Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых» (ВлГУ)

Институт экономики и туризма

УТВЕРЖДАЮ:
Директор института
_____ Козлов Д.А.
«11» сентября 2023 года

ФОНД ОЦЕНОЧНЫХ МАТЕРИАЛОВ (СРЕДСТВ)

ПО ДИСЦИПЛИНЕ

«ИНСТРУМЕНТАЛЬНЫЕ СИСТЕМЫ РАБОТЫ С ДАННЫМИ»

наименование дисциплины

направление подготовки / специальность

01.03.05 Статистика

(код и наименование направления подготовки (специальности)

направленность (профиль) подготовки

«Бизнес-аналитика»

(направленность (профиль) подготовки))

Владимир, 2023

1. ПЕРЕЧЕНЬ КОМПЕТЕНЦИЙ И ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ

Формируемые компетенции (код, содержание компетенции)	Планируемые результаты обучения по дисциплине, в соответствии с индикатором достижения компетенции		Наименование оценочного средства
	Индикатор достижения компетенции *(код, содержание индикатора)*	Результаты обучения по дисциплине	
1	2	3	4
ОПК-3. Способен осознанно применять методы математической и дескриптивной статистики для анализа количественных данных, в том числе с применением необходимой вычислительной техники и стандартных компьютерных программ, содержательно интерпретировать полученные результаты, готовить статистические материалы для докладов, публикаций и других аналитических материалов	ОПК-3.1. Знает современный статистический и математический инструментарий для решения профессиональных задач ОПК-3.2. Умеет использовать информационно-коммуникационные технологии и программные средства для анализа количественных данных ОПК-3.3. Владеет навыками интерпретации полученных результатов анализа количественных данных и подготовки материалов для докладов, публикаций и других аналитических материалов	Знать инструментальные системы работы с данными, применяемые для математической и дескриптивной статистики в анализе количественных данных; Уметь использовать инструментальные системы работы с данными, применяемые для математической и дескриптивной статистики в анализе количественных данных; Владеть навыками применения инструментальных систем работы с данными для решения профессиональных задач для математической и дескриптивной статистики в анализе количественных данных	Тестовые и ситуационные задания. Практически задания
ОПК-4. Способен понимать принципы работы современных информационных технологий и использовать их для решения задач профессиональной деятельности	ОПК-4.1. Знает принципы работы современных информационных технологий ОПК-4.2. Умеет выбирать информационные технологии ОПК-4.3. Владеет навыками использования современных информационных технологий при решении задач профессиональной деятельности	*Знать:* - основные инструментальные системы работы с данными, используемые в профессиональной деятельности; *Уметь:* - использовать основные инструментальные системы работы с данными в профессиональной деятельности; - осуществлять выбор инструментальных систем работы с данными для решения профессиональных задач *Владеть* - навыками применения инструментальных систем работы с данными для решения профессиональных задач	Тестовые и ситуационные задания. Практически задания

2. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ДЛЯ ПРОВЕДЕНИЯ ТЕКУЩЕГО КОНТРОЛЯ УСПЕВАЕМОСТИ ПО ДИСЦИПЛИНЕ

Рейтинг-контроль 1

Тестовое задание. Критерии оценки.
Максимальное количество баллов – 15.
За каждое правильно выполненное задание – 1,5 балла.

1. Укажите один верный ответ:
Под «большими данными» понимают:

а. наборы данных, размер которых превосходит возможности типичных баз данных по занесению, хранению, управлению и анализу информации
б. наборы данных: от 1000 мегабайт (1 гигабайт) до сотен гигабайт,
в. наборы данных: от 1000 гигабайт (1терабайт) до нескольких терабайт,
г. наборы данных: от нескольких терабайт до сотен терабайт,

2. Вставьте пропущенное слово:

Big Data - это серия подходов, инструментов и методов … структурированных и неструктурированных данных огромных объёмов и значительного многообразия для получения воспринимаемых человеком результатов, эффективных в условиях непрерывного прироста, распределения по многочисленным узлам вычислительной сети, альтернативных традиционным системам управления базами данных и решениям класса Business Intelligence

3. Укажите несколько правильных ответов:

Укажите направления, на которых сосредотачиваются для решения вопросов управления большими данными:

а. Volume – объем;
б. Velocity – скорость;
в. Variety – разнообразие;
г. Utility – полезность;

4. Укажите несколько правильных ответов:

Основными принципами работы с большими данными являются:

а. Горизонтальная масштабируемость.
б. Вертикальная масштабируемость.
в. Отказоустойчивость.
г. Локальность данных.

5. Сопоставьте названия фаз жизненного «пути» данных (или, по-другому, истории данных) внутри организации и их описание

Наименование:

1. Data Capture;
2. Data Maintenance;
3. Data Synthesis;
4. Data Usage
5. Data Publication
6. Data Archival
7. Data Purge

Описание:

а. создание или сбор значений данных, которые еще не существуют и никогда не существовали в компании
б. передача данных в точки, где происходит синтез данных и их использование в форме, наиболее подходящей для этих целей.
в. создание ценности из данных через индуктивную логику (занимается логическими процессами умозаключений от частного к общему - индукция), использование других данных в качестве входных данных.
г. применение данных как информации для задач, которые должно ставить и выполнять предприятие
д. отправка данных в место за пределами предприятия, например, ежемесячных отчетов клиентам, после чего эти данные де-факто невозможно отозвать
е. копирование данных в среду, где они хранятся, до тех пор, пока не понадобятся снова для активного использования и удаления из всех активных производственных сред
ж. удаление каждой копии элемента данных с предприятия

6. Укажите один верный ответ:

В широком смысле интеллектуальный анализ данных – это современная концепция анализа данных, предполагающая, что:

а. данные могут быть неточными, неполными (содержать пропуски), противоречивыми, разнородными, косвенными, и при этом иметь гигантские объёмы; поэтому понимание данных в конкретных приложениях требует значительных интеллектуальных усилий;
б. сами алгоритмы анализа данных могут обладать «элементами интеллекта», в частности, способностью обучаться по прецедентам, то есть делать общие выводы на основе частных наблюдений;
в. процессы переработки сырых данных в информацию, а информации в знания уже не могут быть выполнены по старинке вручную, и требуют нетривиальной автоматизации;
г. все вышеперечисленное

7. **Вставьте пропущенное слово**

… является ключевым элементом Big Data и представляет собой совокупность подходов, инструментов и методов обнаружения в больших массивах данных, накапливающихся в информационных системах компаний, ранее неизвестных, нетривиальных, практически полезных и доступных для интерпретации знаний, необходимых для принятия управленческих решений

8. Вставьте пропущенное слово

Операционная аналитика – это интегрированные автоматические процессы принятия решений, предписывающие и реализующие действия в пределах «времени принятия решения».

9. Укажите несколько правильных ответов:

Способы машинного обучения делятся на несколько категорий:

а. обучение с учителем;
б. обучение без учителя;
в. обучение с подкреплением;
г. обучение с редукцией

10. Сопоставьте наименование способа машинного обучения и его описание

Наименование:
1) обучение с учителем;
2) обучение без учителя;
3) обучение с подкреплением

Описание:

а. обучение осуществляется принудительно с помощью примеров «стимул-реакция»;
б. обучение осуществляется спонтанно без вмешательства извне);
в. обучение осуществляется при взаимодействии с некоторой средой).

Ключи к тесту

№ вопроса	Ответ	№ вопроса	Ответ
1	а)	**6**	г)
2	обработки	**7**	Data Mining
3	а) б) в)	**8**	Операционная
4	а) в) г)	**9**	а) б) в)
5	1а) 2б) 3в) 4г) 5д) 6е)	**10**	1а) 2б) 3в)

Рейтинг-контроль 2

Тестовое задание. Критерии оценки.
Максимальное количество баллов – 15.
За каждое правильно выполненное задание – 1,5 балла.

1. Укажите несколько правильных ответов:

Предварительная подготовка данных включает в себя:

а. очистку

б. отбор экземпляров
в. нормализацию
г. преобразование данных
д. выделение признаков
е. отбор признаков

2. Вставьте пропущенное слово

Очистка данных используется для обнаружения, исправления или удаления … записей в наборе данных

3. Вставьте пропущенное слово

… данных используется для стандартизации диапазона значений независимых переменных или признаков данных (например, сведение к интервалам [0, 1] или [-1, +1]);

4. Сопоставьте наименование методов, которые применяются на этапе предварительной обработки данных с их описанием

Наименование:

1. Очистка данных;
2. Нормализация данных;
3. Преобразование данных;
4. Выделение признаков;
5. Уплотнение данных.

Описание:

а. Очистка данных используется для обнаружения, исправления или удаления ошибочных записей в наборе данных;
б. Нормализация данных используется для стандартизации диапазона значений независимых переменных или признаков данных;
в. Преобразование данных используется для приведения данных в формат, который ожидает аудитория;
г. Выделение признаков используется для преобразования входных данных в набор признаков, которые они хорошо представляют;
д. Уплотнение данных используется для преобразования числовых данных в исправленный, упорядоченный и упрощённый вид. Это помогает уменьшить количество и/или размерность данных.

5. Укажите несколько правильных ответов:

Наиболее типичные предметные области, подлежащие очистке и исправлению в корпоративных информационных системах

а. сведения о лицах
б. сведения об организациях,
в. адресная информация,
г. контактная информация,

6. Укажите несколько правильных ответов:

Для обработки пропущенных значений используются следующие методы:

а. исключение пропущенных значений,
б. присваивание значения null,
в. присваивание статического значения (например, 0 или среднего арифметического),
г. вычисление значения на основании предполагаемого или теоретического распределения,
д. независимое моделирование значения.

7. Укажите несколько правильных ответов

Для обработки противоречий используют следующие методы:

а. исключение противоречивых значений,
б. объединение записей с агрегированием числовых значений выходных атрибутов,
в. вычисление значения на основании предполагаемого или теоретического распределения,

г. независимое моделирование значения.

8. Укажите несколько правильных ответов

Для обработки аномальных значений и выбросов используют следующие методы:

а. исключение аномальных значений и выбросов,
б. подавление аномальных значений и выбросов,
в. вычисление значения на основании предполагаемого или теоретического распределения,
г. независимое моделирование значения.

9. Укажите несколько правильных ответов

Оптимизация данных как элемент предобработки включает

а. снижение размерности
б. выявление и исключение незначащих признаков
в. вычисление значения на основании предполагаемого или теоретического распределения,
г. независимое моделирование значения.

10. Укажите несколько правильных ответов

Уменьшение размерности данных может быть осуществлено методами

а. выбора признаков
б. выделения признаков
в. синтеза признаков
г. удаления признаков

Ключи к тесту

№ вопроса	Ответ	№ вопроса	Ответ
1	а) б) в) г) д) е)	**6**	а) б) в) г) д)
2	ошибочных	**7**	а) б)
3	Нормализация	**8**	а) б)
4	1а) 2б) 3в) 4г) 5д)	**9**	а) б)
5	а) б) в) г)	**10**	а) б)

Рейтинг-контроль 3

Тестовое задание. Критерии оценки.

Максимальное количество баллов – 30.

За каждое правильно выполненное задание – 3 балла.

г) лицо, оказывающее услуги по предоставлению вычислительной мощности для размещения информации в информационной системе, постоянно подключенной к сети "Интернет"

1. Вставьте пропущенное слово:

Целями **…** в зависимости от решаемой задачи являются:

1) расширение знаний о предметной области;

2) сжатие данных (замена множества данных, входящим в кластер, одним представителем);

3) обнаружение новизны (данных, которые невозможно отнести ни к одному из кластеров).

2. Укажите несколько правильных ответов:

Регрессионный анализ – одна из важнейших областей машинного обучения, представляющий собой набор статистических методов исследования регрессии, позволяющих:

а) определить степень детерминированности вариации критериальной переменной предикторами;

б) предсказать значение критериальной переменной с помощью предикторов;

в) определить вклад отдельных предикторов в вариацию критериальной переменной;

г) предсказать, к какой категории принадлежит тот или иной объект, на основе его

характеристик

3. Укажите несколько правильных ответов:

Одним из основных методов обучения персептрона является метод коррекции ошибки. Существует следующие модификации метода коррекции ошибок, которые отличаются между собой в зависимости способом выбора величины и знака подкрепления:

а) метод коррекции ошибок без квантования;

б) метод коррекции ошибок с квантованием;

в) метод коррекции ошибок со случайным знаком подкрепления;

г) метод коррекции ошибок со случайными возмущениями.

4. Вставьте пропущенное слово:

Пассивно-агрессивные алгоритмы – это семейство алгоритмов, используемых для крупномасштабного обучения. Их особенность заключается в том, что обучение осуществляется не на основе одновременно поступающего на вход пакета данных, а на основе входных данных, которые поступают в **…** порядке (модель машинного обучения обновляется шаг за шагом по мере поступления новых данных).

5. Сопоставьте названия методов и их описание

Методы:

1) методы опорных векторов;

2) метод дерева решений;

3) метод случайного леса;

4) наивный байесовский метод.

Описание методов:

а) набор схожих алгоритмов для автоматической классификации объектов, при котором осуществляется перевод исходных векторов в пространство более высокой размерности и осуществляется поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве

б) алгоритм классификации объектов, при котором осуществляется построение древовидной структуры, состоящий из решающих правил вида «если ..., то ...», которые генерируются автоматически в процессе обучения на основании обобщения множества отдельных наблюдений.

в) алгоритм автоматической классификации, который предполагает использование большого ансамбля решающих деревьев, каждое из которых само по себе даёт результат очень невысокого качества, но за счёт их большого количества качество результата повышается до приемлемого уровня.

г) алгоритм классификации, который предполагает использование теоремы Байеса со строгими (наивными) предположениями о статистической независимости признаков

6. Расположите в правильной последовательности:

Наивный байесовский метод – алгоритм классификации, который предполагает использование теоремы Байеса со строгими (наивными) предположениями о статистической независимости признаков

Классификация с применением наивного байесовского метода проводится в следующем порядке:

1) для классифицируемого объекта вычисляются функции правдоподобия каждого из классов;

2) по функциям правдоподобия каждого из классов вычисляются апостериорные вероятности классов;

3) объект относится к тому классу, для которого апостериорная вероятность максимальна

7. Укажите один неверный ответ:

Персептрон является простейшим видом искусственных нейронных сетей. В основе персептрона лежит математическая модель восприятия информации мозгом. Персептрон в своем виде представляет систему из элементов следующих разных типов:

а) сенсоров;

б) нейронов;

в) ассоциативных элементов;
г) реагирующих элементов.

8. Вставьте пропущенное слово:

Одним из основных методов обучения персептрона является метод коррекции …, который представляет собой такой метод обучения, при котором вес связи не изменяется до тех пор, пока текущая реакция персептрона остается правильной (при появлении неправильной реакции вес (величина подкрепления) изменяется на единицу, а знак подкрепления (+/-) определяется противоположным от знака ошибки).

9. Вставьте пропущенное слово:

Метод k-ближайших соседей – метрический алгоритм для автоматической … объектов, при котором объект присваивается тому классу, который является наиболее распространённым среди k соседей данного элемента, классы которых уже известны.

10. Укажите несколько правильных ответов:

К особенностям метода k-ближайших соседей относят:
а) не требуется осуществления обучения перед выполнением прогнозов в реальном времени (в результате метод является легкореализуемым, позволяет добавлять новые данные, имеет высокую скорость выполнения);
б) может быть применим к выборкам с большим количеством атрибутов (многомерным);
в) перед применением необходимо определить функцию расстояния (метрику).
г) строить разделяющую поверхность с использованием только небольшого подмножества точек, лежащих в зоне, критической для разделения, тогда как остальные, верно, классифицируемые наблюдения обучающей выборки вне этой зоны игнорируются

Ключи к тесту

№ вопроса	Ответ	№ вопроса	Ответ
1	кластеризации	**6**	1-2-3
2	а) б) в)	**7**	б)
3	а) б) в) г)	**8**	ошибки
4	последовательном	**9**	классификации
5	1а) 2б) 3в) 4г)	**10**	а) б) в)

Иные оценочные материалы для проведения текущего контроля успеваемости

Практико-ориентированные и ситуационные задания

Критерии оценки практико-ориентированного или ситуационного задания

Оценка	Критерии оценивания
5 баллов	задание выполнено, сделаны в целом корректные выводы.
4 балла	задание в целом выполнено, но допущены одна-две незначительных ошибки логического или фактического характера, сделаны выводы
3 балла	задание выполнено отчасти, допущены ошибки логического или фактического характера, предпринята попытка сформулировать выводы
2 балла	допущены серьезные ошибки логического и фактического характера, выводы отсутствуют
1 балл	содержание задания не осознано, продукт неадекватен заданию
0 баллов	задание не выполнено

Ситуационное задание 1.

На основании учебных данных о стоимости жилья в г. Бостоне необходимо

сегментировать объекты недвижимости по трех любым их признакам.

Кластеризацию провести одним из методов итеративной кластеризации, иерархической кластеризации и плотностной кластеризации. Сделать выводы об эффективности применения каждого из них

Ситуационное задание 2.

На основании данных о транзакциях по кредитным картам необходимо выявить те из них, которые являются мошенническими.

Источник данных: https://www.kaggle.com/mlg-ulb/creditcardfraud

В наборе данных признаки V1, V2,… V28 – это основные компоненты, значения которых в целях безопасности были преобразованы; признак «Time» содержит секунды, прошедшие между каждой транзакцией и первой транзакцией в наборе данных; признак «Amount» – это сумма транзакции; признак «Class» – принадлежность транзакции к одному из двух классов (1 –мошенничество).

Необходимо:

1) загрузить набор данных;

2) провести исследование данных;

3) выполнить предварительную обработку данных (при необходимости);

4) выбрать метод построения модели классификации (решение обосновать);

5) выбрать метрику оценки качества модели (решение обосновать);

6) осуществить обучение (на случайной выборке в размере 90%.) и тестирование (на случайной выборке в размере 10%.) модели классификации с использованием различных комбинаций её параметров;

7) оценить качество различных по параметрам вариантов модели по выбранной метрике;

8) осуществить классификацию на основании варианта модели с лучшим значением метрики.

Ситуационное задание 3.

На основании данных о розничных продажах в электронной торговле США необходимо спрогнозировать индекс ECOMPCTSA (доля розничных продаж в электронной торговле в общем объеме продаж).

Горизонт прогнозирования выбран: временной интервал в 12 месяцев

Источник данных: https://fred.stlouisfed.org/series/ECOMPCTSA

Необходимо:

1) загрузить набор данных;

2) провести исследование данных;

3) выполнить предварительную обработку данных (при необходимости);

4) выбрать несколько методов построения модели регрессии (решение обосновать);

5) выбрать метрику оценки качества модели (решение обосновать);

6) осуществить обучение и тестирование моделей регрессии с использованием различных комбинаций их параметров;

7) оценить качество различных по параметрам вариантов моделей по выбранной метрике;

8) спрогнозировать целевой показатель на основании варианта модели с лучшим значением метрики на заданном горизонте.

Задачи

Критерии оценки решения задачи

Оценка	Критерии оценивания
5 баллов	задача решена полностью, используется простой, логичный и адекватный методический инструментарий, в представленном решении обоснованно

	получен правильный ответ.
4 балла	задача решена полностью, но нет достаточного обоснования или при верном решении допущена вычислительная ошибка, не влияющая на правильную последовательность рассуждений, и, возможно, приведшая к неверному ответу.
2 балла	задача решена частично.
0 баллов	решение неверно или отсутствует.

Задания для семестровой контрольной работы

Номер варианта заданий определяется по последней цифре списка группы в аудиторном журнале.

При решении *задачи* в ходе выполнения письменной контрольной работы не следует описывать теорию вопроса, однако расчёты или умозаключения должны сопровождаться пояснениями их сути, а также анализом как исходных данных, так и получаемых результатов. Теоретические вопросы в каждом варианте требуют подробного изложения.

Задание 1.

Дать письменный подробный ответ на поставленные вопросы.

№	Теоретические вопросы
1	Методы оптимизации данных: снижение размерности, выявление и исключение незначащих признаков Классификация как задача машинного обучения.
2	Сущность моделирования данных как этапа проведения интеллектуального анализа данных. Линейные модели классификации.
3	Модели классификации с нелинейными разделяющими поверхностями Сущность машинного обучения
4	Модели кластеризации. Способы машинного обучения: обучение с учителем, обучение без учителя, обучение с подкреплением.
5	Кластерный анализ как задача машинного обучения Модели машинного обучения для решения прикладных задач: регрессии, классификации, кластеризации, понижения размерности данных, выявления аномалий.
6	Сущность понятия «Big Data». Модели регрессии
7	Принципы и подходы к управлению Big Data. Сущность регрессионного анализа.
8	Содержание и задачи процесса управления большими данными. Линейные модели регрессии.
9	Проблемы использования Big Data Нелинейные модели регрессии
10	Системы хранения больших данных Модели классификации.

Задание 2

Выполните следующие задания:

1. Сформулируйте задачи, возникающие в заданном по вариантам виде деятельности, которые можно было бы решить с использованием машинного обучения (необходимо выделить как минимум по одной задачи регрессии, классификации и кластеризации).

2. Опишите каждую задачу по следующей схеме: сущность задачи, класс задачи, состав признаков (наименование, тип данных, ограничения на значения), состав меток (наименование, тип данных, ограничения на значения).

Варианты видов деятельности: 1) банковская деятельность; 2) электронная коммерция; 3) риэлтерская деятельность; 4) информационная безопасность; 5) розничная торговля; 6) сельское хозяйство;7) транспортные услуги; 8) туристические услуги; 9) услуги страхования.

Задание 3

Выполните следующие задания:

1. На основании вариантов видов деятельности, найдите во внешних открытых репозитариях несколько наборов данных им соответствующих и сделайте их описание по следующей схеме:

а) наименование репозитария с указанием его интернет-адреса,

б) краткое и полное наименование набора данных,

в) идентификатор набора данных,

г) краткая характеристика (решаемые задачи, первоисточник данных и т.д.),

д) количественные параметры набора данных (размер выборки),

е) описание признаков и меток.

2. Проведите анализ данных в наборе на доступность, точность, взаимосвязанность, полноту, непротиворечивость, однозначность, релевантность, надежность и своевременность.

Варианты видов деятельности: 1) банковская деятельность; 2) электронная коммерция; 3) риэлтерская деятельность; 4) информационная безопасность; 5) розничная торговля; 6) сельское хозяйство;7) транспортные услуги; 8) туристические услуги; 9) услуги страхования.

Задание 4

Выполните следующие задания:

1. На основании набора данных о покупках в розничном магазине, занимающемся продажей подарков и сувениров через интернет-сайт, осуществите кластеризацию заказов и покупателей. Источник данных: https://archive.ics.uci.edu/ml/datasets/online+retail

2. На основании данных о транзакциях по кредитным картам необходимо выявить те из них, которые являются мошенническими. Источник данных: https://www.kaggle.com/mlg-ulb/creditcardfraud

3. На основании данных о производстве конфет в США с января 1972 по настоящий момент необходимо спрогнозировать индустриальный продуктовый индекс (IPG3113N – универсальный индекс уровня производства, который измеряется как % от уровня производства 2012 года). Горизонт прогнозирования: временной интервал в 24 месяца. Источник данных: https://fred.stlouisfed.org/series/IPG3113N

3. ПРОМЕЖУТОЧНАЯ АТТЕСТАЦИЯ ПО ДИСЦИПЛИНЕ

Перечень вопросов к зачёту

1. Сущность интеллектуального анализа данных.
2. Интеллектуальный анализ данных в бизнесе (описание стандартных задач).
3. Технологии интеллектуального анализа.
4. Обзор программных решений для интеллектуального анализа данных.
5. Основные этапы проведения интеллектуального анализа данных и их краткая характеристика.
6. Основные методы анализа данных.
7. Внешние и внутренние источники данных: системы управления основными и вспомогательными процессами предприятия, специализированные базы данных.
8. Импорт, экспорт, интеграция данных из внешних и внутренних источников.
9. Обзор ключевых операций с наборами однородных и неоднородных данных.
10. Методы описательного анализа данных.
11. Понятие качества данных.
12. Критерии качества данных: доступность, точность, взаимосвязанность, полнота, непротиворечивость, однозначность, релевантность, надежность и своевременность.
13. Сущность очистки данных.

14. Методы очистки данных при обработке пропущенных значений, дубликатов, противоречий, аномальных значений и выбросов, шумов, фиктивных значений и ошибок ввода данных.

15. Сущность оптимизация данных.

16. Методы оптимизации данных: снижение размерности, выявление и исключение незначащих признаков.

17. Сущность моделирования данных как этапа проведения интеллектуального анализа данных.

18. Сущность машинного обучения.

19. Способы машинного обучения: обучение с учителем, обучение без учителя, обучение с подкреплением.

20. Модели машинного обучения для решения прикладных задач: регресии, классификации, кластеризации, понижения размерности данных, выявления аномалий.

21. Модели регрессии.

22. Сущность регрессионного анализа.

23. Линейные модели регрессии.

24. Нелинейные модели регрессии

25. Модели классификации.

26. Классификация как задача машинного обучения.

27. Линейные модели классификации.

28. Модели классификации с нелинейными разделяющими поверхностями.

29. Модели кластеризации.

30. Кластерный анализ как задача машинного обучения.

31. Сущность понятия «Big Data».

32. Принципы и подходы к управлению Big Data.

33. Содержание и задачи процесса управления большими данными.

34. Проблемы использования Big Data.

35. Системы хранения больших данных

Зачёт проводится в устной форме.

Максимальное количество баллов, которое студент может получить на зачёте, в соответствии с Положением о рейтинговой системе комплексной оценки знаний обучающихся в ВлГУ составляет 40 баллов.

Оценка в баллах	**Критерии оценивания компетенций**
Менее 20 баллов	Студент не знает значительной части программного материала (менее 50% правильно выполненных заданий от общего объема работы), допускает существенные ошибки, неуверенно, с большими затруднениями выполняет практические работы, не подтверждает освоение компетенций, предусмотренных рабочей программой
20 баллов	Студент показывает знания только основного материала, но не усвоил его деталей, допускает неточности, недостаточно правильные формулировки, в целом, не препятствует усвоению последующего программного материала, нарушения логической последовательности в изложении программного материала, испытывает затруднения при выполнении практических работ, подтверждает освоение компетенций, предусмотренных рабочей программой на минимально допустимом уровне
30 баллов	Студент показывает 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, владеет необходимыми навыками и приемами их выполнения, допуская некоторые неточности; демонстрирует хороший уровень освоения материала, информационной и коммуникативной культуры и в целом подтверждает освоение компетенций, предусмотренных рабочей программой
40 баллов	Студент глубоко и прочно усвоил программный материал, исчерпывающе, последовательно, четко и логически стройно его излагает, умеет тесно увязывать теорию с практикой, свободно справляется с задачами, вопросами и другими видами применения знаний, причем не затрудняется

	с ответом при видоизменении заданий, использует в ответе материал монографической литературы, правильно обосновывает принятое решение, владеет разносторонними навыками и приемами выполнения практических задач, подтверждает полное освоение компетенций, предусмотренных рабочей программой

Максимальная сумма баллов, набираемая студентом по дисциплине в течение семестра равна 100.

Оценка в баллах	**Оценка по шкале**	**Обоснование**	**Уровень сформированности компетенций**
91 - 100	«зачтено»	Теоретическое содержание курса освоено полностью, без пробелов необходимые практические навыки работы с освоенным материалом сформированы, все предусмотренные программой обучения учебные задания выполнены, качество их выполнения оценено числом баллов, близким к максимальному	***Высокий уровень***
74-90	«зачтено»	Теоретическое содержание курса освоено полностью, без пробелов, некоторые практические навыки работы с освоенным материалом сформированы недостаточно, все предусмотренные программой обучения учебные задания выполнены, качество выполнения ни одного из них не оценено минимальным числом баллов, некоторые виды заданий выполнены с ошибками	***Продвинутый уровень***
61-73	«зачтено»	Теоретическое содержание курса освоено частично, но пробелы не носят существенного характера, необходимые практические навыки работы с освоенным материалом в основном сформированы, большинство предусмотренных программой обучения учебных заданий выполнено, некоторые из выполненных заданий, возможно, содержат ошибки.	***Пороговый уровень***
60 и менее	«не зачтено»	Теоретическое содержание курса не освоено, необходимые практические навыки работы не сформированы, выполненные учебные задания содержат грубые ошибки	Компетенции не сформированы

4. ИТОГОВЫЕ ТЕСТОВЫЕ ЗАДАНИЯ ПО ДИСЦИПЛИНЕ

№ п/п	Контролируемые разделы (темы)	Тестовые задания с вариантами ответов	Код контролируемой компетенции (или ее части)
1	Понятие инструментальных систем работы с данными и его роль в решении профессиональных задач	**Тесты** **1. Укажите один верный ответ:** Под «большими данными» понимают: д. наборы данных, размер которых превосходит возможности типичных баз данных по занесению, хранению, управлению и анализу информации е. наборы данных: от 1000 мегабайт (1 гигабайт) до сотен гигабайт, ж. наборы данных: от 1000 гигабайт (1терабайт) до нескольких терабайт, з. наборы данных: от нескольких терабайт до сотен терабайт, **2. Вставьте пропущенное слово:** Big Data - это серия подходов, инструментов и методов … структурированных и неструктурированных данных огромных объёмов и значительного многообразия для получения воспринимаемых человеком результатов, эффективных в условиях непрерывного прироста, распределения по многочисленным узлам вычислительной сети, альтернативных традиционным системам управления базами данных и решениям класса Business Intelligence **3. Укажите несколько правильных ответов:** Укажите направления, на которых сосредотачиваются для решения вопросов управления большими данными: д. Volume – объем; е. Velocity – скорость;	ОПК-3 ОПК-4

		ж. Variety – разнообразие; з. Utility – полезность; **4. Укажите несколько правильных ответов:** Основными принципами работы с большими данными являются: д. Горизонтальная масштабируемость. е. Вертикальная масштабируемость. ж. Отказоустойчивость. з. Локальность данных. **5. Сопоставьте названия фаз жизненного «пути» данных (или, по-другому, истории данных) внутри организации и их описание** Наименование: 8. Data Capture; 9. Data Maintenance; 10. Data Synthesis; 11. Data Usage 12. Data Publication 13. Data Archival 14. Data Purge Описание: з. создание или сбор значений данных, которые еще не существуют и никогда не существовали в компании и. передача данных в точки, где происходит синтез данных и их использование в форме, наиболее подходящей для этих целей. к. создание ценности из данных через индуктивную логику (занимается логическими процессами умозаключений от частного к общему - индукция), использование других данных в качестве входных данных. л. применение данных как информации для задач, которые должно ставить и выполнять предприятие м. отправка данных в место за пределами предприятия, например, ежемесячных отчетов клиентам, после чего эти данные де-факто невозможно отозвать н. копирование данных в среду, где они хранятся, до тех пор, пока не понадобятся снова для активного использования и удаления из всех активных производственных сред о. удаление каждой копии элемента данных с предприятия **6. Укажите один верный ответ:** В широком смысле интеллектуальный анализ данных – это современная концепция анализа данных, предполагающая, что: д. данные могут быть неточными, неполными (содержать пропуски), противоречивыми, разнородными, косвенными, и при этом иметь гигантские объёмы; поэтому понимание данных в конкретных приложениях требует значительных интеллектуальных усилий; е. сами алгоритмы анализа данных могут обладать «элементами интеллекта», в частности, способностью обучаться по прецедентам, то есть делать общие выводы на основе частных наблюдений; ж. процессы переработки сырых данных в информацию, а информации в знания уже не могут быть выполнены по старинке вручную, и требуют нетривиальной автоматизации; з. все вышеперечисленное 7. **Вставьте пропущенное слово** … является ключевым элементом Big Data и представляет собой совокупность подходов, инструментов и методов обнаружения в больших массивах данных, накапливающихся в информационных системах компаний, ранее неизвестных, нетривиальных, практически полезных и доступных для интерпретации знаний, необходимых для принятия управленческих решений **8. Вставьте пропущенное слово** Операционная аналитика – это интегрированные автоматические процессы принятия решений, предписывающие и реализующие действия в пределах «времени принятия решения». **9. Укажите несколько правильных ответов:** Способы машинного обучения делятся на несколько категорий: д. обучение с учителем; е. обучение без учителя;	

		ж. обучение с подкреплением; з. обучение с редукцией **10. Сопоставьте наименование способа машинного обучения и его описание** Наименование: 1) обучение с учителем; 2) обучение без учителя; 3) обучение с подкреплением Описание: г. обучение осуществляется принудительно с помощью примеров «стимул-реакция»; д. обучение осуществляется спонтанно без вмешательства извне); е. обучение осуществляется при взаимодействии с некоторой средой). **Вопросы** 1. Что представляет собой фаза Data Capture в соответствии с моделью жизненного «пути» данных (или, по-другому, истории данных) внутри организации Малькольма Чисхолма? 2. Что представляет собой фаза Data Maintenance в соответствии с моделью жизненного «пути» данных (или, по-другому, истории данных) внутри организации Малькольма Чисхолма? 3. Что представляет собой фаза Data Usage в соответствии с моделью жизненного «пути» данных (или, по-другому, истории данных) внутри организации Малькольма Чисхолма? 4. Что представляет собой фаза Data Publication в соответствии с моделью жизненного «пути» данных (или, по-другому, истории данных) внутри организации Малькольма Чисхолма? 5. Перечислите проблемы использования Big Data 6. Перечислите категории задач, решаемых с помощью машинного обучения 7. Перечислите основные шаги при построении моделей машинного обучения 8. Что из себя представляет машинное обучение? 9. Перечислите прикладные задачи интеллектуального анализа данных в розничной торговле 10. Перечислите прикладные задачи интеллектуального анализа данных в банковской деятельности 11. Перечислите прикладные задачи интеллектуального анализа данных в области телекоммуникаций 12. Перечислите прикладные задачи интеллектуального анализа данных в области страхования	
2	Сбор данных, описательный анализ и предварительная обработка данных	**Тесты** **11. Укажите несколько правильных ответов:** Предварительная подготовка данных включает в себя: ж. очистку з. отбор экземпляров и. нормализацию к. преобразование данных л. выделение признаков м. отбор признаков **12. Вставьте пропущенное слово** Очистка данных используется для обнаружения, исправления или удаления … записей в наборе данных **13. Вставьте пропущенное слово** … данных используется для стандартизации диапазона значений независимых переменных или признаков данных (например, сведение к интервалам [0, 1] или [-1, +1]); **14. Сопоставьте наименование методов, которые применяются на этапе предварительной обработки данных с их описанием** Наименование: 6. Очистка данных; 7. Нормализация данных; 8. Преобразование данных; 9. Выделение признаков;	ОПК-3 ОПК-4

		10. Уплотнение данных. Описание: е. Очистка данных используется для обнаружения, исправления или удаления ошибочных записей в наборе данных; ж. Нормализация данных используется для стандартизации диапазона значений независимых переменных или признаков данных; з. Преобразование данных используется для приведения данных в формат, который ожидает аудитория; и. Выделение признаков используется для преобразования входных данных в набор признаков, которые они хорошо представляют; к. Уплотнение данных используется для преобразования числовых данных в исправленный, упорядоченный и упрощённый вид. Это помогает уменьшить количество и/или размерность данных. **15. Укажите несколько правильных ответов:** Наиболее типичные предметные области, подлежащие очистке и исправлению в корпоративных информационных системах д. сведения о лицах е. сведения об организациях, ж. адресная информация, з. контактная информация, **16. Укажите несколько правильных ответов:** Для обработки пропущенных значений используются следующие методы: е. исключение пропущенных значений, ж. присваивание значения null, з. присваивание статического значения (например, 0 или среднего арифметического), и. вычисление значения на основании предполагаемого или теоретического распределения, к. независимое моделирование значения. **17. Укажите несколько правильных ответов** Для обработки противоречий используют следующие методы: д. исключение противоречивых значений, е. объединение записей с агрегированием числовых значений выходных атрибутов, ж. вычисление значения на основании предполагаемого или теоретического распределения, з. независимое моделирование значения. **18. Укажите несколько правильных ответов** Для обработки аномальных значений и выбросов используют следующие методы: д. исключение аномальных значений и выбросов, е. подавление аномальных значений и выбросов, ж. вычисление значения на основании предполагаемого или теоретического распределения, з. независимое моделирование значения. **19. Укажите несколько правильных ответов** Оптимизация данных как элемент предобработки включает д. снижение размерности е. выявление и исключение незначащих признаков ж. вычисление значения на основании предполагаемого или теоретического распределения, з. независимое моделирование значения. **20. Укажите несколько правильных ответов** Уменьшение размерности данных может быть осуществлено методами д. выбора признаков е. выделения признаков ж. синтеза признаков з. удаления признаков **Вопросы** 13.Что представляет собой очистка данных? 14. Что представляет собой выделение признаков? 15. Перечислите основные критерии качества данных при использовании	

		технологий интеллектуального анализа 16. Что включает очистка данных при использовании технологий интеллектуального анализа 17. Какие методы используются для обработки противоречий в данных? 18. Какие методы используются для обработки шумов в данных? 19. Перечислите методы выбора признаков, используемые для уменьшения размерности данных при проведении интеллектуального анализа 20. Перечислите методы выделения признаков, используемые для уменьшения размерности данных при проведении интеллектуального анализа	
3	Моделирование данных на основе методов машинного обучения	**Тесты** **21. Вставьте пропущенное слово:** Целями **…** в зависимости от решаемой задачи являются: 1) расширение знаний о предметной области; 2) сжатие данных (замена множества данных, входящим в кластер, одним представителем); 3) обнаружение новизны (данных, которые невозможно отнести ни к одному из кластеров). **22. Укажите несколько правильных ответов:** Регрессионный анализ – одна из важнейших областей машинного обучения, представляющий собой набор статистических методов исследования регрессии, позволяющих: а) определить степень детерминированности вариации критериальной переменной предикторами; б) предсказать значение критериальной переменной с помощью предикторов; в) определить вклад отдельных предикторов в вариацию критериальной переменной; г) предсказать, к какой категории принадлежит тот или иной объект, на основе его характеристик **23. Укажите несколько правильных ответов:** Одним из основных методов обучения персептрона является метод коррекции ошибки. Существует следующие модификации метода коррекции ошибок, которые отличаются между собой в зависимости способом выбора величины и знака подкрепления: а) метод коррекции ошибок без квантования; б) метод коррекции ошибок с квантованием; в) метод коррекции ошибок со случайным знаком подкрепления; г) метод коррекции ошибок со случайными возмущениями. **24. Вставьте пропущенное слово:** Пассивно-агрессивные алгоритмы – это семейство алгоритмов, используемых для крупномасштабного обучения. Их особенность заключается в том, что обучение осуществляется не на основе одновременно поступающего на вход пакета данных, а на основе входных данных, которые поступают в **…** порядке (модель машинного обучения обновляется шаг за шагом по мере поступления новых данных). **35. Сопоставьте названия методов и их описание** Методы: 1) методы опорных векторов; 2) метод дерева решений; 3) метод случайного леса; 4) наивный байесовский метод. Описание методов: а) набор схожих алгоритмов для автоматической классификации объектов, при котором осуществляется перевод исходных векторов в пространство более высокой размерности и осуществляется поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве б) алгоритм классификации объектов, при котором осуществляется построение древовидной структуры, состоящий из решающих правил вида «если ..., то ...», которые генерируются автоматически в процессе обучения на основании обобщения множества отдельных наблюдений. в) алгоритм автоматической классификации, который предполагает использование большого ансамбля решающих деревьев, каждое из которых само по себе даёт результат очень невысокого качества, но за счёт их	ОПК-3 ОПК-4

		большого количества качество результата повышается до приемлемого уровня. г) алгоритм классификации, который предполагает использование теоремы Байеса со строгими (наивными) предположениями о статистической независимости признаков **26. Расположите в правильной последовательности:** Наивный байесовский метод – алгоритм классификации, который предполагает использование теоремы Байеса со строгими (наивными) предположениями о статистической независимости признаков Классификация с применением наивного байесовского метода проводится в следующем порядке: 1) для классифицируемого объекта вычисляются функции правдоподобия каждого из классов; 2) по функциям правдоподобия каждого из классов вычисляются апостериорные вероятности классов; 3) объект относится к тому классу, для которого апостериорная вероятность максимальна **27. Укажите один неверный ответ:** Персептрон является простейшим видом искусственных нейронных сетей. В основе персептрона лежит математическая модель восприятия информации мозгом. Персептрон в своем виде представляет систему из элементов следующих разных типов: а) сенсоров; б) нейронов; в) ассоциативных элементов; г) реагирующих элементов. **28. Вставьте пропущенное слово:** Одним из основных методов обучения персептрона является метод коррекции …, который представляет собой такой метод обучения, при котором вес связи не изменяется до тех пор, пока текущая реакция персептрона остается правильной (при появлении неправильной реакции вес (величина подкрепления) изменяется на единицу, а знак подкрепления (+/-) определяется противоположным от знака ошибки). **29. Вставьте пропущенное слово:** Метод k-ближайших соседей – метрический алгоритм для автоматической … объектов, при котором объект присваивается тому классу, который является наиболее распространённым среди k соседей данного элемента, классы которых уже известны. **30. Укажите несколько правильных ответов:** К особенностям метода k-ближайших соседей относят: а) не требуется осуществления обучения перед выполнением прогнозов в реальном времени (в результате метод является легкореализуемым, позволяет добавлять новые данные, имеет высокую скорость выполнения); б) может быть применим к выборкам с большим количеством атрибутов (многомерным); в) перед применением необходимо определить функцию расстояния (метрику). г) строить разделяющую поверхность с использованием только небольшого подмножества точек, лежащих в зоне, критической для разделения, тогда как остальные, верно, классифицируемые наблюдения обучающей выборки вне этой зоны игнорируются **Вопросы** 21. Перечислите способы машинного обучения 22. Перечислите виды задач, решаемые с помощью машинного обучения 23. В чем суть метода стохастического градиентного спуска? 24. Перечислите основные области классификации как задачи машинного обучения 25.Перечислите свойство линейной регрессии 26. Перечислите классы нелинейной регрессии 27. Перечислите области классификации 28. Перечислите допущения методы логистической регрессии 29. Перечислите цели кластеризации	

		30. В чем сущность итеративной кластеризация?	

Критерии формирования оценок

Критерии оценки результатов тестирования
(max – 15 баллов за тест)

Баллы оценки	Критерии оценки
0-15	Студент получает 0,25 балл за каждый правильный ответ на тест из 60 вопросов.

Критерии оценки решения практико-ориентированных заданий
(max – 25 баллов за выполнение всех заданий)

Баллы оценки	Критерии оценки
Задание 1-5	Студент получает 5 баллов за полностью решенную задачу

Критерии сформированности компетенции

Оценка в баллах	Оценка	Уровень сформированности компетенции
39-40 баллов	отлично	***Высокий уровень***
30-38 баллов	хорошо	***Продвинутый уровень***
20-29 баллов	удовлетворительно	***Пороговый уровень***
Менее 20 баллов	неудовлетворительно	Компетенции не сформированы

Ключи к тестам
а, б, в, г
1а, 2б, 3в, 4г

№ вопроса	Ответ	№ вопроса	Ответ
1	а)	**16**	а) б) в) г) д)
2	обработки	**17**	а) б)
3	а) б) в)	**18**	а) б)
4	а) в) г)	**19**	а) б)
5	1а) 2б) 3в) 4г) 5д) 6е)	**20**	а) б)
6	г)	**21**	кластеризации
7	Data Mining	**22**	а) б) в)
8	Операционная	**23**	а) б) в) г)
9	а) б) в)	**24**	последовательном
10	1а) 2б) 3в)	**25**	1а) 2б) 3в) 4г)
11	а) б) в) г) д) е)	**26**	1-2-3
12	ошибочных	**27**	б)
13	Нормализация	**28**	ошибки
14	1а) 2б) 3в) 4г) 5д)	**29**	классификации
15	а) б) в) г)	**30**	а) б) в)

Ключи к вопросам

1. Самой большой проблемой больших данных являются затраты на их обработку. Это и дорогостоящее оборудование (его приходится регулярно обновлять для поддержания минимальной работоспособности при увеличении объема данных), и значительные расходы на заработную плату квалифицированным специалистам, обслуживающих огромные массивы информации. Вторая проблема связана с профессионализмом самого аналитика, поскольку ему необходимо обрабатывать большое количество информации. Следующая

проблема - потеря информации. Меры предосторожности требуют не ограничиваться простым однократным резервированием данных. Не менее важной проблемой является проблема конфиденциальности Big Data. При переходе большинства сервисов по обслуживанию клиентов на онлайн-использование данных, очень легко стать мишенью для киберпреступников
2. Data Capture – создание или сбор значений данных, которые еще не существуют и никогда не существовали в компании. Сюда относят: а). Data Acquisition – покупка данных, предложенных внешними компаниями; б). Data Entry – генерация данных ручным вводом, при помощи мобильных устройств или программного обеспечения; в). Signal Reception – получение данных с помощью телеметрии (интернет-вещей).
3. Data Maintenance – передача данных в точки, где происходит синтез данных и их использование в форме, наиболее подходящей для этих целей. Фаза часто включает в себя такие задачи, как перемещение, интеграция, очистка, обогащение, изменение данных, а также процессы экстракции (извлечения)-преобразования-нагрузки; Смешение и интеграция данных нужны, если есть несколько разных источников данных, и нужно анализировать эти данные в комплексе.
4. Data Usage – применение данных как информации для задач, которые должно ставить и выполнять предприятие. В частности, использование данных как элемент управления данными, предполагает решение такой задачи как выяснение того, является ли законным использование данных в том виде, в котором хочет бизнес. Речь идет о так называемом «разрешенном использовании данных», поскольку могут существовать регулирующие или контрактные ограничения на фактическое использование данных, а эти ограничения необходимо соблюдать
5. Data Publication – отправка данных в место за пределами предприятия, например, ежемесячных отчетов клиентам, после чего эти данные де-факто невозможно отозвать. Неверные значения данных не могут быть исправлены, поскольку они уже недоступны для предприятия. Управление данными может потребоваться, чтобы помочь решить, как будут обрабатываться неверные данные, которые были отправлены клиентам.
6. Все задачи, решаемые с помощью машинного обучения, относятся к одной из следующих категорий: 1) регрессия (является частным случаем задач прогнозирования, выполняется с помощью обучения с учителем на этапе тестирования); 2) классификация (выполняется с помощью обучения с учителем на этапе собственно обучения); 3) кластеризация (выполняется с помощью обучения без учителя); 4) понижение размерности данных (выполняется с помощью обучения без учителя); 5) выявление аномалий (выполняется с помощью обучения без учителя).
7. Моделирование является итеративным процессом. Процесс построения большинства моделей состоит из следующих шагов: 1) выбор метода моделирования и переменных для включения в модель; 2) выполнение модели; 3) диагностика и сравнение моделей; 4) выбор лучшей модели на основании ряда критериев (метрик, которые различаются между собой при решении различных аналитических задач).
8. Машинное обучение – класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение в процессе применения решений множества сходных задач. В общем, задача обучения рассматривает набор из n выборок данных, а затем пытается предсказать свойства неизвестных данных.
9. Для розничной торговли среди прикладных задач интеллектуального анализа данных выделяют (современные предприятия розничной торговли собирают сведения о каждой покупке каждого своего покупателя, используя карты клиента, а также компьютеризованные системы контроля и оплаты): 1) анализ портрета покупателя и формирование портрета целевой аудитории (позволяют сделать индивидуализированные предложения отдельным покупателям или их группам, а также определить способы продвижения товаров); 2) анализ покупательской корзины и формирование предложений относительной дальнейших покупок

(позволяет выработать стратегии создания запасов товаров, а также определить способы их раскладки в торговых залах).
10. Для банковской деятельности среди прикладных задач интеллектуального анализа данных выделяют: 1) анализ платежных транзакций и выявление мошенничества с банковскими картами; 2) анализ портрета клиента и их сегментирование (выявление потенциально неблагонадежных заёмщиков, формирование портрета целевой аудитории и т.д.).
11. Для деятельности в области телекоммуникаций среди прикладных задач интеллектуального анализа данных выделяют (предприятия данного сектора экономики собирают информацию о параметрах использования связи, доступа в интернет, геолокации и др.): 1) анализ портрета клиента и формирование портрета целевой аудитории (позволяют сделать индивидуализированные предложения отдельным покупателям или их группам, определить способы продвижения услуг, а также использовать при оказании рекламных услуг); 2) анализ поведения клиента и формирование предложений по повышению уровня его лояльности.
12. Для деятельности в области страхования среди прикладных задач интеллектуального анализа данных выделяют (предприятия данного сектора экономики собирают информацию о социальном и имущественном положении своих клиентов и страховых случаях): 1) анализ портрета клиента и формирование портрета целевой аудитории (позволяют сделать индивидуализированные предложения отдельным потребителям или их группам, определить способы продвижения услуг); 2) анализ страховых событий и выявление мошенничества с возмещением вреда по страховкам; 3) анализ риска и формирование страховых продуктов, отвечающих требованиям доходности.
13. Очистка данных (англ. Data cleansing) — процесс выявления и исправления ошибок, несоответствий данных с целью улучшения их качества, иногда классифицируется как составная часть интеллектуального анализа данных.
14. Выделение признаков — это разновидность абстрагирования, процесс снижения размерности, в котором исходный набор исходных переменных сокращается до более управляемых групп (признаков) для дальнейшей обработки, оставаясь при этом достаточным набором для точного и полного описания исходного набора данных
15. Качество данных – обобщенное понятие, отражающее степень их пригодности к решению определенной задачи. Основными критериями качества данных являются: доступность, точность, взаимосвязанность, полнота, непротиворечивость, однозначность, релевантность, надежность и своевременность.
16. Очистка данных включает обработку пропущенных значений, дубликатов, противоречий, аномальных значений и выбросов, шумов, фиктивных значений и ошибок ввода данных
17. Для обработки противоречий используют следующие методы: исключение противоречивых значений (применяется в том случае, если причина противоречия вызвано ошибкой) и объединение записей с агрегированием числовых значений выходных атрибутов (применяется в том случае, если данные отражают реальные события).
18. Для обработки шумов (обусловленных воздействием случайных факторов флуктуаций значений признаков) используют метод их описания с помощью различных математических моделей в соответствии с их временной, спектральной и пространственной структурой с последующим исключением из основного набора данных.
19. Методы выбора признаков оставляют некоторое подмножество исходного набора признаков, избавляясь от признаков избыточных и слабо информативных. К методам выбора признаков относят: 1) методы фильтров (измеряют релевантность признаков на основе функции μ, и затем решают по правилу κ, какие признаки оставить в результирующем множестве); 2) оберточные методы (находят подмножество искомых признаков последовательно, используя некоторый классификатор в качестве источника оценки качества выбранных признаков); 3) встроенные методы (для выбора признаков используется непосредственно структуру некоторого классификатора) и другие.

20. Методы выделения признаков составляют из уже исходных признаков новые, все также полностью описывающие пространство набора данных, но уменьшая его размерность и теряя в репрезентативности данных, т.к. становится непонятно, за что отвечают новые признаки. К методам выделения признаков относят: 1) метод главных компонент; 2) метод разложения по усеченным сингулярным значениям; 3) метод неотрицательного матричного разложения; 4) метода линейного дискриминантного анализа.
21. Способы машинного обучения делятся на несколько категорий: 1) обучение с учителем (обучение осуществляется принудительно с помощью примеров «стимул-реакция»); 2) обучение без учителя (обучение осуществляется спонтанно без вмешательства извне); 3) обучение с подкреплением (обучение осуществляется при взаимодействии с некоторой средой).
22. Все задачи, решаемые с помощью машинного обучения, относятся к одной из следующих категорий: 1) регрессия (является частным случаем задач прогнозирования, выполняется с помощью обучения с учителем на этапе тестирования); 2) классификация (выполняется с помощью обучения с учителем на этапе собственно обучения); 3) кластеризация (выполняется с помощью обучения без учителя); 4) понижение размерности данных (выполняется с помощью обучения без учителя); 5) выявление аномалий (выполняется с помощью обучения без учителя).
23. Метод стохастического градиентного спуска – методом оптимизации в машинном обучении. Стохастический градиентный спуск – это метод итерации для оптимизации целевой функции с подходящими свойствами гладкости (например, дифференцируемости или субдифференцируемости). В данном методе градиент оптимизируемой функции считается на каждом шаге как градиент от одного, случайно выбранного элемента.
24. Как правило, классификацию можно разбить на две области: а) бинарная классификация, в которой требуется сгруппировать результат в одну из двух групп (0 или 1, истинное или ложное, положительное или отрицательное); б) мультиклассовая классификация, где требуется сгруппировать результат в одну из нескольких (более двух) групп.
25. Свойства линейной регрессии: 1) легкость моделирования; 2) высокая полезность при создании не очень сложной зависимости, а также при небольшом количестве данных; 3) интуитивно-понятна; 4) чувствительность к выбросам.
26. Различают два класса нелинейных регрессий: а) регрессии, нелинейные относительно включенных в анализ объясняющих переменных, но линейные по оцениваемым параметрам (например, полиномы различных степеней); б) регрессии, нелинейные по оцениваемым параметрам (например, логарифмическая, показательная, степенная функции).
27. Как правило, классификацию можно разбить на две области: а) бинарная классификация, в которой требуется сгруппировать результат в одну из двух групп (0 или 1, истинное или ложное, положительное или отрицательное); б) мультиклассовая классификация, где требуется сгруппировать результат в одну из нескольких (более двух) групп.
28. Логистическая регрессия – это фундаментальный метод классификации, который применяется в основном для решения бинарных задач. Он принадлежит к группе линейных классификаторов и чем-то похож на полиномиальную и линейную регрессию. Логистическая регрессия выполняется быстро, она относительно не сложна и удобна для интерпретации результатов. Логистическая регрессия требует довольно больших размеров выборки. Логистическая регрессия имеет следующие допущения: 1) для двоичной логистической регрессии требуется, чтобы зависимая переменная была двоичной; 2) для бинарной регрессии значение зависимой переменной равное 1 должно представлять желаемый результат; 3) в модель следует включать только значимые переменные; 4) входные переменные должны быть независимыми друг от друга
29. Кластерный анализ – многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы. Задача кластеризации относится к классу задач машинного обучения без учителя. Целями кластеризации в зависимости от решаемой задачи являются: 1)

расширение знаний о предметной области; 2) сжатие данных (замена множества данных, входящим в кластер, одним представителем); 3) обнаружение новизны (данных, которые невозможно отнести ни к одному из кластеров).
30. Итеративная кластеризация предполагает объединение объектов в кластеры на основе их «близкого» (по некоторой метрике, например на основании евклидова расстояния, манхэттенского расстояния, расстояния Чебышева и т.д.) расположения относительно образовавшегося центра. Итеративная кластеризация разделяет объекты данных на неперекрывающиеся группы. Другими словами, ни один объект не может быть членом более чем одного кластера, и каждый кластер должен иметь хотя бы один объект.

Примечание.

В соответствии с нормативно-правовыми актами для лиц с ограниченными возможностями здоровья при необходимости тестирование может быть проведено только в письменной или устной форме, а также могут быть использованы другие материалы контроля качества знаний, предусмотренные рабочей программой дисциплины.

Разработчик Виноградов Д.В., ст. преподаватель кафедры «Бизнес–информатика и экономика»

Фонд оценочных материалов (средств) рассмотрен и одобрен на заседании кафедры «Бизнес – информатика и экономика»
протокол № 1 от «30» августа 2023 года.
Заведующий кафедрой: д.э.н., профессор Тесленко И.Б.

Фонд оценочных материалов (средств) рассмотрен и одобрен на заседании учебно-методической комиссии направления 01.03.05 Статистика
протокол № 1 от «05» сентября 2023 года.
Председатель комиссии: к.э.н., доцент Яресь О.Б.