% не выполнил поставленные в практической работе задачи, не способен ответить на вопросы, касающиеся теоретической составляющей в объеме содержания учебной дисциплины, относящейся к практической работе.

Дополнительные вопросы

1. Понятие Data Mining.
2. Данные, информация и знания.
3. Единая методология обнаружения знаний.
4. Задача регрессии.
5. Задача классификации.
6. Задача кластеризации.
7. Задача анализа ассоциаций и последовательностей.
8. Наборы и типы данных.
9. Форматы хранения данных.
10. Качество данных. Очистка данных.
11. Методы отбора значимых признаков.
12. Фильтры. Оболочки.
13. Индукция деревьев решений.
14. «Обрезка» деревьев: предредукция и постредукция.
15. Решающие правила.
16. Алгоритмы ID3, CART, C4.5.
17. Алгоритмы ограниченного перебора.
18. Метод опорных векторов. Линейная и нелинейная разделимость.
19. Байесовская классификация.
20. Наивная байесовская классификация.
21. Типологический и таксономический анализ.
22. Статистические методы кластеризации. EM-алгоритм.
23. Метод k-средних. Меры расстояний.
24. Иерархические методы кластеризации.
25. Визуализация кластеров. Дендрограммы.
26. Диаграммы рассеивания.
27. Самоорганизующиеся карты Кохонена.
28. Концептуальная кластеризация.
29. Алгоритм Cobweb.
30. Графовые методы кластеризации.
31. Выделение связных компонент.
32. Нечеткая кластеризация.
33. FCM-алгоритм.
34. Меры интересности: поддержка, достоверность, лифт, уверенность.
35. Алгоритм Apriori. Задача анализа рыночных корзин.
36. Матрица несоответствий.
37. Метрики качества: правильность, полнота, точность, F-мера, чувствительность, AUC.
38. Проблема переобучения.
39. Стратификация данных.
40. Диаграмма выигрыша.
41. Ансамбли (комитеты) моделей. Бэггинг. Бэггинг с рандомизацией.
42. Бустинг (усиление) ансамбля классификаторов.
43. Метаклассификаторы. Стэкинг.
44. Понятия социальной сети и социального графа.
45. Позиционный и ролевой анализ в социальной сети.

Методические рекомендации по выполнению практических работ

Целью практических работ является приобретение практических навыков использования математических моделей, методов и алгоритмов в области технологий анализа больших данных; усвоение полученных знаний студентами, а также формирование у них мотивации к самообразованию за счет активизации самостоятельной познавательной деятельности.

Все работы выполняются студентами в рамках 4-х академических часов, которые отведены учебным планом.

Итогом работы является защита полученных результатов. Защита проводится индивидуально в форме собеседования и проверке полученных навыков работы с системой на компьютере.

10.7. Тест № 1

1. Явление «информационный потоп» подчиняется закону
   - ○ Паркинсона
   - ○ Мура
   - ○ Мёрфи
   - ○ Ньютона

2. Основу интеллектуального анализа данных составляют
   - ○ методы машинного обучения
   - ○ статистические методы
   - ○ единая методология обнаружения знаний
   - ○ все вышеперечисленное

3. Задача, решение которой впервые было предложено в рамках интеллектуального анализа данных:
   - ○ классификация
   - ○ регрессия
   - ○ кластеризация
   - ○ анализ ассоциаций

3. Дерево решений на рисунке 1 содержит ___ корней, ___ узлов проверки,

____ ветвей, ___ терминальных вершин, ___ листьев.

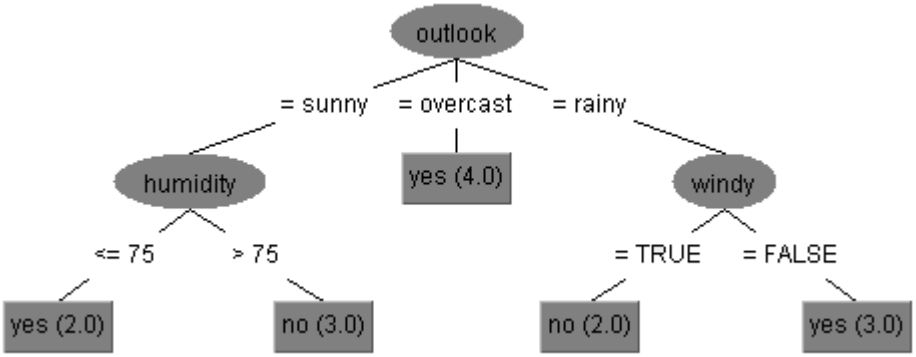


Рис.1. Дерево решений

4. По рисунку 1 определите атрибут с максимальным информационным выигрышем
   - ○ outlook
   - ○ sunny
   - ○ overcast
   - ○ rainy

5. Ограничение глубины дерева может быть следующим критерием:
   - ○ останова
   - ○ расщепления
   - ○ замены
   - ○ любым из вышеперечисленных

6. Какое множество используется на этапе постредукции деревьев решений?
   - ○ весь набор данных

- обучающее
- тестовое
- подтверждающее

7. Какой критерий расщепления используется в алгоритме ID3?
- ограничение глубины дерева
- энтропия
- информационный выигрыш
- индекс Gini

8. Какой алгоритм позволяет работать с непрерывной целевой переменной:
- ID3
- CART
- C4.5
- J4.8

9. Самую надежную оценку классификатора можно получить, используя следующий способ проверки:
- 10-перекрестная проверка
- 10-перекрестная проверка со стратификацией
- десятикратная 10-перекрестная проверка
- перекрестная проверка методом «без одного»

10. Установите соответствие между методом кластеризации и классом используемых алгоритмов.

1. K-средних	а. Инкрементальный
2. Cobweb	б. Графовый
3. EM	в. Статистический
4. ARHP	г. Итеративный

11. Кластер Y*NY можно охарактеризовать как

- монотетический
- расплывчатый монотетический
- политетический
- расплывчатый политетический

12. Оптимальное разбиение на кластеры методом Cobweb подразумевает
- нулевое значение функции полезности
- минимизацию значения функции полезности
- усреднение значения функции полезности
- максимизацию значения функции полезности

13. При перераспределении объектов между кластерами в алгоритме *k*-средних используется следующая метрика:
- евклидово расстояние
- квадрат евклидова расстояния
- расстояние «городских кварталов»
- любая из вышеперечисленных

14. Метод опорных векторов можно охарактеризовать как
- метод классификации вида «обучение с учителем»
- метод классификации вида «обучение без учителя»
- метод кластеризации
- метод регрессии

15. В случае линейной разделимости оптимальный классификатор на основе метода опорных векторов подразумевает

- Максимизацию расстояния между векторами и минимизацию ошибки обучения
- Максимизацию расстояния между векторами и нулевую ошибку обучения
- Минимизацию расстояния между векторами и минимизацию ошибки обучения
- Минимизацию расстояния между векторами и нулевую ошибку обучения

16. Бинарному классификатору с какой точностью следует отдать предпочтение?

- 40%
- 50%
- 60%
- 70%

17. Какая оценка эффективности классификатора наиболее достоверна?

- точность
- чувствительность
- доминирование в ROC-пространстве
- площадь под ROC-кривой

18. На основе матрицы несоответствий рассчитайте следующие значения:

Ложная тревога (ошибка первого рода) _____
Точность (правильность) _____
Чувствительность (TP Rate) _____
F-мера Ван Ризбергена _____

23	9
1	7

19. Из каких соображений назначается стоимость (A) ложной тревоги/ ошибки первого рода и (B) промаха/ошибки второго рода для директ-рассылки маркетинговых материалов?

0	B
A	0

A _____

B _____

20. Оптимальное число репликаций в бутстреп-выборке:

- 90%
- 2/3
- 63,2%
- 36,8%

21. Выпуклая ROC-поверхность определяется следующими классификаторами:
- случайными
- «идеальными»
- доминирующими
- субоптимальными

22. По рисунку 2 отранжируйте классификаторы по доминированию в ROC-пространстве:

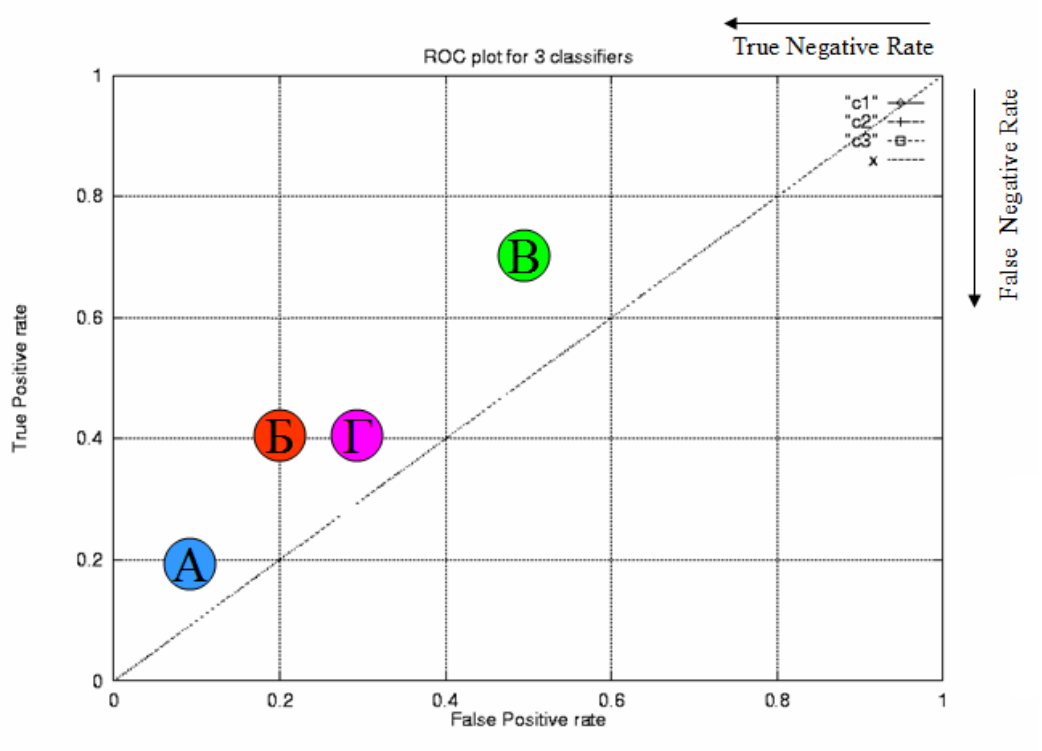


Рис.2. ROC-пространство

1. ___ 2. ___ 3. ___ 4. ___

23. По рисунку 2 оцените значение метрики выпадения для комбинированного классификатора с полнотой 0,6.

Выпадение _____

24. Имеется несколько моделей с коррелируемыми ошибками. Какой комитетный метод не улучшит результатов классификации?
- Бэггинг
- Бэггинг с рандомизацией
- Бустинг
- Ни один из вышеперечисленных

25. Выберите метод ансамблирования классификаторов NaiveBayes и J4.8.
- Бустинг
- Стэкинг
- Любой из вышеперечисленных
- Ни один из вышеперечисленных

26. Пусть {1,2,3,4} является часто встречающимся набором элементов с поддержкой 40%. Что можно сказать о наборе {1,3,4}?

- Часто встречающийся с поддержкой 40%
- Часто встречающийся с поддержкой ≥40%
- Не часто встречающийся с поддержкой ≤40%
- Часто встречающийся с поддержкой ≤40%

27. На основе таблицы транзакций рассчитайте значения следующих мер для ассоциативного правила A → C.

Транзакция	Элементы
t_1	A,B,C
t_2	A,C
t_3	A,C,D
t_4	A,E
t_5	D,E

Поддержка _____
Достоверность _____
Лифт _____
Уверенность _____

28. Для каких задач ассоциативные правила используются в мерчендайзинге?

- Перекрестные продажи (cross-selling)
- Повышение продаж (up-selling)
- Управление лояльностью покупателей
- Для всех вышеперечисленных

Критерии оценки:

Тестирование оценивается дифференцированно по балльной шкале:

- выполнено без ошибок и недочетов 85-100% от общего объема заданий – выставляется от 18 до 20 баллов;
- выполнено без ошибок и недочетов 71-84% от общего объема заданий – выставляется от 15 до 17 баллов;
- выполнено без ошибок и недочетов 60-70% от общего объема заданий – выставляется от 13 до 14 баллов;
- выполнено без ошибок и недочетов 50-59% от общего объема заданий – выставляется от 10 до 12 баллов.

10.1. Тест № 2

1. **Текст задания:**

Рассмотрим следующий обучающий набор для модели деревьев решений:

Номер примера	Классификация	Атрибут 1	Атрибут 2
1	+	T	T
2	+	T	T
3	-	T	F
4	+	F	F

5	-	F	T
6	-	F	T

Рассчитайте информационный выигрыш для атрибута 2.

Ответ (ключ):

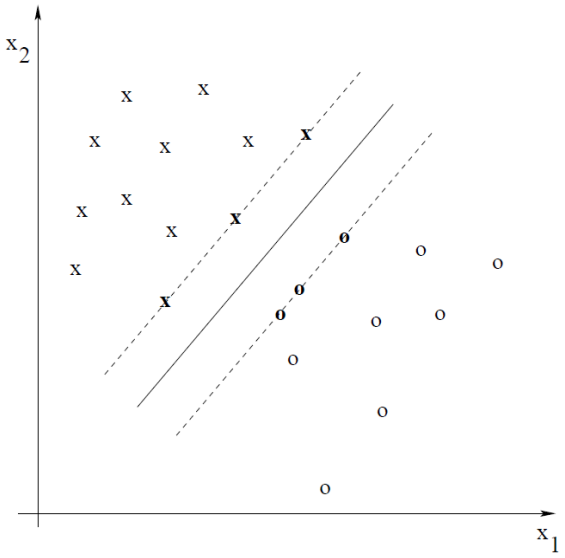
0

Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

2. **Текст задания:**

Оцените ошибку перекрестной проверки для разделяющей прямой на основе метода опорных векторов.



Ответ (ключ):

0

Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

3. **Текст задания:**

Для каких из следующих наборов данных целесообразно использовать скрытую Марковскую модель?

A. Набор данных о последовательности генов.
B. База данных IMDB с рейтингами кинофильмов.
C. Исторические данные фондового рынка.
D. Данные о суточных осадках.

Ответ (ключ):

A, C, D

Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

4. **Текст задания:**

Если параметр C стремится к бесконечности, какое из следующих утверждений верно для метода опорных векторов?

A. Оптимальной гиперплоскостью, если она существует, будет та, которая полностью разделяет данные.
B. Данные сможет разделить классификатор с мягким зазором.
C. Ни одно из перечисленных.

Ответ (ключ):

A

Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

5. **Текст задания:**

Какая из следующих метрик не может применяться для оценки эффективности модели логистической регрессии?

A. AUC-ROC
B. Правильность (Accuracy)
C. Logloss
D. Mean-Squared-Error (MSE)

Ответ (ключ):

D

Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

6. **Текст задания:**

Исследования показывают, что прослушивание музыки во время учебы может улучшить вашу память. Чтобы продемонстрировать это, исследователь дает 36 студентам колледжа стандартный тест на проверку кратковременной памяти, пока они слушают фоновую музыку. Средняя оценка по результатам эксперимента составляет 28. При нормальных обстоятельствах (без музыки) средняя оценка равна 25, а стандартное отклонение 6.

Какой в этом случае будет нулевая гипотеза?

A. Прослушивание музыки во время учебы не повлияет на память.
B. Прослушивание музыки во время учебы может ухудшить память.
C. Прослушивание музыки во время учебы может улучшить память.
D. Прослушивание музыки во время учебы не улучшит память, но может ухудшить ее.

Ответ (ключ):

D

Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

7. **Текст задания:**

Вы создали модель машинного обучения в Python, которую хотите "заморозить" сейчас, чтобы вернуться к ней позже. Библиотека Pickle была импортирована как pkl. Какая из следующих команд может выполнить эту задачу?

A. push(model, “file”)
B. save(model, “file”)
C. dump(model, “file”)
D. freeze(model, “file”)

Ответ (ключ):

C

Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

8. **Текст задания:**

Что из перечисленного верно для «белого шума»?

A. Нулевое среднее.
B. Нулевая автокорреляция.
C. Нулевая автокорреляция, кроме лага 0.
D. Квадратичное отклонение.

Ответ (ключ):

C

Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

9. **Текст задания:**

Какое минимальное количество переменных требуется для выполнения кластеризации?

A. 0
B. 1
C. 2
D. 3

Ответ (ключ):

B

Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

Текст задания:

10. Каким будет результат выполнения следующей функции в R?

```
>    f <- function(num = 1) {
            hello <- "Hello, world!\n"
      for(i in seq_len(num)) {
            cat(hello)
      }
      chars <- nchar(hello) * num
      chars
}
> f()
```

A. Hello, world!
14

B. Hello, world!\n
14

C. Hello, world!
13

D. Hello, world!\n
13

Ответ (ключ):

A

Критерии и параметры оценивания:

Правильный ответ – 5 баллов, неправильный ответ – 0 баллов

11. **Текст задания:**

Рассмотрим многоклассовую задачу классификации для прогнозирования качества вина на основе его характеристик. Данные загружаются во фрейм данных df:

	fixed acidity	**volatile acidity**	**citric acid**	**residual sugar**	**chlorides**	**free sulfur dioxide**	**total sulfur dioxide**	**density**
0	7.4	0.70	0.00	1.9	0.076	11	34	0.9978
1	7.8	0.88	0.00	2.6	0.098	25	67	0.9968
2	7.8	0.76	0.04	2.3	0.092	15	54	0.9970
3	11.2	0.28	0.56	1.9	0.075	17	60	0.9980
4	7.4	0.70	0.00	1.9	0.076	11	34	0.9978

Столбец качества в настоящее время имеет значения от 1 до 10, но мы хотим перейти к бинарной классификации. Вы хотите использовать пороговое значение 5. Если качество выше 5, булева переменная принимает значение 1, иначе 0.

Numpy был импортирован как np. Какой из следующих фрагментов кода в Python поможет вам выполнить эту задачу?

A. Y = df[quality].values

Y = np.array([1 if y >= 6 else 0 for y in Y])

B. Y = df[quality].values()

Y = np.array([0 if y >= 6 else 1 for y in Y])

C. Y = df[quality]

Y = np.array([0 if y >= 6 else 1 for y in Y])

D. Ни один из перечисленных

Ответ (ключ):

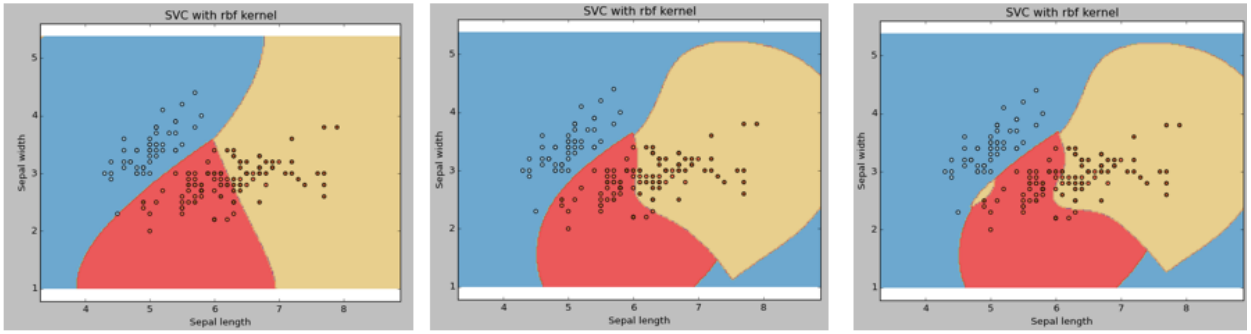
A

Критерии и параметры оценивания:

Правильный ответ – 10 баллов, неправильный ответ – 0 баллов

12. **Текст задания:**

Рассмотрим графики для различных значений *C* в методе опорных векторов. Какой из следующих вариантов лучше всего объясняет значения *C* для трех изображений ниже в случае радиальной ядерной функции?



A. *C*1 = *C*2 = *C*3
B. *C*1 > *C*2 > *C*3
C. *C*1 < *C*2 < *C*3
D. Ни один из перечисленных.

Ответ (ключ):

C

Критерии и параметры оценивания:

Правильный ответ – 10 баллов, неправильный ответ – 0 баллов

13. **Текст задания:**

Какой из следующих методов перекрестной проверки лучше подходит для временных рядов?

A. K-перекрестная проверка.
B. Перекрестная проверка методом «без одного».
C. Перекрестная проверка со стратифицированной перетасовкой.
D. Перекрестная проверка с прямой цепочкой.

Ответ (ключ):

D

Критерии и параметры оценивания:

Правильный ответ – 10 баллов, неправильный ответ – 0 баллов

14. **Текст задания:**

На выборах N кандидатов конкурируют друг с другом, и избиратели голосуют за любого из кандидатов. Избиратели не общаются друг с другом во время голосования. Какая из следующих комитетных моделей похожа на описанную процедуру выборов?

A. Bagging
B. Boosting
C. Stacking
D. AdaBoost.M1

Ответ (ключ):

A

Критерии и параметры оценивания:

Правильный ответ – 10 баллов, неправильный ответ – 0 баллов

15. **Текст задания:**

Предположим, вы настроили сложную регрессионную модель на наборе данных. Теперь вы используете гребневую регрессию с параметром настройки лямбда, чтобы уменьшить сложность модели. Выберите вариант, который описывает связь смещения и дисперсии с лямбдой.

A. В случае очень большой лямбды смещение низкое, а дисперсия низкая.
B. В случае очень большой лямбды смещение низкое, а дисперсия высокая.
C. В случае очень большой лямбды смещение высокое, а дисперсия низкая.
D. В случае очень большой лямбды смещение высокое, а дисперсия высокая.

Ответ (ключ):

C

Критерии и параметры оценивания:

Правильный ответ – 10 баллов, неправильный ответ – 0 баллов

Критерии оценки:

Тестирование оценивается дифференцированно по балльной шкале:

- выполнено без ошибок и недочетов 85-100% от общего объема заданий – выставляется от 18 до 20 баллов;
- выполнено без ошибок и недочетов 71-84% от общего объема заданий – выставляется от 15 до 17 баллов;
- выполнено без ошибок и недочетов 60-70% от общего объема заданий – выставляется от 13 до 14 баллов;
- выполнено без ошибок и недочетов 50-59% от общего объема заданий – выставляется от 10 до 12 баллов.