Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего образования
«Сибирский государственный индустриальный университет»

Кафедра прикладных информационных технологий и программирования

УТВЕРЖДАЮ
Директор института
информационных технологий и
автоматизированных систем
_____ А.В. Зимин
подпись
«____» _____ 20___ г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Обработка и анализ данных

09.03.03 «Прикладная информатика»
(направленность (профиль): «Искусственный интеллект и машинное обучение»)

Квалификация выпускника
Бакалавр

Форма обучения
Очная форма

Срок обучения: 4 года

Год начала подготовки 2025

Новокузнецк
2025

СибГИУ
Подлинник электронного документа хранится в системе электронного согласования документов СибГИУ
ДОКУМЕНТ ПОДПИСАН ЭП
Сертификат: daba56d7b892efca3bf5c04ed0251c98b19b099c
Владелец: Директор ИИТиАС Зимин Алексей Валерьевич
Действителен: с 15 января 2025 г. по 10 апреля 2026 г.
Дата и время: 10-04-2025 15:26:08 ПРОВЕРИТЬ ПОДПИСЬ

1 Цели и задачи освоения учебной дисциплины

Целями учебной дисциплины являются:

– формирование целостного представления, понимания места и роли анализа данных, а также освоение методов и алгоритмов анализа и обработки данных при решении практических задач с использованием современных информационных технологий.

Задачами учебной дисциплины являются:

– научить обучающихся извлекать полезную информацию из сложных рядов данных;;
– проводить проверку воспроизводимости и статистической значимости получаемых оценок;;
– визуализировать данные;;
– использовать методы и алгоритмы интеллектуального анализа данных для решения практических задач;;
– получить навыки применения программных продуктов, предназначенных для анализа данных.

2 Место учебной дисциплины в структуре ООП по направлению подготовки (специальности)

Учебная дисциплина относится к учебным дисциплинам обязательной части **Блока 1 «Дисциплины (модули)»** ООП по направлению подготовки (специальности) 09.03.03 «Прикладная информатика».

Учебная дисциплина базируется на предварительном усвоении обучающимися учебных дисциплин:

– Математика;
– Теория вероятностей и математическая статистика;
– Основы программирования;
– Программирование;
– Численные методы решения инженерных задач.

Учебная дисциплина дополняет знания и умения, получаемые по одновременно изучаемым и последующим учебным дисциплинам:

– Теория искусственных нейронных сетей и машинное обучение;
– Основы работы в Business Intelligence системах;
– Моделирование;
– Проектная деятельность 2;
– Проектная деятельность 3;
– Научно-исследовательская работа;
– Подготовка к сдаче и сдача государственного экзамена.

3 Планируемые результаты обучения по учебной дисциплине

Процесс изучения учебной дисциплины направлен на формирование следующих компетенций:

– Общепрофессиональные компетенции

Наименование категории (группы) ОПК	Код и наименование ОПК	Код и наименование индикатора достижения ОПК	Планируемые результаты обучения
	ОПК-1: Способен применять естественнонаучные и общеинженерные знания, методы математического анализа и моделирования, теоретического и экспериментального исследования в профессиональной деятельности	ОПК-1.2 Использует естественнонаучные и общеинженерные методы и инструменты при решении практических задач	– знать: область применения для анализа и обработки данных в научно-исследовательских и прикладных задачах. – уметь: проводить предварительную подготовку и анализ данных с использованием современных ИКТ.
		ОПК-1.3 Применяет методы математического анализа и моделирования в теоретических и экспериментальных исследованиях профессиональной деятельности	– знать: методы и алгоритмы статистического и интеллектуального анализа данных. – уметь: применять методы и алгоритмы статистического и интеллектуального анализа данных при решении практических задач.
	ОПК-3: Способен решать стандартные задачи профессиональной деятельности на основе информационной и библиографической культуры с применением информационно-коммуникационных технологий и с учетом основных требований информационной безопасности	ОПК-3.1 Находит и анализирует информацию, необходимую для решения задач профессиональной и научно-исследовательской деятельности, с применением современных инфокоммуникационных технологий	– знать: методы и алгоритмы сбора; преобразования, визуализации данных с использованием ИКТ и специализированного программного обеспечения. – уметь: применять ИКТ и специализированного программного обеспечения при сборе, преобразовании, визуализации данных при решении прикладных и научно-исследовательских задач.

4 Объем и содержание учебной дисциплины

Учебные занятия по учебной дисциплине проводятся в форме контактной работы и в форме самостоятельной работы обучающихся.

Контактная работа включает в себя занятия лекционного типа (лекции), занятия семинарского типа (семинары, практические занятия, практикумы, лабораторные работы, коллоквиумы), промежуточную аттестацию обучающихся и иные формы взаимодействия обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации ООП на иных условиях, в том числе при проведении промежуточной аттестации обучающихся. Контактная работа может проводиться с применением электронного обучения, дистанционных образовательных технологий.

Объем учебной дисциплины

Семестр / курс		***ИТОГО***	***5 семестр***
Форма промежуточной аттестации			*экзамен*
Трудоёмкость	*академ. час.*	***216***	*216*
	зачетных единиц	***6***	*6*
Лекции, *академ. час.*		***32***	*32*
в форме практической подготовки		***0***	*0*
Лабораторные работы, *академ. час.*		***16***	*16*
в форме практической подготовки		***0***	*0*
Практические занятия, *академ. час.*		***32***	*32*
в форме практической подготовки		***0***	*0*
Курсовая работа / проект, *академ. час.*		***0***	*0*
в форме практической подготовки		***0***	*0*
Консультации, *академ. час.*		***0***	*0*
в форме практической подготовки		***0***	*0*
Самостоятельная работа, *академ. час.*		***100***	*100*
в форме практической подготовки		***0***	*0*
Контроль, *академ. час.*		***36***	*36*
в форме практической подготовки		***0***	*0*

Содержание учебной дисциплины

Раздел 1 Прикладной статистический анализ;

Тема 1.1 Закон распределения случайной величины и способы его оценки. (Закона распределения случайной величины в интегральной и дифференциальной форме. Статистический ряд. Построение гистограмм. Виды законов распределения и гистограмм. Проверка правдоподобия гипотез с использованием статистических критериев.);

Тема 1.2 Статистические характеристики. (Выборочное среднее, дисперсия, среднеквадратическое отклонение, моменты различных порядков и их свойства. Статистическая гипотеза и способы ее проверки.);

Тема 1.3 Многовариантное усреднение данных (Алгоритмы усреднения данных, включая скользящее арифметическое среднее, экспоненциальное средне, релейно-экспоненциальное среднее, медианное среднее и др. Методика выбора настроечных коэффициентов . Использование алгоритмов скользящих средних для построения прогностических моделей);

Тема 1.4 Корреляционный анализ (Парные коэффициенты корреляции. Авто-, взаимно корреляционные функции. Корреляционные матрицы. Частные и множественные коэффициенты корреляции. Непараметрические коэффициенты корреляции. Условные средние. Авто-, взаимно корреляционное отношение. Показатель степени нелинейности. Непараметрический корреляционный анализ.);

Раздел 2 Интеллектуальный анализ данных;

Тема 2.1 Введение в интеллектуальный анализ данных (Задачи интеллектуального анализа данных и их классификация. Модели и методы Data Mining.);

Тема 2.2 Задача ассоциации. (Ассоциативные правила. Алгоритм a priori.);

Тема 2.3 Задача кластеризации. (Введение в кластеризацию. Алгоритмы k-means и g-means.);

Тема 2.4 Задача классификации и регрессии. (Статистические методы классификации и регрессии. Метод k-ближних соседей, правила классификации.).

5 Перечень тем лекций

№ раздела / темы дисциплины	Темы лекций	Трудоемкость, *академ.час*	
		всего	**в форме практической подготовки**
Раздел 1.	Прикладной статистический анализ		
Тема 1.1.	Закон распределения случайной величины и способы его оценки.	2	
Тема 1.2.	Статистические характеристики.	6	
Тема 1.3.	Многовариантное усреднение данных	6	
Тема 1.4.	Корреляционный анализ	6	
Раздел 2.	Интеллектуальный анализ данных		
Тема 2.1.	Введение в интеллектуальный анализ данных.	2	
Тема 2.2.	Задача ассоциации.	2	
Тема 2.3.	Задача кластеризации.	4	
Тема 2.4.	Задача классификации и	4	

	регрессии.		
Итого:		**32**	**0**

6 Перечень тем практических занятий (семинаров)

№ раздела / темы дисциплины	Темы практических занятий (семинаров)	Трудоемкость, *академ.час*	
		всего	**в форме практической подготовки**
Тема 1.1.	Закон распределения случайной величины и способы его оценки	2	
Тема 1.2.	Статистические характеристики	6	
Тема 1.3.	Многовариантное усреднение данных	8	
Тема 1.4.	Корреляционный анализ	8	
Тема 2.3.	Задача кластеризация	4	
Тема 2.4.	Задача классификация и регрессия	4	
Итого:		**32**	**0**

7 Перечень тем лабораторных работ

№ раздела / темы дисциплины	Темы лабораторных работ	Трудоемкость, *академ.час*	
		всего	**в форме практической подготовки**
Раздел 1.	Знакомство с аналитической платформой	2	
Раздел 1.	Основы работы с аналитической платформой	2	
Раздел 1.	Предварительная обработка и исследование данных	4	
Тема 2.1.	Задача поиска ассоциативных плавил	2	
Тема 2.3.	Задача кластеризации	2	
Тема 2.4.	Задача классификации	2	
Тема 2.4.	Задача регрессии	2	
Итого:		**16**	**0**

8 Перечень тем курсовых работ (проектов)

№ раздела / темы дисциплины	Темы курсовых работ (проектов)	Трудоемкость, *академ.час*	
		всего	**в форме практической подготовки**
	Отсутствуют		
Итого:		**0**	**0**

9 Виды самостоятельной работы

№ раздела / темы дисциплины	Виды самостоятельной работы	Трудоемкость, *академ.час*	
		всего	**в форме практической подготовки**
Раздел 1.	1. Изучение лекционного материала; 2. Подготовка к лабораторной работе; 3. Подготовка к практическому занятию; 4. Прохождение тестирования.	50	
Раздел 2.	1. Изучение лекционного материала; 2. Подготовка к лабораторной работе; 3. Подготовка к практическому занятию; 4. Прохождение тестирования.	50	
Контроль	*Подготовка к экзамену*	36	
Итого:		**136**	**0**

10 Матрица компетентностных задач по учебной дисциплине

Контролируемые разделы (темы) дисциплины	Контролируемые компетенции	Оценочные средства
Раздел 1.	ОПК-1; ОПК-3	1. Вопросы для устного обсуждения (опроса); 2. Наблюдение за работой на лабораторной работе; 3. Наблюдение за работой на практическом занятии; 4. Оценка участия в работе на лекционном занятии; 5. Тестовые задания
Раздел 2.	ОПК-1; ОПК-3	1. Вопросы для устного обсуждения (опроса); 2. Наблюдение за работой на лабораторной работе; 3. Наблюдение за работой на практическом занятии; 4. Оценка участия в работе на лекционном занятии; 5. Тестовые задания
Экзамен.	ОПК-1; ОПК-3	1. Вопросы к экзамену; 2. Задачи к экзамену

11 Входной контроль

11.1 Цель входного контроля − определить начальный уровень подготовленности обучающихся и выстроить индивидуальную траекторию обучения. В условиях личностно-ориентированной образовательной среды результаты входного оценивания обучающегося используются как начальные значения в индивидуальном профиле академической успешности обучающегося.

11.2 Описание оценочных средств

Форма проведения входного контроля – тестирование.

Длительность тестирования – 30 минут. Тестирование проводится в течение первых двух недель изучения учебной дисциплины на практическом занятии.

Шкала оценивания результатов тестирования

За каждый правильный ответ выставляется 1 балл. Максимальное количество – 20 баллов.

Оценка формируется в соответствии с критериями таблицы:

Максимальный балл	Проходной балл	Оценка
20	не менее 18	отлично
17	не менее 14	хорошо
13	не менее 10	удовлетворительно
9	-	неудовлетворительно

Основываясь на результатах тестирования, формируется индивидуальная траектория обучающегося (группы обучающихся), исходя из следующих соображений:

– 14 – 20 баллов – обучающийся готов к изучению учебной дисциплины без дополнительного повторения изученного ранее материала;

– 10 – 13 баллов – обучающийся в целом готов к изучению учебной дисциплины, требуется повторение отдельных вопросов(тем) изученного ранее материала, необходимого для успешного освоения учебной дисциплины;

– менее 10 баллов – обучающийся имеет слабую подготовку к изучению учебной дисциплины, необходимы корректирующие занятия и консультации.

Тестовые задания для входного контроля знаний обучающихся приведены в приложении.

12 Текущий контроль

12.1 Текущий контроль знаний используется для оперативного и регулярного управления учебной деятельностью (в том числе самостоятельной) обучающихся. Данный вид контроля стимулирует у обучающихся стремление к систематической самостоятельной работе по изучению учебной дисциплины и позволяет оценить степень освоения изученного материала.

Текущий контроль успеваемости осуществляется в течение семестра, в ходе повседневной учебной работы по индивидуальной инициативе педагогического работника, в том числе посредством использования СУО «Moodle».

Формы проведения текущего контроля – вопросы для устного обсуждения (опроса); наблюдение за работой на лабораторной работе; наблюдение за работой на практическом занятии; оценка участия в работе на лекционном занятии; тестовые задания.

12.2 Описание оценочных средств

Шкалы оценивания (методики оценки)

12.2.1 Рекомендации по оцениванию устных ответов обучающихся

С целью контроля и подготовки обучающихся к изучению новой темы в начале занятия лекционного типа (лекции) педагогическим работником проводится индивидуальный или фронтальный устный опрос по выполненным заданиям предыдущего раздела (темы).

Критерии оценки:

– правильность ответа по содержанию задания (учитывается количество и характер ошибок при ответе);

– полнота и глубина ответа (учитывается количество усвоенных фактов, понятий и т.п.);

– сознательность ответа (учитывается понимание излагаемого материала);

– логика изложения материала (учитывается умение строить целостный, последовательный рассказ, грамотно пользоваться специальной терминологией);

– рациональность использованных приемов и способов решения поставленной учебной задачи (учитывается умение использовать наиболее прогрессивные и эффективные способы достижения цели);

– своевременность и эффективность использования наглядных пособий и технических средств при ответе (учитывается умение грамотно и с пользой применять наглядность и демонстрационный опыт при устном ответе);

– использование дополнительного материала (обязательное условие);

– рациональность использования времени, отведенного на задание (не одобряется затянутость выполнения задания, устного ответа во времени).

Оценка «отлично» ставится, если обучающийся:

– полно и аргументировано отвечает по содержанию задания;

– обнаруживает понимание материала, может обосновать свои суждения, применить знания на практике, привести необходимые примеры, в том числе самостоятельно составленные;

– излагает материал последовательно и правильно.

Оценка «хорошо» ставится, если обучающийся дает ответ, удовлетворяющий тем же требованиям, что и для оценки «отлично», но допускает 1 – 2 ошибки, которые сам же исправляет.

Оценка «удовлетворительно» ставится, если обучающийся обнаруживает знание и понимание основных положений данного задания, но:

– излагает материал неполно и допускает неточности в определении понятий или формулировке правил;

– не умеет достаточно глубоко и доказательно обосновать свои суждения и привести свои примеры;

– излагает материал непоследовательно и допускает ошибки.

Оценка «неудовлетворительно» ставится, если обучающийся обнаруживает незнание ответа на соответствующее задание, допускает ошибки в формулировке определений и правил, искажающие их смысл, беспорядочно и неуверенно излагает материал. Оценка «неудовлетворительно» отмечает такие недостатки в подготовке обучающегося, которые являются серьезным препятствием к успешному освоению последующего материала.

12.2.2 Рекомендации по оцениванию результатов тестирования обучающихся

Тестирование проводится по изученным разделам (темам) учебной дисциплины для выявления уровня знаний и умений обучающегося, мотивирования обучающегося к активизации самостоятельной работы по усвоению учебного материала, выявления и устранения пробелов в знаниях, повышения дисциплины и организации учебной деятельности обучающегося.

Тестирование проводится в СУО «Moodle» в рамках часов, отведенных на самостоятельную работу обучающихся. Для тестирования используется 20 вопросов из базы типовых тестовых заданий по изученным разделам (темам) учебной дисциплины. Обучающийся выбирает один ответ из предложенных вариантов ответа. В случае выбора неправильного ответа обучающийся получает за данный вопрос 0 баллов; в случае правильного ответа – 1 балл.

Оценка формируется в соответствии с критериями таблицы:

Максимальный балл	Проходной балл	Оценка
20	не менее 18	отлично
17	не менее 14	хорошо
13	не менее 10	удовлетворительно
9	–	неудовлетворительно

12.2.3 Рекомендации по оцениванию участия обучающихся в работе на лекционных занятиях

Формирование у обучающихся добросовестного и ответственного отношения к собственному профессиональному образованию проявляется в обязательном его присутствии на лекционном занятии.

На лекционных занятиях происходит формирование информационной компетентности обучающихся, в том числе способности фиксировать информацию с помощью конспектирования учебного материала. Преподаватель на лекционных занятиях акцентирует внимание обучающихся на том, как необходимо оформлять конспект лекций в тетради. По окончании лекционных занятий преподаватель осуществляет проверку тетрадей обучающихся и оценивает их.

Критерии оценки участия обучающихся в работе на лекционных занятиях

Наименование параметра оценки	Максимальное количество баллов за одно лекционное занятие
Наличие у обучающихся добросовестного и ответственного отношения к собственному профессиональному образованию, выраженное в посещении лекций	0,5
Умение конспектировать учебный материал в тетради, в том числе: - материал изложен последовательно, четко и аккуратно - в конспекте имеются изученные темы, основные термины, схемы	0,5
Всего	1

Если по результатам оценки участия обучающийся в лекционных занятиях в среднем за семестр обучающийся набрал не менее 0,5 баллов, то ему выставляется оценка «зачтено», менее 0,5 баллов – «не зачтено».

12.2.4 Рекомендации по оцениванию в процессе наблюдения за работой на практическом занятии

Наблюдение за работой на практическом занятии – важнейший компонент контроля деятельности обучающихся и качества их обучения.

Критерии оценки в процессе наблюдения за работой на практическом занятии:

– самостоятельность выполнения задания и работы с методическими указаниями (учитывается работа в течение занятия, быстрота и способность нахождения в методических указаниях нужной информации);

– правильность выполнения работы (учитывается последовательность выполняемых действий, умение работать с инструментами и/или средствами вычислительной техники);

– рациональность использования времени, отведенного на задание (не одобряется затянутость выполнения работы с учетом индивидуальных особенностей обучающихся);

– активность обучающегося при выполнении практической работы (учитывается индивидуальная работа и работа в группе).

Критерии оценки по результатам наблюдения за работой на практическом занятии

Наименование параметра оценки	Максимальное количество баллов
Самостоятельно выполняет задания; умеет работать с методическими указаниями.	2,5
Указанные в методических указаниях действия выполняет в правильной последовательности; умеет работать с инструментами и/или средствами вычислительной техники.	1,25
Рационально использует время, отведенное на задание; проявляет активность при выполнении практической работы.	1,25
Итого	**5**

Максимальное количество – 5 баллов соответствует оценке «отлично», 4 балла – «хорошо», 3 балла – «удовлетворительно», 2 балла – «неудовлетворительно».

12.2.5 Рекомендации по оцениванию в процессе наблюдения за работой на лабораторном занятии

Наблюдение за работой на лабораторном занятии – важнейший компонент контроля деятельности обучающихся и качества их обучения.

Критерии оценки в процессе наблюдения за работой на лабораторном занятии:

– самостоятельность выполнения задания и работы с методическими указаниями (учитывается работа в течение занятия, быстрота и способность нахождения в методических указаниях нужной информации);

– правильность выполнения работы (учитывается последовательность выполняемых действий, умение работать с инструментами и/или средствами вычислительной техники);

– рациональность использования времени, отведенного на задание (не одобряется затянутость выполнения работы с учетом индивидуальных особенностей обучающихся);

– активность обучающегося при выполнении лабораторной работы (учитывается индивидуальная работа и работа в группе).

Критерии оценки по результатам наблюдения за работой на лабораторном занятии

Наименование параметра оценки	Максимальное количество баллов
Самостоятельно выполняет задания; умеет работать с методическими указаниями.	2,5
Указанные в методических указаниях действия выполняет в правильной последовательности; умеет работать с инструментами и/или средствами вычислительной техники.	1,25
Рационально использует время, отведенное на задание; проявляет активность при выполнении лабораторной работы.	1,25
Итого	**5**

Максимальное количество – 5 баллов соответствует оценке «отлично», 4 балла – «хорошо», 3 балла – «удовлетворительно», 2 балла – «неудовлетворительно».

Задания и вопросы для текущего контроля знаний обучающихся по разделам (темам) учебной дисциплины приведены в приложении.

13 Промежуточная аттестация

13.1 Оценочные материалы для промежуточной аттестации обучающихся по учебной дисциплине предназначены для оценки степени достижения запланированных результатов обучения по завершению изучения учебной дисциплины в установленной учебным планом форме и позволяют определить качество усвоения изученного материала.

Итоговой формой контроля сформированности компетенций у обучающихся по учебной дисциплине является: экзамен.

Оценочные материалы для промежуточной аттестации обучающихся включают: вопросы к экзамену; задачи к экзамену.

13.2 Оценивание обучающегося на экзамене

Оценка	Требования к знаниям
отлично	Оценка «отлично» выставляется обучающемуся, если он глубоко и прочно усвоил программный материал, исчерпывающе, последовательно, четко и логически его излагает, умеет тесно увязывать теорию с практикой, свободно справляется с задачами, вопросами и другими видами применения знаний, причем не затрудняется с ответом при

Оценка	Требования к знаниям
	видоизменении заданий, использует в ответе материал различной литературы, правильно обосновывает принятое нестандартное решение, обладает приемами выполнения практических задач по формированию компетенций.
хорошо	Оценка «хорошо» выставляется обучающемуся, если он твердо знает материал, грамотно и по существу излагает его, не допуская существенных неточностей в ответе на вопрос, правильно применяет теоретические положения при решении практических вопросов и задач, обладает необходимыми умениями и приемами их выполнения.
удовлетворительно	Оценка «удовлетворительно» выставляется обучающемуся, если он имеет знания только основного материала, но не усвоил его деталей, допускает неточности, недостаточно правильные формулировки, испытывает сложности при выполнении практических работ и затрудняется связать теорию вопроса с практикой.
неудовлетворительно	Оценка «неудовлетворительно» выставляется обучающемуся, который не знает значительной части программного материала, неуверенно отвечает, допускает серьезные ошибки, не имеет представлений по методике выполнения практической работы. Как правило, оценка «неудовлетворительно» ставится обучающимся, которые не могут продолжить обучение без дополнительных занятий по данной учебной дисциплине.

Задания и вопросы для промежуточной аттестации обучающихся приведены в приложении.

14 Учебно-методическое и информационное обеспечение учебной дисциплины

а) литература:

1 Мельниченко, А. С. Математическая статистика и анализ данных : учебное пособие / А. С. Мельниченко. – Москва : МИСиС, 2018. – 45 с. – ISBN 978-5-906953-62-9. – URL: https://www.studentlibrary.ru/book/ISBN9785906953629.html (дата обращения: 05.04.2025);

2 Замятин, А. В. Интеллектуальный анализ данных : учебное пособие / А. В. Замятин. – Москва : Томского ГУ, 2020. – 196 с. – ISBN 978-5-94621-898-6. – URL: https://www.studentlibrary.ru/book/ISBN9785946218986.html (дата обращения: 05.04.2025);

3 Кремер, Н. Ш. Теория вероятностей и математическая статистика : учебник и практикум для вузов / Н. Ш. Кремер. — 5-е изд., перераб. и доп. — Москва : Издательство Юрайт, 2025. — 538 с. — ISBN 978-5-534-10004-4. – URL: https://urait.ru/bcode/565694 (дата обращения: 05.04.2025);

4 Третьяк, Л. Н. Обработка экспериментальных данных: основы теории и практики : учебник для вузов / Л. Н. Третьяк, А. Л. Воробьев ; под общей редакцией Л. Н. Третьяк. — 2-е изд., испр. и доп. — Москва : Юрайт, 2025. — 212 с. — ISBN 978-5-534-08623-2. – URL: https://urait.ru/bcode/563557 (дата обращения: 05.04.2025);

5 Миркин, Б. Г. Введение в анализ данных : учебник и практикум / Б. Г. Миркин. — Москва : Юрайт, 2024. — 174 с. — ISBN 978-5-9916-5009-0. – URL: https://urait.ru/bcode/536117 (дата обращения: 05.04.2025);

6 Статистика : учебник для вузов / под редакцией В. С. Мхитаряна. — 3-е изд., перераб. и доп. — Москва : Юрайт, 2025. — 503 с. — ISBN 978-5-534-18687-1. – URL: https://urait.ru/bcode/568978 (дата обращения: 05.04.2025);

7 Анализ данных : учебник для вузов / под редакцией В. С. Мхитаряна. — Москва : Юрайт, 2025. — 448 с. — ISBN 978-5-534-19964-2. – URL: https://urait.ru/bcode/560311 (дата обращения: 05.04.2025);

8 Сапрыкин, О. Н. Интеллектуальный анализ данных : учебное пособие / О. Н. Сапрыкин. — Самара : Самарский университет, 2020. — 80 с. — ISBN 978-5-7883-1563-8. – URL: https://e.lanbook.com/book/188906 (дата обращения: 05.04.2025);

9 Хруничев, Р. В. Прикладные статистические методы анализа : учебное пособие / Р. В. Хруничев. — Рязань : РГРТУ, 2023. — 80 с. – URL: https://e.lanbook.com/book/380498 (дата обращения: 05.04.2025);

10 Васильев, Е. П. Интеллектуальный анализ данных в технологиях принятия решений : учебное пособие / Е. П. Васильев, В. И. Орешков. — Рязань : РГРТУ, 2023. — 180 с. — ISBN 978-5-7722-0344-6. – URL: https://e.lanbook.com/book/380378 (дата обращения: 05.04.2025);

11 Вольфсон, М. Б. Анализ данных : учебно-методическое пособие / М. Б. Вольфсон. — Санкт-Петербург : СПбГУТ им. М.А. Бонч-Бруевича, 2023. — 69 с. – URL: https://e.lanbook.com/book/381533 (дата обращения: 05.04.2025).

б) ресурсы информационно-телекоммуникационной сети «Интернет»:

1 Электронный каталог : сайт / Научно-техническая библиотека СибГИУ. – Новокузнецк, [199 –]. – URL: http://libr.sibsiu.ru. – URL: https://libr.sibsiu.ru;

2 Электронная библиотека // Научно-техническая библиотека СибГИУ : сайт. – Новокузнецк, [200 –]. – URL: http://library.sibsiu.ru/LibrELibraryFullText.asp. – Режим доступа: для авторизир. пользователей. – URL: https://library.sibsiu.ru/LibrELibraryFullText.asp;

3 Консультант студента : электронно-библиотечная система / ООО «КОНСУЛЬТАНТ СТУДЕНТА». – Москва, [200 –]. – URL: http://www.studentlibrary.ru. – Режим доступа: для авторизир. пользователей;

4 ЛАНЬ : электронно-библиотечная система : [коллекция «Единая профессиональная база знаний для технических вузов - Издательство Лань" ЭБС»] / ООО «Издательство ЛАНЬ». – Санкт-Петербург, [200 –]. – URL: http://e.lanbook.com. – Режим доступа: для авторизир. пользователей;

5 Образовательная платформа ЮРАЙТ / ООО «Электронное издательство ЮРАЙТ». – Москва, [200 –]. – URL: https://urait.ru. – Режим доступа: для авторизир. пользователей;

6 Университетская библиотека онлайн : электронно-библиотечная система / ООО «Директ-Медиа». – Москва, [200 –]. – URL: https://biblioclub.ru. – Режим доступа: для авторизир. пользователей. – URL: http://www.biblioclub.ru;

7 Электронная библиотека УМЦ ЖДТ : [коллекция «Эксплуатация железных дорог»] / ФГБУ ДПО «Учебно-методический центр по образованию на железнодорожном транспорте». – Москва, [2013 –]. – URL: https://umczdt.ru/books/. – Режим доступа: для авторизир. пользователей;

8 НАУЧНАЯ ЭЛЕКТРОННАЯ БИБЛИОТЕКА eLIBRARY.RU : база данных / ООО «НЭБ». – Москва, [200 –]. – URL: http://elibrary.ru. – Режим доступа: по подписке;

9 Электронные периодические издания ИВИС : универсальная база данных / ООО «ИВИС». – Москва, [200 –]. – URL: http://eivis.ru. – Режим доступа: по подписке.

в) лицензионное и свободно распространяемое программное обеспечение:

– 7-Zip;
– Adobe Acrobat Reader;
– Astra Linux Special Edition;
– Deductor Academic;
– Kaspersky Endpoint Security;
– Loginom Community;
– Microsoft Office;
– Microsoft Windows;
– OnlyOffice;
– Р7-Офис.

г) базы данных и информационно-справочные системы:

1 ГАРАНТ : справочно-правовая система / ООО «Правовой центр «Гарант». – Кемерово, [200 –]. – Режим доступа: компьютерная сеть Сиб. гос. индустр. ун-та.;

2 КонсультантПлюс : справочно-правовая система / ООО «Информационный центр АНВИК». – Новокузнецк, [199 –]. – Режим доступа: компьютерная сеть библиотеки Сиб. гос. индустр. ун-та.;

3 Техэксперт : информационно-справочная система / ООО «Группа компаний «Кодекс». – Кемерово, [200 –]. – Режим доступа: компьютерная сеть Сиб. гос. индустр. ун-та.

15 Материально-техническое обеспечение учебной дисциплины

Материально-техническое обеспечение учебной дисциплины включает учебные аудитории, оснащенные оборудованием, компьютерной техникой, и техническими средствами обучения, в том числе:
- учебную аудиторию для проведения занятий лекционного типа, оборудованную учебной доской, экраном и мультимедийным проектором;
- учебную аудиторию для проведения занятий семинарского типа (практических и лабораторных занятий), оснащенную учебной доской, компьютерной техникой с возможностью подключения к сети Интернет и обеспечением доступа в электронную информационно-образовательную среду;
- помещения для самостоятельной работы, оснащенные компьютерной техникой с возможностью подключения к сети Интернет и обеспечением доступа в электронную информационно-образовательную среду, научно-техническую библиотеку СибГИУ;
- учебную аудиторию (помещения) для групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации;

Рабочая программа дисциплины составлена в соответствии с требованиями ФГОС ВО по направлению подготовки (специальности) 09.03.03 «Прикладная информатика».

Составитель(и):

доцент Кораблина Татьяна Валентиновна (кафедра прикладных информационных технологий и программирования).

Рабочая программа дисциплины рассмотрена и утверждена на заседании кафедры.

Согласована:

Генеральный директор
Общества с ограниченной ответственностью
Научно-производственная фирма
"Радиан-М" (ООО НПФ "Радиан-М")
к.т.н., доцент Е.Г. Руденкова

Приложение

Тестовые задания для входного контроля знаний обучающихся

Задание 1 (2 балла). В лотерее разыгрывается 1000 билетов. Из них на один билет выпадает выигрыш 10000 руб., на 10 билетов – выигрыш по 2000 руб., на 50 билетов – выигрыш по 500 руб., на 100 билетов – выигрыш по 100 руб., остальные билеты без выигрыша. Найти вероятность выиграть: а) не менее 500 руб.; б) на менее 1000 руб.

Задание 2 (2 балла). В корзине 15 яблок, среди них 5 красных. Достали 5 яблок (без возвращения). Найти вероятность того, что попались ровно 4 красных.

Задание 3 (2 балла). Имеется три одинаковые коробки с карандашами. В первой коробке два красных и один синий; во второй – три красных и один синий; в третьей – два красных и два синих. Выбирается наугад одна из коробок и вынимается из нее карандаш. Найти вероятность того, что карандаш красный.

Задание 4 (2 балла). Из колоды в 52 карты (4 масти по 13 карт в каждой) наугад выбирается шесть карт. Найти вероятность того, что среди выбранных карт будет по две карты каких-то трех мастей.

Задание 5 (2 балла). Из коробки, в которой лежит 5 белых, 6 красных и 4 черных шара, наугад достается 5 шаров. Найти вероятность того, что будут вынуты шары: а) только одного цвета; б) двух цветов; в) всех цветов.

Задание 6 (2 балла). В лотерее разыгрывается 20 билетов. Из них на один билет выпадает выигрыш 10000 руб., на 1 билет – выигрыш по 2000 руб., на 1 билет – выигрыш по 500 руб., остальные билеты без выигрыша. Построить ряд распределения случайной величины β – суммарного выигрыша по двум купленным билетам.

Задание 7 (2 балла). Случайная величина β имеет следующую таблицу распределения:

β	–2	–1	0	1	2
$p(\beta)$	0,1	0,15	0,4	$\boldsymbol{b}$	0,05

Найти значение $\boldsymbol{b}$, математической ожидание и дисперсию случайно величины β.

Задание 8 (3 балла). Случайная величина X подчинена закону распределения с плотностью:

$$f(x)=\begin{cases}0 & \text{при } x<1,\\ a\cdot x & \text{при } 1<x<3,\\ 0 & \text{при } x>1.\end{cases}$$

Найти значение постоянной a, вычислить математическое ожидание и

дисперсию случайной величины X.

Задание 9 (3 балла). Функция распределения непрерывной случайной величины X задана выражением

$$F(x)=\begin{cases}0 & \text{при } x\le 0,\\ ax^2 & \text{при } 0<x\le 3,\\ 1 & \text{при } x>3.\end{cases}$$

Найти значение постоянной *a*, найти вероятность P попадания величины X на участок от 1 до 2, найти плотность распределения $f(x)$.

Задания и вопросы для текущего контроля знаний обучающихся по разделам (темам) учебной дисциплины

1 Устный опрос на лекции

Раздел 1 Прикладной статистический анализ

Тема 1.1 Закон распределения случайной величины и способы его оценки

1 Что называется генеральной совокупностью?

2 Что называется выборкой?

3 Что такое закон распределения?

4 Какие есть формы записи закона распределения?

5 Как строится вариационный ряд?

6 Какое распределение называется выборочным?

7 Как строится гистограмма? График выборочной функции распределения?

8 Что такое нормальное распределение? В какой типичной ситуации оно появляется?

9 Перечислите основные распределения, используемые в статистических расчетах?

10 Какие критерии используются для проверки гипотез о виде распределения?

Тема 1.2 Статистические характеристики

1 Как вычисляется выборочное среднее? Выборочная дисперсия? Выборочное стандартное отклонение?

2 Перечислите свойства выборочного среднего.

3 Перечислите свойства дисперсии.

4 Что такое мода и как она определяется?

5 Как вычисляется и что характеризует коэффициент асимметрии выборки? Коэффициент эксцесса?

6 Какие из рассмотренных статистических характеристик относятся к характеристикам положения?

7 Какие из рассмотренных статистических характеристик относятся к

характеристикам разброса?

8 Какие оценки параметров называются точечными? Перечислите основные свойства точечных оценок.

9 Какие оценки называются интервальными и для чего они рассчитываются?

Тема 1.3 Многовариантное усреднение данных

1 Дайте определения выборочного среднего, текущего среднего и скользящего среднего?

2 Как будет вести себя текущее среднее по мере увеличения объема выборки?

3 Объясните назначение скользящих средних. Влияние каких компонент временного ряда устраняется с их помощью?

4 Поясните, когда целесообразно использовать простые скользящие средние, а для каких временных рядов предпочтительнее применение взвешенных.

5 Приведите алгоритм расчета простых скользящих средних.

6 В чем отличие алгоритма расчета взвешенных скользящих средних от простых?

7 Сколько значений теряется при использовании скользящей средней с длиной интервала сглаживания $L^*=2p+1$? Какие приемы восстановления потерянных отсчетов после реализации процедур сглаживания используются на практике?

8 Что выступает в качестве настроечного параметра степени сглаженности в алгоритмах скользящего среднего, скользящего медианного среднего, скользящего экспоненциального среднего?

9 Какие существуют способы выбора наилучшего настроечного параметра для алгоритмов сглаживания?

10 Какие из рассмотренных алгоритмов обладают свойством помехозащищенности?

Тема 1.4 Корреляционный анализ

1 Что означает некоррелированность случайных величин X и Y?

2 В каком случае коэффициент корреляции равен по модулю единице?

3 Как проверить значимость коэффициента парной корреляции?

4 Как строится корреляционная матрица? Какими свойствами обладает корреляционная матрица?

5 Для чего определяются частные коэффициенты корреляции?

6 Для чего определяются множественные коэффициенты корреляции?

7 Для чего рассчитывается корреляционное отношение?

8 С помощью какого показателя можно оценить наличие и силу линейных связей между величинами X и Y?

9 С помощью какого показателя можно оценить наличие и силу всех

видов связей между величинами X и Y?

10 С помощью какого показателя можно оценить наличие и силу нелинейных связей между величинами X и Y?

11 Что такое остаточная дисперсия? Что она характеризует?

12 С помощью какого критерия можно выявить связь между двумя порядковыми признаками?

13 В каком случае используются коэффициенты Спирмена и Кэнделла?

Раздел 2 Интеллектуальный анализ данных

Тема 2.1 Введение в интеллектуальный анализ данных

1 Дайте определение понятию Data Mining.

2 Какими свойствами должны обладать скрытые знания «добытые» с помощью Data Mining?

3 Какие задачи решаются методами Data Mining?

4 На стыке каких наук и направлений находятся методы Data Mining?

5 Какие этапы включает в себя интеллектуальный анализ?

6 В каких областях человеческой деятельности используются методы Data Mining?

7 Что повлияло на распространение технологии Data Mining?

Тема 2.2 Задача ассоциации

1 В чем заключается задача поиска ассоциативных правил?

2 В чем заключается секвенциальный анализ?

3 В чем отличие поиска ассоциативных правил от секвенциального анализа?

4 Что такое поддержка ассоциативного правила?

5 Что такое достоверность ассоциативного правила?

6 Что такое улучшение ассоциативного правила?

7 Какие этапы предполагает задача поиска ассоциативных правил?

8 В чем заключается идея алгоритма Apriori поиска ассоциативных правил?

Тема 2.3 Задача кластеризации

1 В чем заключается задача кластеризации?

2 В чем отличие задачи кластеризации от задачи классификации?

3 Что используется в качестве меры «похожести» объектов при решении задачи кластеризации?

4 Какими способами можно представлять результаты кластеризации?

5 В чем заключается алгоритм k-means?

6 В чем достоинства и недостатки алгоритма k-means?

7 Даны два объекта с признаками (1; 3) и (3; 3) соответственно. Определите между ними Евклидово расстояние и расстояние

Манхэттена.

8 Как определить расстояние для категориальных признаков?

Тема 2.4 Задача классификации и регрессии

1 В чем заключается задача классификации?

2 В чем заключается задача регрессии?

3 В чем заключаются этапы решения задачи классификации и регрессии?

4 Какие алгоритмы используются для решения задачи классификации?

5 Из каких частей состоят классификационные правила?

6 Что из себя представляют деревья решений?

7 В чем заключается идея алгоритма классификации 1R-алгоритм?

8 При решении каких практических задач используются алгоритмы классификации и регрессии Data Mining?

3 Тестирование по результатам освоения лекционного материала и самостоятельной работы

Тестирование проводится для определения степени усвоения материалов лекций, а также выполнения самостоятельной работы в рамках часов, отведенных на самостоятельную работу обучающихся, в СУО «Moodle».

Раздел 1 Прикладной статистический анализ

Тема 1.1 Закон распределения случайной величины и способы его оценки

1 В каком диапазоне изменяется значение вероятности случайно величины? (*один ответ*)

а) [0, 1]
б) [–1, 1]
в) [–∞, ∞]
г) [0, ∞]
д) [0, 0,5];
е) нет правильного ответа

2 Случайные величины могут быть: (*несколько ответов*)

а) дискретные;
б) непрерывные;
в) статистические;
г) хаотические.

3 В каких осях строится гистограмма? (*один ответ*)

а) ось абсцисс – значение случайной величины, ось ординат - частота попадания случайной величины в интервал;

б) ось абсцисс – значение случайной величины, ось ординат -

количество данных, попавших в интервал;

в) ось абсцисс – значение случайной величины, ось ординат - среднее значение случайной величины в интервале;

г) ось абсцисс – количество данных в интервале, ось ординат - частота попадания случайной величины в интервал;

д) ось абсцисс – количество данных в интервале, ось ординат - среднее значение случайной величины в интервале;

4 Для эмпирической оценки закона распределения случайной величины необходимо (*один ответ*):

а) построить гистограмму

б) построить график временного ряда данных

в) рассчитать среднее значение случайно величины

г) рассчитать среднее значение и дисперсию

5 Что называют статистической гипотезой? (*один ответ*)

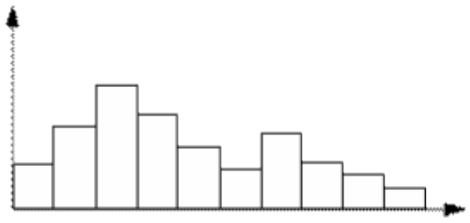
а) предположение о теоретических значениях параметров распределения

б) утверждение о виде неизвестного распределения

в) предположение о виде неизвестного распределения или о параметрах известного распределения

г) утверждение о параметрах известного распределения

6 На рисунке изображена гистограмма, построенная по экспериментальным данным. Каков характер соответствующего закона распределения?



а) унимодальный

б) полимодальный

в) антимодальный

7 Для ответа на какой вопрос используется критерий согласия χ^2 (хи-квадрат)? (*один ответ*)

а) Правильно определен доверительный интервал при заданном значении вероятности *p*?

б) Правильно подобраны теоретические параметры закона распределения?

в) Согласуются между собой статистическое распределение и выбранная теоретическая кривая закона распределения *f*(*x*)?

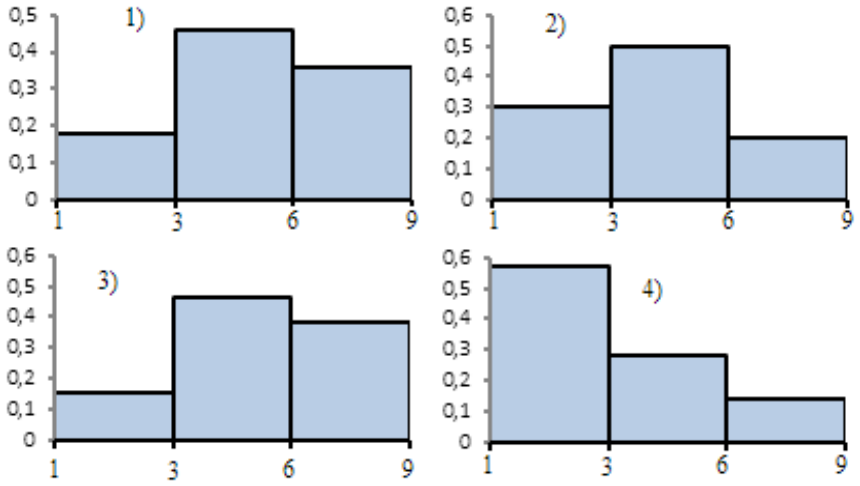
8 Что необходимо сделать после визуального анализа построенной гистограммы? (*один ответ*)

а) проверить, согласуются ли экспериментальные данные с гипотезой о том, что случайная величина подчинена теоретическому закону распределения;

б) подобрать теоретические значения параметров закона распределения;

в) построить статистический ряд.

9 На каком из рисунков изображена гистограмма ряда данных 5, 1, 2, 7, 5, 2, 3, 6, 8, 5, 3, 4, 7, 4, 9, 5, 5, 6, 7, 6, 3, 6, 7, 8, 5



Введите номер рисунка _____

10 Определите порядок следования шагов при построении гистограммы:

№ шага	Описание шага
	определить количество данных попавших в каждый интервал
	определить диапазон изменения случайной величины и разбить его на интервалы
	разделить диапазон изменения случайной величины на интервалы
	построить столбиковую диаграмму
	рассчитать частоту попадания случайной величины в каждвй интервал

11 По данным 5, 1, 2, 7, 5, 2, 3, 6, 8, 5, 3, 4 построить вариационный ряд. Значения ряда вводить через пробел. _____

Тема 1.2 Статистические характеристики

1 Статистические характеристики – это (*один ответ*)

а) закон распределения случайной величины.

б) числовые параметры, которые описывают случайную величину полностью

в) числовые параметры, которые характеризуют степень разбросанности значений случайной величины;

г) числовые параметры, которые отражают существенные черты

случайной величины.

д) ничего из перечисленного;

е) все перечисленное;

2 Мода – это (*один ответ*)

а) наиболее вероятное значение случайной величины;

б) средняя величина, характеризующая ряд данных при последовательном увеличении его реализации

в) средняя величина, характеризующая участок ряда данных при его скольжении вдоль реализации

г) максимально значение случайно величины;

д) минимальное значение случайной величины;

е) некоторое значение, расположенное между минимальным и максимальным значением случайной величины.

3 Коэффициент вариации – это (*один ответ*)

а) некоторое значение, расположенное между минимальным и максимальным значением случайной величины;

б) мера тесноты линейной связи между двумя случайными величинами;

в) наиболее вероятное значение случайной величины;

г) мера рассеяния случайной величины относительно среднего значения;

д) характеристика закона распределения случайной величины с точки зрения его симметрии;

е) характеристика закона распределения случайной величины с точки зрения его «сглаженности»;

ж) квадрат отклонения случайной величины от среднего.

4 Дисперсия – это (*один ответ*)

а) наиболее вероятное значение случайной величины;

мера рассеяния случайной величины относительно среднего значения;

б) мера тесноты линейной связи между двумя случайными величинами;

в) максимально значение случайно величины;

г) некоторое значение, расположенное между минимальным и максимальным значением случайной величины.

д) минимальное значение случайной величины;

5 Какие из приведенных высказываний являются свойствами дисперсии? (*один ответ*)

а) Дисперсия произведения независимой случайной величины на постоянную равна произведению дисперсии случайных величины на квадрат постоянной.

б) Дисперсия суммы независимых случайных величин равно сумме

дисперсий этих случайных величин

в) Дисперсия постоянной равна нулю.

г) Дисперсия произведения случайных величин равно произведению дисперсий этих случайных величин

д) Дисперсия суммы случайных величин равно сумме дисперсий этих случайных величин;

6 Какая из характеристик меры рассеяния случайной величины имеет ту же размерность, что и сама случайная величина? (*один ответ*)

а) среднеквадратическое отклонение;
б) дисперсия;
в) коэффициент вариации;
г) коэффициент асимметрии;
д) эксцесс;
е) коэффициент корреляции.

7 Эксцесс характеризует (*один ответ*)

а) некоторое значение, расположенное между минимальным и максимальным значением случайной величины;

б) рассеяние случайной величины относительно среднего значения;

в) закон распределения случайной величины с точки зрения его симметрии по сравнению с нормальным законом;

г) тесноту линейной связи между двумя случайными величинами;

д) наиболее вероятное значение случайной величины;

е) закон распределения случайной величины с точки зрения его «сглаженности» по сравнению с нормальным законом;

ж) квадрат отклонения случайной величины от среднего.

8 Какие из статистических характеристик относятся к характеристикам положения (*несколько ответов*)

а) выборочное среднее;
б) коэффициент вариации;
в) медиана;
г) эксцесс;
д) мода;
е) межквартильный размах.

9 Какие из статистических характеристик относятся к характеристикам рассеяния (*несколько ответов*)

а) выборочное среднее;
б) коэффициент вариации;
в) медиана;
г) эксцесс;
д) мода;

е) межквартильный размах.

10 Найти медиану для числовой последовательности 24, 97, 12, 86, 96, 25, 11, 58, 16. _____

11 Найти моду для числовой последовательности 10, 4, 10, 4, 3, 6, 10, 8, 8. _____

12 Найти дисперсию для числовой последовательности 6, 2, 2, 7, 4, 3, 7, 5, 6. (Ответ округлить до целого числа). _____

Тема 1.3 Многовариантное усреднение данных

1 Какого вида бывают помехи? (*несколько ответов*)

а) импульсная помеха;
б) средневзвешенная помеха;
в) рекуррентная помеха;
г) скалярная помеха;
д) флуктуационная помеха.

2 Степень сглаженности для алгоритма скользящего медианного среднего определяется (*один ответ*)

а) размером выборки;
б) размером окна скольжения.
в) значением настроечного параметра α.
г) значением весовых коэффициентов;
д) значением настроечного параметра β.

4 Алгоритмический подход к сглаживанию рядов данных (*один ответ*)

а) состоит в допущении, что можно задать общий вид функции, описывающей регулярную, неслучайную составляющую
б) состоит в допущении, что можно задать общий вид функции и дает алгоритм расчета неслучайной составляющей в любой заданный момент времени
в) не предполагает описание динамики неслучайной составляющей с помощью единой функции и дает лишь алгоритм расчета неслучайной составляющей в любой заданный момент времени.
г) не предполагает описание динамики неслучайной составляющей с помощью единой функции, а позволяет определить только вид приближенной функции в форме обобщенного многочлена

5 Настроечный параметр *α* алгоритма релейно-экспоненциального среднего определяет (*один ответ*)

а) значения весовых коэффициентов;
б) степень сглаженности;
в) диапазон возможного изменения сглаженного сигнала;
г) помехозащищенность алгоритма.

6 Настроечный параметр α изменяется в диапазоне (*один ответ*)

а) от 0 до 1;
б) диапазон ограничен только размером выборки;
в) от -1 до 1;
г) от -2 до 2.

7 Помехозащищенными алгоритмами сглаживания являются (*несколько ответов*)

а) скользящего экспоненциального среднего
б) скользящего релейно-экспоненциального среднего.
в) скользящего медианного среднего.
г) скользящего арифметического среднего;
д) скользящего взвешенного арифметического среднего;

8 При сглаживании ряда данных с использованием алгоритма скользящего арифметического среднего более гладким ряд получится при использовании окна скольжения *L** размером (*один ответ*)

а) *L**=12;
б) *L**=5;
в) *L**=10;
г) *L**=4.

9 Причинами возникновения выбросов являются (*несколько ответов*)

а) неисправность измерительных приборов;
б) ошибки оператора;
в) аварийный режим работы технического или технологического оборудования;
г) неправильное использование алгоритмов сглаживание;
д) неправильные настройки алгоритмов сглаживания.

10 При сглаживании временного ряда с окном скольжения *L**=11 (*один ответ*).

а) теряются первые пять и последние пять значений;
б) теряются первые и последние одиннадцать значений;
в) теряются только первые одиннадцать значений;
г) теряются только последние одиннадцать значений;
д) теряются только последние пять значений;
е) исходный и сглаженный ряд имеют одинаковое количество отсчетов.

11 При сглаживании ряда данных с использованием алгоритма скользящего экспоненциального среднего более гладким ряд получится при использовании настроечного параметра *α* (*один ответ*).

а) *α* =0,13;
б) *α* =0,21;

в) α =0,72;
г) α =0,93.

Тема 1.4 Корреляционный анализ

1 В каком диапазоне изменяется значение парного коэффициента корреляции? (*один ответ*)

а) [0, 1];
б) [-1, 1];
в) диапазон не ограничен;
г) [0, ∞].

2 В каком диапазоне изменяется значение частного коэффициента корреляции? (*один ответ*)

а) [0, 1];
б) [-1, 1];
в) диапазон не ограничен;
г) [0, ∞].

3 В каком диапазоне изменяется значение частного множественного корреляции? (*один ответ*)

а) [0, 1];
б) [-1, 1];
в) диапазон не ограничен;
г) [0, ∞].

4 Вы рассчитали коэффициент корреляции и его значение оказалось близкое к нулю. Какое из высказываний в данной ситуации является верным? (*один ответ*)

а) Можно утверждать, что величины независимые.
б) Зная только значение коэффициента корреляции, нельзя утверждать, что случайные величины являются независимыми.
в) Между случайными величинами есть сильная статистическая связь.

5 Вы рассчитали коэффициент взаимной корреляции r = 0,9. Какое из высказываний в данной ситуации является верным? (*один ответ*)

а) Между двумя отсчетами одной случайно величины есть сильная линейная связь.
б) Между двумя случайными величинами линейная связь отсутствует.
в) Между двумя случайными величинами есть сильная линейная связь.
г) Между двумя случайными величинами есть сильная нелинейная связь.
д) Случайные величины независимы.

6 Две величины функционально связаны по нелинейному закону, тогда для них

а) корреляционное отношение =0, коэффициент корреляции =1;
б) корреляционное отношение =1, коэффициент корреляции =0;
в) корреляционное отношение =-1, коэффициент корреляции =-1;
г) корреляционное отношение =1, коэффициент корреляции =1.

7 Найти частный коэффициенты корреляции $r_{xz/y}$, если известны значения парных коэффициентов корреляции r_{xy} = 0,6; r_{xz} = 0,9; r_{yz} = 0,4. (Ответ округлить до одного знака после запятой, например, 1,1). _____

8 Найти множественный коэффициенты корреляции $r_{x/yz}$, если известны значения парных коэффициентов корреляции r_{xy} = 0,6; r_{xz} = 0,9; r_{yz} = 0,4 (ответ округлить до двух знаков после запятой, например, 1,11). _____

9 Найти значение показателя степени нелинейности, если корреляционное отношение η=0,9, а коэффициент корреляции r=0,2. (Ответ округлить до двух знаков после запятой, например, 1,11). _____

10 Найти значение корреляционного отношения, если межгрупповая выборочная дисперсия равна 2; выборочная дисперсия равна 6. (Ответ округлить до двух знаков после запятой, например, 1,11). _____

11 Найти значение коэффициента детерминации, если коэффициент корреляции 0,7.(Ответ округлить до двух знаков после запятой, например, 1,11). _____

12 Если взаимно корреляционное отношение η=0,9 и коэффициент взаимной корреляции r=0,09, то это означает, что (*один ответ*).

а) между случайными величинами есть сильная линейная связь;
б) между случайными величинами есть сильная нелинейная связь;
в) между случайными величинами линейная связь отсутствует;
г) между отсчетами одной случайно величины линейная связь отсутствует;
д) случайные величины независимы.

13 У вас есть четыре ряда данных X1, X2, X3, X4, необходимо определить наличие и силу линейных связей между величиной X2 и всеми остальными величинами вместе X1, X3, X4. Какую характеристику следует рассчитать? (*один ответ*).

а) парный коэффициент корреляции
б) корреляционное отношение
в) частный коэффициент корреляции
г) множественный коэффициент корреляции

д) коэффициент вариации

14 У вас есть четыре ряда данных X1, X2, X3, X4, необходимо определить наличие и силу линейной связи между двумя случайными величинами X2 и X4, исключив влияние величин X1 и X3 на них. Какую характеристику следует рассчитать? (*один ответ*).

а) частный коэффициент корреляции
б) множественный коэффициент корреляции
в) парный коэффициент корреляции
г) коэффициент вариации

15 Определить коэффициент автокорреляции со сдвигом 1 для ряда: 24, 97, 12, 86, 96, 25, 11, 58, 16 (Ответ округлить до двух знаков после запятой, например, 1,11). _____

16 Определить коэффициент взаимной корреляции со сдвигом 0 для рядов данных X={0,70; 0,51; 0,50; 0,60; 0,70; 0,72; 0,67} и Y={0,29; 0,22; 0,20; 0,26; 0,28; 0,26; 0,20} (Ответ округлить до двух знаков после запятой, например 1,11). _____

17 Для независимых случайных величин (*один ответ*).

а) корреляционное отношение =0;
б) корреляционное отношение >0;
в) корреляционное отношение <0;
г) корреляционное отношение =-1.

18 Найти условные средние, если число групп *q* равно 3 (*один ответ*).

x	3	8	6	9	1	9	10	0	7	11
y	79	11	38	59	35	95	47	96	88	24

а) $\widetilde{x}(q_1) = 96{,}7;\ \ \widetilde{x}(q_2) = 90{,}0;\ \ \widetilde{x}(q_3) = 17{,}7;$

б) $\widetilde{x}(q_1) = 90{,}2;\ \ \widetilde{x}(q_2) = 70{,}0;\ \ \widetilde{x}(q_3) = 72{,}0;$

в) $\widetilde{x}(q_1) = 71{,}3;\ \ \widetilde{x}(q_2) = 69{,}5;\ \ \widetilde{x}(q_3) = 18{,}8;$

г) $\widetilde{x}(q_1) = 70{,}0;\ \ \widetilde{x}(q_2) = 63{,}0;\ \ \widetilde{x}(q_3) = 17{,}5;$

д) $\widetilde{x}(q_1) = 34{,}3;\ \ \widetilde{x}(q_2) = 91{,}2;\ \ \widetilde{x}(q_3) = 42{,}7;$

19 Имеется две случайные величины, закон распределения которых не является нормальным. Какой показатель следует рассчитать для изучения связей между ними? (*несколько ответом*).

а) коэффициент корреляции Пирсона;
б) коэффициент корреляции Кенделла;
в) коэффициент корреляции Спирмена;
г) коэффициент вариации.

20 Вы рассчитали коэффициент корреляции Спирмена с доверительной вероятности α = 0,95 и подучили значение ρ=0,27.

Критическое значение $\rho_{0,95}$=0,653. Какой вывод следует сделать относительно значимости корреляции? (*один ответ*)

а) Корреляцию следует признать значимой.
б) Корреляцию следует признать незначимой.
в) О признании значимости корреляции информации недостаточно.

Раздел 2 Интеллектуальный анализ данных

Тема 2.1 Введение в интеллектуальный анализ данных

1 О какой из задач идет речь: «_____ позволяет определить по известным характеристикам объекта значение некоторого его параметра». (*один ответ*)

а) задача классификации;
б) задача кластеризации;
в) задача регрессии;
г) задача поиска ассоциативных правил.

2 О задачах какого типа идет речь? «____ задачи уделяют внимание улучшению понимания анализируемых данных». (*один ответ*)

а) Анализирующие;
б) Предсказательные;
в) Описательные;
г) Качественные.

3 К какому типу относится задача: «Известна частота появления определенных слов. Требуется определить, является ли сообщение спамом»? (*один ответ*)

а) Задача регрессии;
б) Задача кластеризации;
в) Задача классификации;
г) Задача поиска ассоциативных правил.

4 К какому типу относится задача: «Компанию, предоставляющую услуги, интересует, какими услугами клиенты предпочитают пользоваться в совокупности»? (*один ответ*)

а) Задача регрессии;
б) Задача кластеризации;
в) Задача классификации;
г) Задача поиска ассоциативных правил.

5 К какому типу относится задача: «Какой диагноз может быть поставлен пациенту, если известна совокупность симптомов»? (*один ответ*)

а) Задача регрессии;

б) Задача кластеризации;
в) Задача классификации;
г) Задача поиска ассоциативных правил.

6 Какие модели относятся к предсказательным? (*несколько ответов*)

а) итоговые модели;
б) игровые модели;
в) модели последовательностей;
г) модели классификации;
д) ассоциативные модели;
е) модели исключений;

7 Расположите в порядке следования этапы Knowledge Discovery in Databases (KDD).

№ этапа	Описание шага
	Преобразование данных
	Очистка данных
	Интерпретация результатов
	Формирование выборки
	Построение модели Data Mining

8 Расположите в порядке следования этапы построение модели.

№ этапа	Описание шага
	Построение модели, объясняющей имеющиеся факты
	Формулирование цели моделирования
	Тестирование модели и интерпретация результатов
	Сбор и подготовка данных
	Применение полученной модели

9 Какими свойствами должна обладать обучающая выборка? (*несколько ответов*)

а) быть репрезентативной
б) удовлетворять условиям качества
в) быть разбита на две части - обучающую и контрольную
г) отражать правила и закономерности исследуемого процесса
д) быть закодирована на машинном языке

10 Критериями выбора алгоритма являются (*несколько ответов*):

а) устойчивость;
б) производительность;
в) используемые методы;
г) язык программирования;
д) точность.

11 Какого вида является переменная, номинальная или ординальная?

Переменная	Возможные значения	Вид
Образование	Среднее, Среднее специальное, Высшее, Ученая степень	

Цвет	Синий, Красный, Зеленый	
Провайдер	Мегафон, МТС, Билайн	
Карта клиента	Серебряная, Золотая, Платиновая	
Наличие машины	Да, Нет	
Оценка	Неудовлетворительно, Удовлетворительно, Хорошо, Отлично	

12 В какой шкале измеряются переменные, в шкале наименований или в шкале порядка?

Переменная	**Возможные значения**	**Шкала**
Образование	Среднее, Среднее специальное, Высшее, Ученая степень	
Цвет	Синий, Красный, Зеленый	
Провайдер	Мегафон, МТС, Билайн	
Карта клиента	Серебряная, Золотая, Платиновая	
Наличие машины	Да, Нет	
Оценка	Неудовлетворительно, Удовлетворительно, Хорошо, Отлично	

13 Какая из переменных является дискретной? (*несколько ответов*)

а) Вес.
б) Уровень образования
в) Площадь квартиры.
г) Оценка за экзамен

14 Сценарий, создаваемый в аналитической платформе Loginom состоит из (*один ответ*):

а) узлов
б) компонентов
в) функций

15 Через что узлы сценария в аналитической платформе Loginom обмениваются данными? (*один ответ*).

а) порты
б) коннекторы
в) стрелки

16 Как определить, что узел сценария в аналитической платформе Loginom активирован? (*один ответ*).

а) у портов сплошной, а не прерывистый контур
б) узел мигает
в) контур узла зеленого цвета

17 Порт в аналитической платформе Loginom, взаимодействующий с набором переменных, графически отображается … (*один ответ*).

а) прямоугольником

б) точкой
в) полукругом

18 Как в аналитической платформе Loginom быстро создать независимую копию узла с такими же настройками? (*один ответ*).

а) с помощью команды *Копировать узел*
б) с помощью команды *Создать производный компонент*.

19 Сценарий в аналитической платформе Loginom представляет собой… (*один ответ*).

а) древовидную структуру с ветвями
б) направленный граф
в) нумерованный иерархический список

20 При импорте текстового файла в аналитической платформе Loginom у части столбцов конвертация типов данных происходит с потерями. Что может помочь для исправления ситуации? (*несколько ответов*).

а) изменить метку
б) изменить вид данных
в) изменить тип данных
г) изменить десятичный разделитель
д) уменьшить значение параметра Кол-во строк для анализа

21 Для имеющихся экспериментальных данных, которые могут быть описаны функцией $y=a\cdot\exp(b\cdot x)$, необходимо подобрать параметры a и b. Какой инструмент Excel можно использовать для решения задачи? (*один ответ*).

а) Анализ данных
б) Фильтр
в) Поиск решения
г) Анализ «что если»

Тема 2.2 Задача ассоциации

1 Какой из показателей является мерой точности правила? (*один ответ*)

а) достоверность
б) поддержка;
в) лифт;
г) улучшение;
д) левередж.

2 Какой из показателей является обобщенной мерой связи двух предметных наборов (*один ответ*)

а) достоверность;

б) поддержка;
в) лифт;
г) улучшение;
д) левередж.

3 Какой из показателей показывает полезнее ли правило случайного угадывания? (*один ответ*)

а) достоверность;
б) поддержка;
в) лифт;
г) улучшение;
д) левередж.

4 Что означает значение лифта больше 1? (*один ответ*)

а) условие чаще появляется в транзакциях, содержащих следствие, чем в остальных;
б) условие реже появляется в транзакциях, содержащих следствие, чем в остальных;
в) вероятнее предсказать наличие набора, содержащего следствие, с помощью правила, чем угадать случайно;

5 Что означает равенство 1 значение улучшения? (*один ответ*)

а) условие чаще появляется в транзакциях, содержащих следствие, чем в остальных;
б) условие реже появляется в транзакциях, содержащих следствие, чем в остальных;
в) вероятнее предсказать наличие набора, содержащего следствие, с помощью правила, чем угадать случайно;

6 Если к предметному набору, не являющемуся частым, добавить некоторый нового предмет, то он станет популярным? (*один ответ*)

а) да;
б) нет;
в) имеющейся информации недостаточно для ответа на вопрос

7 Количество транзакций, содержащих и условие A и следствие B равно 78, общее количество транзакций 156. Чему будет равна поддержка? (*один ответ*)

а) 0,25;
б) 0,5;
в) 0,75;
г) 2.

8 Какие из однопредметных наборов являются популярными, если граничная частота определена *f*=5? (*несколько ответов*)

№ транзакции	сыр	творог	сметана	кефир	молоко	хлеб	конфеты

1	1	1	0	0	0	0	0
2	0	0	0	1	0	0	0
3	0	0	0	0	1	1	1
4	0	0	1	0	1	1	0
5	0	1	1	0	1	0	1
6	0	0	1	1	1	0	0
7	0	1	1	0	1	0	0
8	0	1	1	0	0	1	1
9	1	0	1	0	1	1	1
10	0	0	0	0	1	1	0

а) сыр;
б) творог;
в) сметана;
г) кефир;
д) молоко;
е) хлеб;
ж) конфеты.

9 Количество транзакций, содержащих и условие А и следствие В равно 50, количество транзакций, содержащих только условие А = 250, общее количество транзакций 1500. Чему будет равно значение достоверности?

а) 0,25;
б) 0,2;
в) 6;
г) 5.

10 Для ассоциативного правила *молоко* → *хлеб* определить достоверность?

№	Транзакция
1	сыр, творог, сметана
2	кефир, конфеты
3	конфеты, молоко, хлеб
4	молоко, хлеб, сметана, картофель, конфеты
5	молоко, печенье, творог, конфеты, сметана
6	молоко, печенье, кефир, сметана
7	молоко, творог, сметана
8	печенье, творог, хлеб, сметана, конфеты
9	молоко, хлеб, сыр, сметана, конфеты
10	молоко, картофель, хлеб

Ответ ввести в виде десятичного числа, округлив до двух знаков после запятой _____ .

Тема 2.3 Задача кластеризации

1 Кластеризацию можно определить, как (*один ответ*)

а) группировка объектов по похожести их свойств; каждый кластер

состоит из схожих объектов, а объекты разных кластеров существенно отличаются.

б) процедура, которая любому объекту $x \in X$ ставит в соответствие метку набору $y \in Y$.

в) исследование взаимной связи между событиями, которые происходят совместно.

г) установление закономерности между входными и выходными переменными

2 Даны два объекта с признаками (1; 3) и (3; 3) соответственно. Определите между ними Евклидово расстояние. (Результат округлить до одного знака после запятой, например, 1,0). _____

3 По какой формуле рассчитывается расстояние между категориальными признаками? (*один ответ*)

а) $d(X,Y) = \sum_i (x_i^2 - y_i^2)$;

б) $d(x_i,y_i) = \begin{cases} 0, \text{если } x_i = y_i, \\ 1, \text{если } x_i \neq y_i. \end{cases}$

в) $d(X,Y) = \sqrt{\sum_i (x_i^2 - y_i^2)}$

г) $d(X,Y) = \sum_i |x_i - y_i|$.

4 Что можно использовать в качестве критерия остановки работы алгоритма кластеризации? (*несколько ответов*)

а) Координаты центроидов перестают изменяться.

б) Ошибка стала меньше установленного значения.

в) Количество объектов в кластере превысило заданное значение.

г) Количество объектов в кластере меньше заданного значения.

5 Даны два объекта с признаками (1; 3; 2) и (3; 3; 1) соответственно. Определите между ними расстояние Манхеттена. (Результат округлить до одного знака после запятой, например, 1,0). _____

6 Признак цвет имеет значения: желтый, зеленый, синий. Какая метрика должна быть использована для определения расстояния между объектами? (*один ответ*)

а) Евклидово расстояние;

б) расстояние Манхеттена

в) Функция отличия.

7 В результате очередного шага в кластер попали объекты с признаками (1; 3; 2), (3; 3; 1), (2; 2; 4), (1; 3; 2). Определить координаты центроида кластера. (Результат округлить до одного знака после запятой, например, 1,0). _____.

8 Даны два объекта с признаками (2; 6; зелёный) и (1; 3; красный) соответственно. Определите между ними Евклидово расстояние.

Результат округлить до одного знака после запятой, например, 1,0. _____

9 При решении задачи кластеризации заранее известно, сколько кластеров будет сформировано?

а) да;
б) нет;
в) можно определить только вероятность получения того или иного количества кластеров.

10 В результате очередного шага кластеризации объекты распределились по двум кластерам с центрами m_1 и m_2 в соответствии с данными в таблице. Рассчитать ошибку, результат округлить до одного знака после запятой, например, 1,0. _____

№ объекта	Расстояние от m_1	Расстояние от m_2	Принадлежность кластеру
1	2,0	2,3	1
2	1,8	2,2	1
3	3,6	2,8	2
4	4,4	3,6	2
5	1,0	1,4	1
6	3,1	2,2	2
7	1,0	1,2	1
8	1,5	1,0	2

Тема 2.4 Задача классификации и регрессии

1 Какие методы классификации относятся к статистическим методам классификации и регрессии? (*несколько ответов*)

а) логистическая регрессия;
б) линейная регрессия;
в) деревья решений;
г) нейронные сети;
д) метод k ближайших соседей.

2 Какой показатель используется оценки качества классификации?

$$C = \frac{\text{кол - во правильно класифицированных примеров в данном узле}}{\text{кол - во ошибочно классифицированных примеров в данном узле}}$$

(*один ответ*)

а) поддержка;
б) улучшение;
в) достоверность;
г) левередж.

3 Вам поручено провести исследования в крупной торговой сети и выделить покупателей с определенными предпочтениями для лучшего удовлетворения их потребностей? Какую из задач интеллектуального анализа вы будете решать? (*один ответ*)

а) кластеризации;
б) построение линии регрессии;
в) поиска ассоциативных правил;
г) классификации.

4 Отделу аналитики банка, которым вы руководите, поручено разработать скоринговую модель для оценки кредитоспособности клиентов? Какую из задач интеллектуального анализа вы будете решать?

а) кластеризации;
б) классификации;
в) поиска ассоциативных правил;
г) построение линии регрессии.

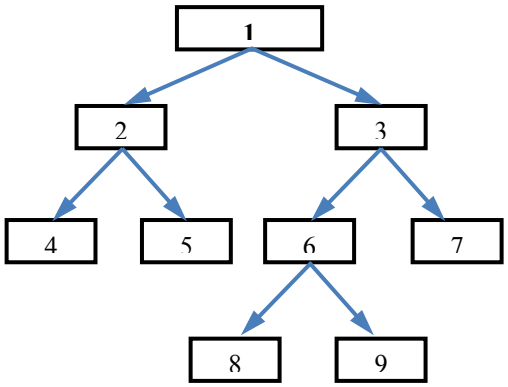
5 При установлении закономерности между входными и выходными переменными и входные, и выходные переменные модели являются непрерывными. Какую из задач интеллектуального анализа вы будете решать?

а) классификации;
б) регрессии;
в) поиска ассоциативных правил;
г) кластеризации;
д) регрессии.

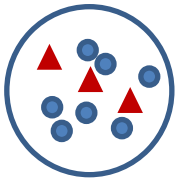
6 К методам машинного обучения для решения задачи классификации и регрессии относятся

а) линейная регрессия;
б) логистическая регрессия;
в) деревья решений;
г) нейронные сети;
д) метод k ближайших соседей.

7 Какие из узлов дерева являются листьями _____ .

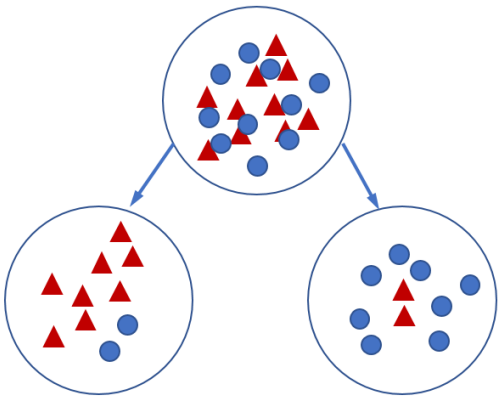


8 При решении задачи классификации после очередного разбиения получено множество. Чему равен индекс Джини?



а) 0,58;
б) 0,85;
в) 0,24;
г) 0,5.

9 Для приведенного разбиения определить энтропию узла нижнего уровня. Ответ округлить до двух знаков после запятой, например, 1,22



_____

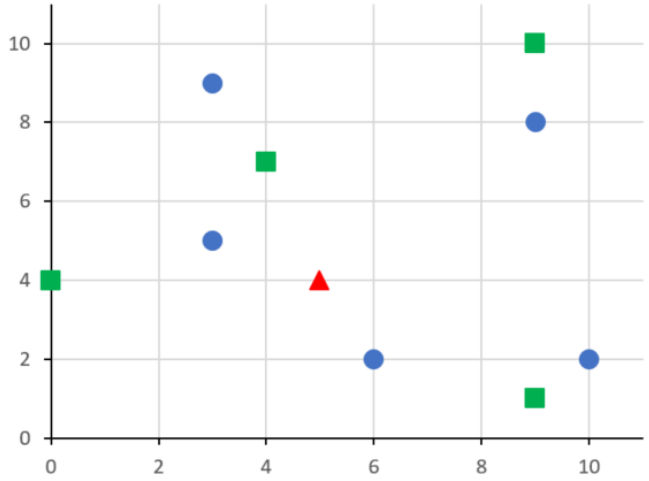
10 Определить ошибку правила «Если (A1=A), то (C=C1)». Ответ записать в виде обыкновенной дроби, например, 2/14

_____

№ п/п	A1	A2	A3	C
1	A	70	Да	C1
2	B	90	Да	C1
3	A	90	Да	C2

4	C	80	Да	C2
5	A	85	Нет	C2
6	C	80	Нет	C1
7	B	65	Да	C1
9	A	95	Нет	C2
10	A	70	Нет	C1

11 На основании использования метода k ближних соседей к каком классу должен быть отнесен тестовый образец (красный треугольник) при k=3?



а) к классу 1 – синих кругов;
б) к классу 2 – зеленых квадратов;
в) для решения задачи недостаточно данных

12 При установлении закономерности между входными и выходными переменными и входные, и выходные переменные модели являются непрерывными. Какую из задач интеллектуального анализа вы будете решать?

а) классификации;
б) регрессии;
в) поиска ассоциативных правил;
г) кластеризации;
д) регрессии.

4 Выполнение практических работ

Задания для практических работ, теоретический и методический материал для самостоятельного изучения и вопросы для собеседования представлены в методических указаниях к практическим занятиям.

Тема 1.1 Закон распределения случайной величины и способы его оценки

1 Для ряда данных построить две гистограммы при разбиении диапазона изменения величины на 5 и 7 интервалов.

2 По результатам визуального анализа полученных гистограмм выдвиньте гипотезу о законе распределения, которому подчинена рассматриваемая случайная величина.

3 Проверить согласованность теоретического и статистического распределения.

Тема 1.2 Статистические характеристики

1 Для ряда данных рассчитать значения выборочных оценок статистических характеристик: среднего, медианы, дисперсии, среднеквадратического отклонения, коэффициента вариации, коэффициента асимметрии и эксцесса.

2 Для полученных оценок в задании 1 выборочного среднего значения и дисперсии найти их доверительные интервалы с вероятностью p = 95 %.

3 В течение 2 лет после освоения нового типа изделия цех выпускал изделие этого типа в двух модификациях, отличающихся технологическим маршрутом изготовления. В самом начале освоения этого типа изделия один из маршрутов обеспечивал получение более дефицитной группы, отличавшейся большей мощностью. Другой маршрут позволял производить с меньшей себестоимостью изделия несколько пониженной мощности. За 2 года этот маршрут подвергся ряду изменений, которые должны были поднять мощность выпускаемых изделий. При проверке гипотезы *о* том, что изделия, выпускаемые по маршруту более дефицитных групп, действительно имеют среднюю мощность выше, чем изделия, выпускаемые по параллельному маршруту, но имеющими меньшую себестоимость. Было известно, что дисперсия для первого маршрута (маршрут более мощных изделий) равна 3200 мВт^2, а дисперсия для второго маршрута равна 3025 мВт^2. Применяя метод статистических гипотез рассмотреть целесообразность ликвидации первого маршрута.

4 Для изучения эффективности нового препарата железа были выбраны две группы пациентов с анемией. В первой группе пациенты в течение двух недель получали новый препарат, а во второй группе – получали плацебо. После этого было проведено измерение уровня гемоглобина в периферической крови. В первой группе средний уровень гемоглобина составил 115,4±1,2 г/л, а во второй - 103,7±2,3 г/л (данные представлены в формате M±m), сравниваемые совокупности имеют нормальное распределение. При этом численность первой группы составила 34, а второй - 40 пациентов. Необходимо сделать вывод о статистической значимости полученных различий и эффективности нового препарата железа.

Тема 1.3 Многовариантное усреднение данных

1 Ряд данных сгладить, используя следующие алгоритмы:

а) скользящее арифметическое среднее (САС),

б) скользящее взвешенное арифметическое среднее (ВСАС),

в) скользящее медианное среднее (СМС),

г) скользящее экспоненциальное среднее (СЭС).

Для алгоритмов САС и СМС задать окно скольжения L^*=7, для алгоритма СЭС рассчитать значение α по эмпирической формуле

При использовании алгоритмов САС и СМС полученное среднее значение по данным, попавшим в окно скольжения, располагать в середине окна. Для ВСАС задать ряд весов, убывающий от конца окна скольжения к началу.

2 Восстановить краевые значения сглаженного ряда САС, используя средний абсолютный прирост.

3 Вывести графики исходного ряда и полученных сглаженных рядов в одних осях. Сравнить их. Чем они отличаются, почему?

4 Повторить расчеты с окном скольжения L^*=11 и соответствующего α. Построить графики. Сравнить графики, полученные в п. 3. Чем они отличаются, почему?

5 Из рассмотренных алгоритмов помехозащищенными являются алгоритмы скользящего медианного среднего (СМС) и релейно-экспоненциального среднего (РЭС). Ряд с выбросами сгладить с использованием алгоритмов РЭС с рассчитанным значением настроечного параметра β и α =0,3, СМС с окном скольжения эквивалентным α и СЭС с α =0,3. Вывести графики исходного ряда и сглаженных рядов в одних осях. Сравнить их между собой. Чем они отличаются и почему?

6 Выбор наилучшего настроечного параметра сглаживающего алгоритма с помощью двухкомпонентного критерия качества сглаживания и воспроизводимости оценок.

Тема 1.4 Корреляционный анализ

1 Для двух рядов данных рассчитать коэффициент взаимной корреляции и проверить гипотезу о наличии корреляции между случайными величинами с достоверностью α = 0,95.

Построить взаимно корреляционную функцию для сдвигов λ =0, ± 1 1, $\pm$ 2, $\pm$ 3. Оформить результаты в таблице и построить график.

2 Для одного ряда данных рассчитать коэффициенты автокорреляции и построить автокорреляционную функцию для сдвигов λ =0, 1, 2, 3. Оформить результаты в таблице и построить график.

3 Для трех рядов данных рассчитать три частных и три множественных коэффициента корреляции. Вычислить *t*-статистики для проверки значимости коэффициентов корреляции и проверить гипотезу о наличии корреляции между случайными величинами с достоверностью α = 0,95. Проверить правильность вычислений, используя контрольные соотношения.

4 Для ряда пар случайных величин (x_ℓ, y_ℓ) установить наличие корреляции одним из «быстрых» критериев Кенуя или Кокс-Стюарта.

5 Проверить наличие корреляции с помощью коэффициента ρ Спирмена при доверительной вероятности α = 0,95.

Тема 2.3 Задача кластеризация

1 На основании данных в соответствии с вариантом проведите кластеризацию по двум признакам ирисов, используя алгоритм k-means.

2 На основании данных, приведенных в таблице 10, в соответствии с вариантом провести кластеризацию магазинов по трем признакам: X_1 – площадь торгового зала, $м^2$, X_2 – товарооборот на одного продавца, ден. ед., X_3 – уровень рентабельности, %. Построить дендрограмму.

Тема 2.4 Задача классификация и регрессия

1 На основании набора данных (в соответствии с вариантом), используя алгоритм «*k* ближайших соседей», определить к какому из трех классов: Ирис щетинистый (*Iris setosa*), Ирис виргинский (*Iris virginica*), Ирис разноцветный (*Iris versicolor*), относится пример последней записи в наборе данных.

Задачу решить при *k*=1 и *k*=3.

В качестве метрик для определения расстояния между записями использовать Евклидово расстояние и расстояние Манхеттена. Сравнить результаты классификации при использовании разных метрик.

2 Проанализируйте зависимость стоимости продаваемого жилья от набора показателей с помощью регрессии на наборе данных о продажах. Докажите, почему регрессия подходит для использования. Визуализируйте коэффициенты и обученную модель.

3 Делают ли деньги людей счастливыми? Какие факторы влияют на индекс счастья? Ответьте на эти вопросы, проанализировав набор данных с помощью линейной регрессии.

4 По данным варианта построить дерево решений, используя алгоритм ID3. В качестве критерия выбора атрибута разбиение применить прирост энтропии. Используя построенное дерево определить неизвестное значение выходной переменной.

5 Используя алгоритм ID3 постройте дерево решений, позволяющее предсказать, выжил ли пассажир «Титаника» в зависимости от информации о нём.

5 Выполнение лабораторных работ

Задания для лабораторных работ, теоретический и методический материал для самостоятельного изучения и вопросы для собеседования представлены в методических указаниях к лабораторным работам.

Лабораторная работа 1 Знакомство с аналитической платформой

1 Установите Loginom на своем компьютере и убедитесь, что программа запускается.

2 Создайте новый пакет и сохраните его под именем *test.lgp.*

3 Импортируйте данные из текстового файла *Продажи за месяц.txt*. Определить месяцы, в которые объем продаж находится в диапазоне от 6000 до 8000 тыс. руб. Экспортируйте полученные результаты в текстовый файл *Месяцы.txt*.

4 Откройте созданный ранее в пакете *test.lgp* сценарий, состоящий из узла импорта данных из файла *Продажи за месяц.txt*, узла фильтрации и узда экспорта*.*

5 Используя Переменный и узел Калькулятор произведите расчет стоимости товара по текущему курсу валют.

Лабораторная работа 2 Основы работы с аналитической платформой

1 Загрузите файл с данными в аналитическую платформу Loginom. Задайте полям метки.

2 Аналитику необходимо оценить кредитоспособность потенциального заемщика. Предполагается, что заемщики, берущие суммы разного диапазона, ведут себя по-разному, следовательно, модели прогноза должны быть свои для каждой группы, т. е. для дальнейшего построения моделей оценки кредитоспособности различных категорий заемщиков необходимо использовать фильтрацию. Определите группу заемщиков, взявших кредит менее 10000 рублей и живущих в области.

3 Произвести группировку для данных о продажах строительных материалов по сумме и количеству по городам.

4 При оценке качества данных о месячных продажах имеется два выброса и одно пропущенное значение в марте месяце. Используя Мастер *Редактирование выбросов* и Мастер *Заполнение пропусков* отредактируйте выбросы и замените пропущенное значение применяя различные методы. Сравните полученные результаты. Какие методы по вашему мнению больше подходят для исследуемых данных? Используя визуализатор, постройте графики.

5 Тренд ряда данных об объеме месячных продаж имеет сезонную составляющую. Подберите из имеющихся алгоритмов сглаживания обработки такой, который позволяет выделить полезную (медленную) составляющую сигнала с сезонностью наилучшим образом. С помощью визуализатора постройте графики.

6 Используя обработчик *Квантование*, разбить всех клиентов банка на группы в соответствии с размером взятого кредита.

Лабораторная работа 3 Предварительная обработка и исследование данных

1 Произвести обогащение данных с использованием узлов Слияние, Соединение, Дополнение данных, Объединение.

2 Преобразование данных формата Дата/Время.

3 Используя обработчики *Автокорреляция* и *Корреляционный анализ* провести корреляционный анализ данных.

4 Используя обработчик *Факторный анализ* по данным о клиентах банка выявить факторы, влияющие на возврат кредита.

5 Представьте информацию об объеме продаж, используя визуализаторы Диаграмма, Качество данных, Куб, Статистика, Таблица. Сформируйте соответствующие отчеты

Лабораторная работа 4 Задача поиска ассоциативных плавил

Розничная сеть по продаже товаров бытовой химии поставила задачу анализа покупательских корзин для оптимизации размещения товаров на витринах и проведения кросс-продаж. Отдел маркетинга предоставил 5000 чеков, в которых отражены покупки, сделанные клиентами магазинов. Требуется:

1. Предсказать, какие товары покупатели могут выбрать в зависимости от того, что уже есть в их корзинах;
2. Выявить наиболее популярные товарные наборы, состоящие из более чем одного предмета;
3. Предложить рекламные акции типа: «Каждому купившему *A* и *B* – товар *C* в подарок».

Исходные данные представлены в файле *Чеки.txt* двумя полями – *Номер транзакции* и *Товар.* Поскольку номенклатура товаров бытовой химии очень разнообразна, решено ограничиться представлением товаров в обобщенной форме без торговых марок: порошки, моющие средства и т. д. (всего 37 наименований).

Лабораторная работа 5 Задача кластеризации

1 Выполните кластеризацию по данным о параметрах цветков ириса в аналитической платформе Loginom. Обработчик производит кластеризацию объектов на основе алгоритмов *k-means*, если количество кластеров заранее известно, и *g-means*, если количество кластеров определяется автоматически в рамках заданного интервала.

2 По данным абонентов телекоммуникационной компании провести сегментацию клиентов и оценить наиболее и наименее доходные сегменты.

3 Используя алгоритм *k-means* выполните кластеризацию по данным о состоянии банков. Проанализируйте статистически свойства кластеров, определите их мощность.

4 По данным об обеспеченности населения лечебными учреждениями регионов Российской Федерации в 2018 году провести сегментацию регионов и выяснить существуют ли заметные различия в

их обеспеченности населения лечебными учреждениями с помощью алгоритмов *k-means*, *g-means* и *EM кластеризации*.

5 На основании данных о продажах в розничной сети произвести анализ потребительской корзины с помощью кластеризации транзакций для выявления совместно приобретаемых товаров в супермаркете.

Лабораторная работа 6 Задача классификации

Используя данные кредитных историй, постройте скоринговую модель на основе логистическое регрессии. Определите оптимальный порог и порог, при котором чувствительность модели 90%. Настроенную модель проверьте на новых данных. Проанализируйте полученные результаты. В какой класс попали данные заемщики?

Лабораторная работа 7 Задача регрессии

1 Найдите уравнения линейной регрессии по данным о продажах квартир, использовав в качестве входного поля *Количество комнат*, *Жилая площадь*, *Площадь кухни.* Оцените значимость коэффициентов и значимость полученных уравнений регрессии. Постройте графики.

2 Постройте регрессионные модели с помощью узла *Линейная регрессия* и проанализируйте, влияние на стоимость квартиры факторов *Количество комнат* (x_1), *Общая площадь* (x_2), *Жилая площадь* (x_3), *Площадь кухни* (x_4), ручной настройке регрессии.

3 На основе данных о числе амбулаторно-поликлинических организаций в Российской Федерации построить кубическую регрессионную модель. Оценить ее статистическую значимость. Построить диаграмму с исходными и модельными данными.

6 Вопросы к экзамену для промежуточной аттестации

1 Закон распределения случайной величины.

2 Статистический ряд. Построение гистограммы.

3 Статистические характеристики положения.

4 Статистические характеристики рассеяния.

5 Статистические характеристики, характеризующие закон распределения случайной величины.

6 Интервальные оценки статистических характеристик. Доверительная волосность.

7 Доверительный интервал для среднего.

8 Доверительная интервал для дисперсии и среднеквадратического отклонения.

9 Текущее среднее арифметическое. Варианты скользящего арифметического среднего. Взвешенное среднее арифметическое.

10 Медиана. Скользящая медиана.

11 Экспоненциальное среднее. Релейно-экспоненциальное среднее.

12 Использование алгоритмов скользящих средних для построения

прогностических моделей

13 Коэффициент ковариации. Автоковариационная и взаимоковариационная функции.

14 Коэффициент авто- и взаимокорреляции. Автокорреляционная и взаимокорреляционная функции.

15 Частные и множественные коэффициенты корреляции.

16 Методика определения корреляционного отношения. Показатель степени нелинейности

17 Непараметрический корреляционный анализ

18 Классификация задач Data Mining.

19 Основные этапы анализа.

20 Подготовка исходных данных.

21 Задача поиска ассоциативных правил. Алгоритм Apriori.

22 Задача кластеризации. Алгоритмы k-means и g-means.

23 Правила классификации.

24 Статистические методы классификации

25 Построение деревьев решений

26 Построение математических функций.

7 Задачи к экзамену для промежуточной аттестации

1 Для заданных рядов данных независимой Y и зависимой X переменных найти условные средние, если число групп равно q=3.

2 Для заданного ряда данных определить коэффициент автокорреляции при λ=1. Оценить значимость корреляции.

3 Для заданных рядов данных рассчитать парный коэффициент корреляции при λ=0. Оценить значимость корреляции.

4 Найти множественные коэффициенты корреляции, если задана корреляционная матрица. Оценить значимость корреляции.

5 Для заданных рядов данных независимой Y и зависимой X переменных рассчитать корреляционное отношение при λ=0 для рядов через остаточную дисперсию (q=2).

6 Для заданных рядов данных независимой Y и зависимой X переменных рассчитать корреляционное отношение при λ=0 для рядов через межгрупповую выборочную дисперсию (q=4).

7 Для данной числовой последовательности построить гистограмму. Провести визуальный анализ и выдвинуть гипотезу о законе распределения

8 Для ряда данных построить ряд скользящего среднего с окном скольжения 3 и ряд скользящего экспоненциального среднего с настроечным параметром α эквивалентным заданному окну скольжения.

9 Для ряда данных, содержащего выбросы, построить ряд скользящего экспоненциального среднего и релейно-экспоненциального среднего, настроив параметр β.

10 Для ряда данных найти точечные и интервальные оценки среднего, дисперсии и среднеквадратического отклонения, соответствующие доверительной вероятности p = 0,8.

11 Для ряда данных найти точечные и интервальные оценки дисперсии и среднеквадратического отклонения, соответствующие доверительной вероятности p = 0,8.

12 Для ряда данных определить непараметрические статистические характеристики.

13 Найти частные коэффициенты корреляции, если задана корреляционная матрица. Оценить значимость корреляции.

14 Заданный ряд данных разбить на обучающую и контрольную выборку, на обучающей выборке настроить прогностический алгоритм на основе скользящего экспоненциального среднего с горизонтом прогноза τ=1. Проверить модель на контрольной выборке. В качестве критерия использовать среднемодульное отклонение.

15 Для заданного ряда данных подобрать наилучший настроечный параметр α скользящего экспоненциального среднего, исплльзуя двухкомпонентный критерий качества сглаживания и воспроизводимости оценок.

14 Создать проект Loginom. Импортировать в него файл с данными. Проверить ряд на наличие выбросов и отредактировать их, используя различные методы. Визуализировать результаты обработки. Какой метод на ваш лучше подходит для заданного ряда?

15 Создать проект Loginom. Импортировать в него файл с данными. Подберите из имеющихся алгоритмов сглаживания обработки такой, который позволяет выделить полезную (медленную) составляющую сигнала с сезонностью наилучшим образом. С помощью визуализатора постройте графики.

16 Создать проект Loginom. Для данных о клиентах банка необходимо обнаружить и заполнить пропуски, обнаружить и отредактировать выбросы. Выполните квантование данных о возрасте клиентов.

17 Создать проект Loginom. По данным, характеризующим социально-экономическое состояние регионов определите, какие параметры оказывают большее влияние на среднедушевой доход в регионе, а какие можно исключить из рассмотрения.

18 На основании данных, приведенных в таблице, провести кластеризацию по двум признакам, используя алгоритм k-means

№ п/п	Признак 1	Признак 2
1	5,1	3,8
2	7	3,2
3	4,6	3,2
4	6,4	3,2
5	6,9	3,1

6	5,3	3,7
7	5,5	2,3
8	5	3,3

19 На основании данных, приведенных в таблице, найти расстояние Манхэттена и Евклидово расстояние между признаками.

№ п/п	Признак 1	Признак 2	Признак 2
1	100	160	25
2	130	200	30
3	80	180	20
4	40	100	22
5	150	90	15

20 Используя данные транзакций определить двухпредметные и трехпредметные популярные наборы. Набор считается популярным, если он встречается три раза и более.

№	**Транзакция (D)**
1	Капуста, перец, кукуруза
2	Спаржа, кабачки, кукуруза
3	Кукуруза, помидоры, фасоль, кабачки
4	Перец, кукуруза, помидоры, фасоль
5	Фасоль, спаржа, капуста
6	Кабачки, спаржа, фасоль, помидоры
7	Помидоры, кукуруза
8	Капуста, помидоры, перец
9	Кабачки, спаржа, фасоль
10	Фасоль, кукуруза
11	Перец, капуста, фасоль, кабачки
12	Спаржа, фасоль, кабачки
13	Кабачки, кукуруза, спаржа, фасоль
14	Кукуруза, перец, помидоры, фасоль, капуста

21 Используя данные транзакций определить для правил «Если {*фасоль*}, то{*кабачки*}» и «Если {*перец*}, то{*капуста*}» поддержку, достоверность, лифт, левередж. Какое из правил является более полезным?

№	**Транзакция (D)**
1	Капуста, перец, кукуруза
2	Спаржа, кабачки, кукуруза

3	Кукуруза, помидоры, фасоль, кабачки
4	Перец, кукуруза, помидоры, фасоль
5	Фасоль, спаржа, капуста
6	Кабачки, спаржа, фасоль, помидоры
7	Помидоры, кукуруза
8	Капуста, помидоры, перец
9	Кабачки, спаржа, фасоль
10	Фасоль, кукуруза
11	Перец, капуста, фасоль, кабачки
12	Спаржа, фасоль, кабачки
13	Кабачки, кукуруза, спаржа, фасоль
14	Кукуруза, перец, помидоры, фасоль, капуста

22 На основании данных, приведенных в таблице, найти прирост информации при использовании атрибута x_1 для разбиения исходного множества.

№ п/п	x_1	x_2	x_3	x_4	y
	Соперник по турнирной таблице	Место проведения	Наличие лидера в команде	Погодные условия	Результат
1	выше	дома	на месте	осадки	поражение
2	выше	дома	на месте	без осадков	победа
3	выше	дома	пропускает	без осадков	победа
4	ниже	дома	пропускает	без осадков	победа
5	ниже	в гостях	пропускает	без осадков	поражение
6	ниже	дома	пропускает	осадки	победа
7	выше	в гостях	на месте	осадки	поражение
8	ниже	в гостях	на месте	без осадков	?

23 На основании набора данных, используя алгоритм «k ближайших соседей», определить к какому из трех классов, относятся примеры в наборе данных, для которых не определена принадлежность к какому-либо классу. Задачу решить при k=2 и k=3. В качестве метрики для определения расстояния между записями использовать Евклидово расстояние и расстояние Манхэттена. Сравнить полученные результаты.

24 Создать проект Loginom. На основании набора данных построить уравнение множественной линейной регрессии. Визуализируйте коэффициенты и обученную модель.

25 Создать проект Loginom. На данных контрольной выборки обучите модель логистической регрессии. Проверьте модель на контрольной выборке. Визуализируйте полученные результаты.