МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФГБОУ ВО «Уральский государственный экономический университет»

Одобрена
на заседании кафедры информационных технологий и статистики

10 января 2020 г.

протокол № 6

Зав. кафедрой _____ Сурнина Н.М.
(подпись)

Утверждена
Советом по учебно-методическим вопросам и качеству образования

15 января 2020 г.

протокол № 5

Председатель _____ Карх Д.А.
(подпись)

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Наименование дисциплины	Инструменты обработки и анализа корпоративных данных
Направление подготовки	09.04.03 Прикладная информатика
Профиль	Бизнес-модели и цифровые решения
Форма обучения	очная
Год набора	2020

Разработана:
Доцент, к.э.н.

_____ Кислицын Евгений Витальевич
(подпись)

Екатеринбург
2020 г.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Рабочая программа дисциплины является частью основной профессиональной образовательной программы высшего образования - программы магистратуры, разработанной в соответствии с ФГОС ВО

ФГОС ВО	Федеральный государственный образовательный стандарт высшего образования по направлению подготовки 09.04.03 Прикладная информатика (уровень магистратуры) (приказ Минобрнауки России от 19.09.2017г. №916)
ПС	

1. ЦЕЛЬ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Формирование знаний, умений и навыков в области обработки больших данных в корпорациях, применения статистических, математических и эмпирических методов анализа экономической информации и извлечения знаний из больших массивов данных.

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОПОП

Дисциплина относится к вариативной части учебного плана.

3. ОБЪЕМ ДИСЦИПЛИНЫ

Промежуточный контроль	Часов					З.е.
	Всего за семестр	Контактная работа .(по уч.зан.)			Самостоятельная работа в том числе подготовка контрольных и курсовых	
		Всего	Лекции	Лабораторные		
Семестр 3						
Экзамен	180	32	4	28	112	5

4.ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОСВОЕНИЯ ОПОП

В результате освоения ОПОП у выпускника должны быть сформированы компетенции, установленные в соответствии ФГОС ВО.

Общепрофессиональные компетенции (ОПК)

Шифр и наименование компетенции	Индикаторы достижения компетенций
ОПК-5 Способен разрабатывать и модернизировать программное и аппаратное обеспечение информационных и автоматизированных систем;	ОПК-5.1 Знать: Знать современное программное и аппаратное обеспечение информационных и автоматизированных систем; Уметь: модернизировать программное и аппаратное обеспечение информационных и автоматизированных систем для решения профессиональных задач;

ОПК-3 Способен анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями;	ОПК-3.1 Знать: принципы, методы и средства анализа и структурирования профессиональной информации; Уметь: анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров;

Профессиональные компетенции (ПК)

Шифр и наименование компетенции	Индикаторы достижения компетенций
проектный	
ПК-3 Способен проводить анализ корпоративных и отраслевых данных с использованием современных алгоритмов и инструментальных средств	ПК-3.1 Знать: основы статистики, основы теории отраслевых рынков, многомерные статистические методы, принципы корреляционного, регрессионного, факторного и кластерного анализа, теорию нейронных сетей. Уметь: проводить анализ отраслевых рынков и предприятий с использованием статистических алгоритмов и методов машинного обучения. Иметь навыки: обработки и анализа корпоративных данных, работы с инструментальными средствами анализа данных.
ПК-1 Способен проектировать и разрабатывать цифровые решения в области экономики и управления	ПК-1.1 Знать: основы институциональной экономики, технологии и методологии проектирования информационных систем, основы искусственного интеллекта, имитационного моделирования. Уметь: создавать цифровые решения с использованием технологий искусственного интеллекта, имитационные модели, программные средства анализа данных и управления процессами, системы поддержки принятия решений. Иметь навыки: работы со средой имитационного моделирования, с информационно-аналитическими системами, автоматизации прикладных задач с использованием технологий искусственного интеллекта, имитационного моделирования, информационно-аналитических систем.

5. ТЕМАТИЧЕСКИЙ ПЛАН

Тема	Часов						
	Наименование темы	Всего часов	Контактная работа .(по уч.зан.)			Самост. работа	Контроль самостоятельной работы
			Лекции	Лабораторные	Практические занятия		
Семестр 3		144					
Тема 1.	Статистические модели: критерии и методы их оценивания. Инструменты построения статистических моделей	24	2	4		18	
Тема 2.	Регрессионные модели	23	1	4		18	
Тема 3.	Модели классификации	23	1	4		18	
Тема 4.	Ансамблевые модели	22		4		18	
Тема 5.	Методы кластеризации и понижения размерности	28		8		20	

Тема 6.	Нейронные сети	24		4		20	

6. ФОРМЫ ТЕКУЩЕГО КОНТРОЛЯ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ШКАЛЫ ОЦЕНИВАНИЯ

Раздел/Тема	Вид оценочного средства	Описание оценочного средства	Критерии оценивания
Текущий контроль (Приложение 4)			
Темы 1-3	Тест (приложение 4)	Тест состоит из 50-ти вопросов	10 баллов
Темы 4-5	Тест (приложение 4)	Тест состоит из 16-ти вопросов	10 баллов
Темы 6-7	Реферат (приложение 4)	Объем реферата от 15 до 25 страниц	10 баллов
Промежуточный контроль (Приложение 5)			
3 семестр (Эк)	Экзаменационный билет (приложение 5)	Билет состоит из 2 теоретических вопросов и 1 практического задания	100 баллов

ОПИСАНИЕ ШКАЛ ОЦЕНИВАНИЯ

Показатель оценки освоения ОПОП формируется на основе объединения текущей и промежуточной аттестации обучающегося.

Показатель рейтинга по каждой дисциплине выражается в процентах, который показывает уровень подготовки студента.

Текущая аттестация. Используется 100-балльная система оценивания. Оценка работы студента в течении семестра осуществляется преподавателем в соответствии с разработанной им системой оценки учебных достижений в процессе обучения по данной дисциплине.

В рабочих программах дисциплин и практик закреплены виды текущей аттестации, планируемые результаты контрольных мероприятий и критерии оценки учебный достижений.

В течение семестра преподавателем проводится не менее 3-х контрольных мероприятий, по оценке деятельности студента. Если посещения занятий по дисциплине включены в рейтинг, то данный показатель составляет не более 20% от максимального количества баллов по дисциплине.

Промежуточная аттестация. Используется 5-балльная система оценивания. Оценка работы студента по окончанию дисциплины (части дисциплины) осуществляется преподавателем в соответствии с разработанной им системой оценки достижений студента в процессе обучения по данной дисциплине. Промежуточная аттестация также проводится по окончанию формирования компетенций.

Порядок перевода рейтинга, предусмотренных системой оценивания, по дисциплине, в пятибалльную систему.

Высокий уровень – 100% - 70% - отлично, хорошо.

Средний уровень – 69% - 50% - удовлетворительно.

Показатель оценки	По 5-балльной системе	Характеристика показателя
100% - 85%	отлично	обладают теоретическими знаниями в полном объеме, понимают, самостоятельно умеют применять, исследовать, идентифицировать, анализировать, систематизировать, распределять по категориям, рассчитать показатели, классифицировать, разрабатывать модели, алгоритмизировать, управлять, организовать, планировать процессы исследования, осуществлять оценку результатов на высоком уровне
84% - 70%	хорошо	обладают теоретическими знаниями в полном объеме, понимают, самостоятельно умеют применять, исследовать, идентифицировать, анализировать, систематизировать, распределять по категориям, рассчитать показатели, классифицировать, разрабатывать модели, алгоритмизировать, управлять, организовать, планировать процессы исследования, осуществлять оценку результатов. Могут быть допущены недочеты, исправленные студентом самостоятельно в процессе работы (ответа и т.д.)
69% - 50%	удовлетворительно	обладают общими теоретическими знаниями, умеют применять, исследовать, идентифицировать, анализировать, систематизировать, распределять по категориям, рассчитать показатели, классифицировать, разрабатывать модели, алгоритмизировать, управлять, организовать, планировать процессы исследования, осуществлять оценку результатов на среднем уровне. Допускаются ошибки, которые студент затрудняется исправить самостоятельно.
49 % и менее	неудовлетворительно	обладают не полным объемом общих теоретическими знаниями, не умеют самостоятельно применять, исследовать, идентифицировать, анализировать, систематизировать, распределять по категориям, рассчитать показатели, классифицировать, разрабатывать модели, алгоритмизировать, управлять, организовать, планировать процессы исследования, осуществлять оценку результатов. Не сформированы умения и навыки для решения
100% - 50%	зачтено	характеристика показателя соответствует «отлично», «хорошо», «удовлетворительно»
49 % и менее	не зачтено	характеристика показателя соответствует «неудовлетворительно»

7. СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

7.1. Содержание лекций

Тема 1. Статистические модели: критерии и методы их оценивания. Инструменты построения статистических моделей Задача машинного обучения. Интеллектуальный анализ данных. Обучение с учителем и без учителя. Вероятностная постановка задачи. Регрессионная функция. Байесов классификатор. Метод главных компонент. Проверка статистических гипотез.
Тема 2. Регрессионные модели Регрессия. Проверка значимости и доверительные интервалы для коэффициентов. Подготовка данных. Переобучение.
Тема 3. Модели классификации Наивный байесовский классификатор. Дискриминантный анализ. МНК для задачи классификации. Логистическая регрессия.

7.2 Содержание практических занятий и лабораторных работ

Тема 1. Статистические модели: критерии и методы их оценивания. Инструменты построения статистических моделей Визуализация данных. Статистический анализ.
Тема 2. Регрессионные модели Линейная регрессия. Функционал качества и градиентный спуск.
Тема 3. Модели классификации Классификация в бинарных пространствах с использованием классических моделей. Бинарные деревья решений. Поиск логических закономерностей в данных. Алгоритмы выделения ассоциативных правил. Анализ последовательностей знаков или событий. Дискриминантный анализ. Метод опорных векторов. Ядерные функции машины опорных векторов. Деревья классификации, случайный лес и логистическая регрессия. Процедуры сравнения эффективности моделей классификации. Ирисы Фишера и метод k-ближайших соседей. Наивный классификатор Байеса. Классификация в линейном дискриминантном пространстве. Нелинейные классификаторы в R. Модель мультиномиального логита.
Тема 4. Ансамблевые модели Решающие деревья. Случайный лес. Градиентный бустинг. Применение ансамблевых моделей.
Тема 5. Методы кластеризации и понижения размерности Задача кластеризации. Группы методов. Метод k-средних. Иерархическая кластеризация. Агломеративный алгоритм. DBSCAN. Оценки качества кластеризации. Метод главных компонент. Сингулярное разложение матрицы и связь с PCA. Применение PCA на данных. Многомерное шкалирование. t-SNE. Рекомендательные системы. Методы коллаборативной фильтрации. Методы с матричными разложениями.
Тема 6. Нейронные сети Обучение нейросети. Сверточные сети. Рекуррентные сети. Современные архитектуры. Введение в TensorFlow. Классификация изображений на Tensorflow.

7.3. Содержание самостоятельной работы

Тема 1. Статистические модели: критерии и методы их оценивания. Инструменты построения статистических моделей Изучение основной и дополнительной литературы, интернет-источников по теме. Разбор практических примеров. Изучение функций языка R. Выполнение практических работ.
Тема 2. Регрессионные модели Изучение основной и дополнительной литературы, интернет-источников по теме. Разбор практических примеров. Изучение функций языка R. Выполнение практических работ.

Тема 3. Модели классификации Изучение основной и дополнительной литературы, интернет-источников по теме. Разбор практических примеров. Изучение функций языка R. Выполнение практических работ.
Тема 4. Ансамблевые модели Изучение основной и дополнительной литературы, интернет-источников по теме. Разбор практических примеров. Изучение функций языка R. Выполнение практических работ.
Тема 5. Методы кластеризации и понижения размерности Изучение основной и дополнительной литературы, интернет-источников по теме. Разбор практических примеров. Изучение функций языка R. Выполнение практических работ.
Тема 6. Нейронные сети Изучение основной и дополнительной литературы, интернет-источников по теме. Разбор практических примеров. Изучение функций языка R. Выполнение практических работ.

7.3.1. Примерные вопросы для самостоятельной подготовки к зачету/экзамену
Приложение 1.

7.3.2. Практические задания по дисциплине для самостоятельной подготовки к зачету/экзамену
Приложение 2.

7.3.3. Перечень курсовых работ
Не предусмотрено.

7.4. Электронное портфолио обучающегося
Материалы не размещаются.

7.5. Методические рекомендации по выполнению контрольной работы
Не предусмотрено.

7.6 Методические рекомендации по выполнению курсовой работы
Не предусмотрено.

8. ОСОБЕННОСТИ ОРГАНИЗАЦИИ ОБРАЗОВАТЕЛЬНОГО ПРОЦЕССА ПО ДИСЦИПЛИНЕ ДЛЯ ЛИЦ С ОГРАНИЧЕННЫМИ ВОЗМОЖНОСТЯМИ ЗДОРОВЬЯ

По заявлению студента

В целях доступности освоения программы для лиц с ограниченными возможностями здоровья при необходимости кафедра обеспечивает следующие условия:

- особый порядок освоения дисциплины, с учетом состояния их здоровья;
- электронные образовательные ресурсы по дисциплине в формах, адаптированных к ограничениям их здоровья;
- изучение дисциплины по индивидуальному учебному плану (вне зависимости от формы обучения);
- электронное обучение и дистанционные образовательные технологии, которые предусматривают возможности приема-передачи информации в доступных для них формах.
- доступ (удаленный доступ), к современным профессиональным базам данных и информационным справочным системам, состав которых определен РПД.

9. ПЕРЕЧЕНЬ ОСНОВНОЙ И ДОПОЛНИТЕЛЬНОЙ УЧЕБНОЙ ЛИТЕРАТУРЫ, НЕОБХОДИМОЙ ДЛЯ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Сайт библиотеки УрГЭУ
http://lib.usue.ru/

Основная литература:

1. Кулаичев А. П.. Методы и средства комплексного статистического анализа данных:учебное пособия для вузов по дисциплинам «Математическая статистика» и «Информатика». - Москва: ИНФРА-М, 2018. - 484 с.

Дополнительная литература:

1. Ниворожкина Л. И., Арженовский С. В., Рудяга А. А., Торопова Н. А., Федосова О. Н., Житников И. В., Трегубова А. А., Федотова Э. А., Ниворожкина Л. И.. Статистические методы анализа данных:учебник. - Москва: РИОР: ИНФРА-М, 2016. - 333 с.

2. Ковалев В. В., Дюкина Т. О., Зуга Е. И., Колычева В. А., Попова И. Н., Смирнова Н. А., Третьяков С. Л., Шаныгин С. И., Подкорытова О. А.. Теория статистики с элементами эконометрики в 2 ч. Часть 2 [Электронный ресурс]:Учебник. - Москва: Издательство Юрайт, 2019. - 348 – Режим доступа: https://www.biblio-online.ru/bcode/434520

10. ПЕРЕЧЕНЬ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, ВКЛЮЧАЯ ПЕРЕЧЕНЬ ЛИЦЕНЗИОННОГО ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ И ИНФОРМАЦИОННЫХ СПРАВОЧНЫХ СИСТЕМ, ОНЛАЙН КУРСОВ, ИСПОЛЬЗУЕМЫХ ПРИ ОСУЩЕСТВЛЕНИИ ОБРАЗОВАТЕЛЬНОГО ПРОЦЕССА ПО ДИСЦИПЛИНЕ

Перечень лицензионное программное обеспечение:

Astra Linux Common Edition. Договр № 1 от 13 июня 2018, акт от 17 декабря 2018. Срок действия лицензии - без органичения срока.

Libre Office. Лицензия GNU LGPL. Срок действия лицензии - без органичения срока.

Язык программирования R.Лицензия GNU GPL 2.Срок действия лицензии - без органичения срока.

R Studio (среда для языка программирования R).Лицензия GNU Affero General Public License v3.Срок действия лицензии - без органичения срока.

Перечень информационных справочных систем, ресурсов информационно-телекоммуникационной сети «Интернет»:

Статистические методы в управлении инновациями
https://openedu.ru/course/ITMOUniversity/INMAN/
R для лингвистов: программирование и анализ данных
https://openedu.ru/course/hse/RLING/

11. ОПИСАНИЕ МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЙ БАЗЫ, НЕОБХОДИМОЙ ДЛЯ ОСУЩЕСТВЛЕНИЯ ОБРАЗОВАТЕЛЬНОГО ПРОЦЕССА ПО ДИСЦИПЛИНЕ

Реализация учебной дисциплины осуществляется с использованием материально-технической базы УрГЭУ, обеспечивающей проведение всех видов учебных занятий и научно-исследовательской и самостоятельной работы обучающихся:

Специальные помещения представляют собой учебные аудитории для проведения всех видов занятий, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации.

Помещения для самостоятельной работы обучающихся оснащены компьютерной техникой с возможностью подключения к сети "Интернет" и обеспечением доступа в электронную информационно-образовательную среду УрГЭУ.

Все помещения укомплектованы специализированной мебелью и оснащены мультимедийным оборудованием спецоборудованием (информационно-телекоммуникационным, иным компьютерным), доступом к информационно-поисковым, справочно-правовым системам, электронным библиотечным системам, базам данных действующего законодательства, иным информационным ресурсам служащими для представления учебной информации большой аудитории.

Для проведения занятий лекционного типа презентации и другие учебно-наглядные пособия, обеспечивающие тематические иллюстрации

Приложение №1

к рабочей программе дисциплины

7.3.1. Примерные вопросы для самостоятельной подготовки к экзамену

1. Вероятностная постановка задачи обучения с учителем. Функция потерь. Средний (ожидаемый) риск. Эмпирический риск. Регрессионная функция и байесов классификатор. Неустранимая (байесовская) ошибка.
2. Ошибки 1-го и 2-го рода. Чувствительность, специфичность, точность, полнота. ROC-кривая. Площадь под ROC-кривой.
3. Принцип минимизации эмпирического риска. Минимизация отступа. Регуляризация.
4. Экспериментальная (эмпирическая) оценка качества обучения. Обучающая, проверочная и тестовая выборки. Метод перекрестного (скользящего) контроля.
5. Метод k ближайших соседей в задачах классификации и восстановления регрессии. Теорема об оценке риска в методе ближайшего соседа. Идея доказательства.
6. Метод наименьших квадратов. Система нормальных уравнений. Псевдорешение.
7. Борьба с переобучением в методе наименьших квадратов. Сокращение числа параметров. Полный перебор всех подмножеств признаков. Жадный (Forward stepwise) алгоритм.
8. Борьба с переобучением в методе наименьших квадратов. Ридж-регрессия (регуляризация). Метод "Лассо".
9. Наивный байесовский классификатор. Сглаживание Лапласа. Использование наивного байесовского классификатора для количественых признаков.
10. Линейный дискриминантный анализ. Квадратичный дискриминантный анализ.
11. Логистическая регрессия. Логистическая функция и softmax
12. Нейронные сети. Алгоритм обучения Back-Propagation.
13. Понятие о глубоком обучении. Сверточные нейронные сети.
14. Машина опорных векторов. Формулировка задачи в виде задачи математического программирования. Двойственная задача. (Случай линейно разделимых и неразделимых классов)
15. Машина опорных векторов. Ядра и спрямляющие пространства.
16. Деревья решений. Алгоритм CART.
17. Случайный лес. Экстремально случайные деревья.
18. Алгоритм XGBoost.
19. Градиентный бустинг деревьев решений.
20. Дилемма "Смещение-разброс". Кривая обучения.
21. Задача понижения размерности. Метод главных компонент.
22. Задача кластеризации. Метод центров тяжести. Метод медоидов.
23. Алгоритм "Ожидание-максимизация".
24. Алгоритм DBSCAN.

25.Алгоритмы иерархической кластеризации.

Приложение 2
к рабочей программе

7.3.2. Практические задания по дисциплине для самостоятельной подготовки к зачету/экзамену

Примерные практические задания к экзамену

1. Пусть g: $R^n \to R^m$, x ∈ Rn. Матрицей Якоби называется матрица

$$\frac{\partial g}{\partial x} = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} & \cdots & \frac{\partial g_2}{\partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial g_m}{\partial x_1} & \frac{\partial g_m}{\partial x_2} & \cdots & \frac{\partial g_m}{\partial x_n} \end{pmatrix}.$$

В частности, если m = 1 (т. е. g(x) — скалярная функция векторного аргумент x), то ∂g/∂x — градиент функции g.
Доказать, что
1) если a ∈ R^n , x ∈ R^n, то ∂(aᵀx)/∂x = a;
2) если A ∈ $R^{m\times n}$, x ∈ R^n , то ∂(Ax)/∂x = A;
3) если A ∈ $R^{n\times n}$, x ∈ R^n , òî ∂(xᵀAx)/∂x = (A + Aᵀ)x; в частности, если Aᵀ = A, то ∂(xᵀAx)/∂x = 2Ax;
4) если x ∈ Rn , то ∂|x|²/∂x = 2x;
5) если g — скалярная функция и под g(x) понимается применение функции g к каждой компоненте вектора x ∈ R^n, то
∂g(x)/∂x = diag(g′(x)),
где diag(a) — диагональная матрица с диагональю a;
6) если h : $R^n \to R^m$, g : $R^m \to R^p$, x ∈ R^n , то
∂g(h(x))/∂x = (∂g(h(x))/∂h)(∂h(x)/∂x).
2. Пользуясь №1, найдите градиент ∂g(β)/∂β и гессиан (∂²g(β))/(∂βᵀ∂β)
функции g(β) = |Xβ − y|². Выведите отсюда, что решение линейной задачи наименьших квадратов β^ = argmin |Xβ − y|² является решением нормальной системы линейных уравнений XᵀXβ = Xᵀy.
3. Дана обучающая выборка

x	1	1	0	0	-1
y	4	4	0	2	6

1) изобразить точки;
2) методом наименьших квадратов построить модель вида f(x) = $\beta_0 + \beta_1 x + \beta_2 x^2$; построить график этой функции;
3) построить модель того же вида методом ридж-регрессии с параметром регуляции λ = 1; построить график этой функции.
Замечание: при ручных вычислениях по методу наименьших квадратов квадратов рекомендуется составить систему XᵀXβ = Xᵀy и решить ее. Регулязированная система (XᵀX + λI)β = Xᵀy, где I — единичная матрица.
4. Рассмотрим задачу восстановления регрессии, в которой y распределен согласно нормальному закону N(Xβ, σ^2 I), а β имеет априорное распределение N(0, τI). Найти апостериорное распределение для β. Доказать, что β^{ridge} есть его математическое ожидание. Найти связь между параметром регуляризации λ и дисперсиями τ , σ^2.
5. Показать, что процедура гребневой регрессии эквивалентна обычному методу наименьших квадратов, примененному к расширенным данным: к центрированной матрице X дописывается матрица (√ λ)I, к вектору y приписывается d нулей.
6. Показать, как (и объяснить почему) задачу квадратичного программирования в методе лассо

$$\widehat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{d} x_{ij}\beta_j \right)^2 \right\},$$

при условии

$$\sum_{j=1}^{d} |\beta_j| \le s$$

можно свести к задаче квадратичного программирования с 2d+1 неизвестными и 2d+1 линейными ограничениями.

7. Метод использования линейной регрессии в задаче классификации заключается в следующем. Сопоставим каждому классу k вектор $(y_1,y_2,...y_K)$, в котором $y_k = 1$? а $y_i = 0$ при $i \neq k$. Собрав вместе индикаторные векторы объектов обучающей выборки, получим матрицу Y размера $N \times K$. Пусть X — матрица размера $N \times (d + 1)$, первый столбец которой состоит из единиц, а последующие представляют собой векторы из обучающей выборки. Применяя метод наименьших квадратов одновременно к каждому столбцу матрицы Y, получаем значения

$\hat{Y} = X(X^\top X)^{-1}X^\top Y.$

Для каждого столбца y_k матрицы Y получим свой столбец коэффициентов $\hat{\beta}_k$. Соберем их в матрицу $\hat{B}$ размера $(d + 1) \times K$. Имеем

$\hat{B} = (X^\top X)^{-1}X^\top Y.$

Объект x будем классифицировать согласно следующему правилу: Вычислим вектор-строку длину K

$g(x) = (1, x)\,\hat{B}.$

Отнесем x к классу

$f(x) = \arg\max_k g_k(x)$

Доказать, что

$$\sum_{k=1}^{K} g_k(x) = 1.$$

Доказать, что в случае K = 2 данный метод эквиваленте решению одной задачи восстановления регрессии. Какой?

8. Дана обучающая выборка

x_1	0	1	0	2	2	2	4	3
x_2	-1	0	0	0	1	0	1	2
y	0	0	0	0	0	1	1	1

1) Методом линейного дискриминантного анализа для каждого класса построить дискриминатную функцию и записать уравнение разделяющей поверхности.

2) Методом квадратичного дискриминантного анализа построить дискриминантные функции. Изобразить точки и разделяющие поверхности (кривые)

9. Задача Фишера сводится к максимизации отношения Рэлея

$\max_a \; (a^\top Ba)/(a^\top Wa).$

Показать, как эта задача сводится к обощенной задаче на собственные значения

$Ba = \lambda W.$

10. Показать, что оптимальная гиперплоскость, разделяющая два множества, является плоскостью, проходящей через середину отрезка, соединяющего пару ближайших точек из выпуклой оболочки каждого из классов, и перпендикулярно ему. Указание: рассмотреть задачу, двойственную к задаче определения оптимальной гиперплоскости.

11. Показать, что в алгоритме SVM задача

$$\min_{\beta,\beta_0,\xi_i} \frac{1}{2}\|\beta\|^2 + \gamma\sum_{i=1}^{N}\xi_i,$$

при ограничениях

$y_i(x_i^\top\beta + \beta_0) \ge 1 - \xi_i\,,\ \xi_i \ge 0\ (i = 1, 2, \ldots, N),$

эквивалентна задаче

$$\min_{\beta,\beta_0} \sum_{i=1}^{N}\left[1 - y_i\left(x_i^\top\beta + \beta_0\right)\right]_+ + \alpha\|\beta\|^2,$$

где $[\cdot]_+$ означает положительную часть, и $\alpha = 1/(2\gamma)$.

12. SVM и задача восстановления регрессии. Для восстановления β , β_0 в модели $f(x) = x^T\beta + \beta_0$. Рассмотрим задачу минимизации функции

$$H(\beta, \beta_0) = \sum_{i=1}^{N} V\left(y_i - f(x_i)\right) + \frac{\alpha}{2}\|\beta\|^2,$$

где

$$V(t) = V_\varepsilon(t) = \begin{cases} 0, & \text{если } |t| < \varepsilon, \\ |t| - \varepsilon & \text{в противном случае.} \end{cases}$$

Доказать, что решение $\beta^$, $\beta^_0$, минимизирующее функцию þ $H(\beta, \beta_0)$, можно представить в виде

$$\widehat{\beta} = \sum_{i=1}^{N} (\widehat{\alpha}_i^* - \widehat{\alpha}_i) x_i, \quad \widehat{f}(x) = \sum_{i=1}^{N} (\widehat{\alpha}_i^* - \widehat{\alpha}_i) \langle x, x_i \rangle + \beta_0,$$

где $a^_i$ и $a^_i*$ являются решением следующей задачи квадратичного программирования:

$$\min_{\alpha_i, \alpha_i^*} \left(\varepsilon \sum_{i=1}^{N} (\alpha_i^* + \alpha_i) - \sum_{i=1}^{N} y_i (\alpha_i^* - \alpha_i) + \frac{1}{2} \sum_{i,j=1}^{N} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) \langle x_i, x_j \rangle \right)$$

при ограничениях

$$0 \le \alpha_i, \alpha_i^* \le \frac{1}{\lambda}, \quad \sum_{i=1}^{N} (\alpha_i^* - \alpha_i) = 0, \quad \alpha_i \alpha_i^* = 0.$$

13. Дана обучающая выборка

x_1	-1	-1	1	1
x_2	-1	1	-1	1
y	1	-1	-1	1

Подобрать ядро и указать соответствующий SVM-классификатор, для которого ошибка на обучающей выборке равна 0.

14. Дана обучающая выборка

x_1	0	0	1	1	0	0	1	1	1	0
x_2	0	1	0	1	1	1	1	1	1	1
y	0	0	0	0	0	1	1	1	1	1

С помощью наивного байесова классификатора оценить вероятность $Pr(Y = 0|X_1 = 1, X_2 = 1)$; $Pr(Y = 1|X_1 = 1, X_2 = 1)$.

15. Дана выборка

x_1	4	0	-2	2
x_2	3	1	-3	-1

Найти главные направления и дисперсии по главным компонентам.
Изобразить точки и главные направления.

16. Дана выборка

x_1	4	0	-1	3	4
x_2	2	-3	-2	1	2
x_3	3	2	2	1	-3

Найти главные направления и дисперсии по главным компонентам.

17. Пусть $\sigma(z) = 1/(1 + e^{-z})$ (сигмоидальная функция). Проверить, что $\sigma' = \sigma(1 - \sigma)$.

18. Пусть в задаче классификации на K классов {1,2,…,K} последний слой нейронной сети вычисляет softmax-функцию:

$$g_k(s_1, s_2, \ldots, s_K) = e^{s_k} \Big/ \sum_{\ell=1}^{K} e^{s_\ell}.$$

В качестве штрафа используется кросс-энтропия (logross-функция):

$$R^{(i)} = -\sum_{k=1}^{K} I(y^{(i)} = k) \ln g_k(s_1, s_2, \ldots, s_K),$$

где $g_k(s_1,s_2,\ldots,s_K)$ — softmax-функция. Доказать, что
1) $\partial g_k/\partial s_\ell = g_k \cdot I(k = \ell) - g_\ell$;
2) $\partial R^{(i)}/\partial gk = - I(y^{(i)})/g_k$;
3) $\partial R^{(i)}\ \partial s_\ell = g_\ell - I(\ell = y^{(i)})$.

19. Предположим, что рассматривается K задач двуклассовой классификации, в каждой из которых по $x \in X$ требуется предсказать $y_K \in \{0,1\}$ (k = 1, 2, …, K) (например, требуется определить, присутствует или отсутствует на изображении x объект k). Обучающая выборка составлена из пар $(x^{(i)},y^{(i)})$, где $y^{(i)} = (y^{(i)}_1, y^{(i)}_2, \ldots, y^{(i)}_K) \in \{0, 1\}^K$ (i=1, 2, …, N). Для решения такой задачи можно использовать нейронную сеть, последний слой которой вычисляет K сигмоидальных функций $g_k(s_k) = 1/(1 + e^{-sk})$ (k = 1, 2, . . . , K).
Пусть в качестве штрафа используется

$$R^{(i)} = -\sum_{k=1}^{K} \left(y_k^{(i)} \ln g_k + (1 - y_k^{(i)}) \ln(1 - g_k) \right).$$

Доказать, что
$\partial R^{(i)}/\partial sk = g_k - y^{(i)}{}_k$.

20. Нейронная сеть с двумя нелинейными слоями вычмсляет функцию saftmax(B(σ(Ax))). Пусть в качестве функции потерь используется logloss. Таким образом, на объекте $x^{(i)}$ потери равны logloss(softmax(B(σ(Ax)))). Пользуясь результатом задачи №1, выпишите матричные формулы для алгоритма backpropagation.

21. Пусть рассматривается задача 2-классовой классификации. Доказать, что если известно сколько в выборке представителей каждого из двух классов, то по любым двум показателям из списка TPR, TNR, PPV, NPV определяются остальные два.

22. Пусть рассматривается задача 2-х классовой классификации. Верно ли, что
1) если у двух классификаторов на одной и той же выборке совпадают PPV (Precision) и совпадают TPR (Recall или Sensitivity), то будут совпадать TNR (Specificity) и NPV;
2) если у двух классификаторов на одной и той же выборке совпадают TNR (Specificity) и совпадают NPV, то будут совпадать PPV(Precision) и TPR (Recall или Sensitivity);
3) совпадение ROC кривых(для двух классификаторов на одной и той же выборке) влечет совпадение Precision-Recall кривых и наоборот?

23. Рассмотрим задачу восстановления регрессии с квадратичной функцией потерь $L(y',y) == (y' - y)^2$ Доказать, что если $f^*(x) = E(Y|X = x)$ (регрессионная функция). Чему тогда равен средний риск $R(f^*)$?

24. Рассмотрим задачу восстановления регрессии с функцией потерь $L(y', y) = |y' - y|$. Доказать, что минимум среднему риску доставляет при этом условная медиана $f(x) = \text{median}(Y|X = x)$.

25. Пусть в задаче классификации с двумя классами {0,1} используется функция потерь L(y ′ , y), такая, что $L(0,0) = L(1,1)=0$, $L(1,0) = l_1$, $L(0,1) = l_0$. Докажите, что в этом случае байесов классификатор $f^*(x)$ удовлетворяет условию

$$f(x) = \underset{y \in \{0,1\}}{\operatorname{argmax}} \ell_y \Pr(y|x).$$

26. Выразить байесов классификатор $f^*(x)$ для классификации с K классами, если функция потерь равна L(y ′ , y) = ℓy′y (y ′ , y = 1, 2, . . . , K).

27. Пусть N точек распределены случайно равномерно в единичной d-мерной гиперсфере. Доказать, что медианное расстояние от центра сферы до ближайшей точки равно

$$\rho(d, N) = \sqrt[d]{1 - \sqrt[N]{1/2}}.$$

Найти предел $\rho(d, N)$ при $d \to \infty$, $N = O(d)$. Какой вывод из этого можно сделать применительно к методу ближайшего соседа при больших d?

28. Bias-variane trade-off. Рассмотрим задачу восстановления зависимости $Y = f*(X)$, где X - случайная величина, а f* - неизвестная детерминированная функция. Пусть $x^{(1)}, x^{(2)}, \ldots, x^{(N)}$ суть независимые реализации величины X. В качестве модельной зависимости возьмем функцию f(x, D), где D = x (1), x(2), . . . , x(N). Разложить $E_D\, f(x, D) - f^*(x)^2$ в сумму квадрата математического ожидания смещения (bias) и дисперсии (variance).

29. Может ли использование коррелированных переменных улучшить качество предсказания? Рассмотрим задачу классификации с двумя классами. Пусть пространство признаков двумерное. Объекты каждого класса имеют нормальное распределение с математическим ожиданием (-1,-1) и (1,1) соответственно и единичной матрицей коварации каждый. Априорные вероятности классов равны ½.

1) Вычислите коэффициент корреляции для переменных x_1, x_2.
2) Найти байесов классификатор и вычислить байесову ошибку для усеченной задачи, рассматривая только одну переменную x_1.
3) Найти байесов классификатор и вычислить байесову ошибку для исходной задачи.
4)Приводит ли использование второй переменной к уменьшению ошибки?

28. Рассмотрим задачу классификации с двумя классами 0 и 1. Пусть пространство признаков двумерное. Объекты каждого класса имеют нормальное распределение с математическим ожиданием (0,0) и (1,1) соответственно и матрицей ковариации

$$\mathbf{\Sigma} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Априорные вероятности классов равны $\Pr\{Y = 0\} = 1/3$ и $\Pr\{Y = 1\} = 2/3$.

1) Найти уравнение разделяющей поверхности байесова классификатора.
2) Найти собственное разложение матрицы Σ.
3) Перейти к новым координатам, оси которых совпадают с собственными векторами матрицы Σ.
4) Выписать уравнение разделяющей поверхности байесова классификатора в новых координатах.

29. Влияние шума на качество предсказания. Пусть пространство признаков одномерное и обучающая выборка состоит из двух объектов $x^{(0)} = 0$, $x^{(1)} = 1$. Добавим к объектам шумовой признак, распределенный равномерно на отрезке [-1,1]. Какова вероятность, что объект x = (0.32,0) окажется ближе (по евклидову расстоянию) к объекту $x^{(1)}$, чем к $x^{(0)}$? (Шум к x не добавляется. Только к $x^{(0)} = 0$ и $x^{(1)} = 1$.)

30. Выпуклым полиэдром (или выпуклым многогранником) в пространстве R^d называется пересечение конечного числа полупространств, т.е. множество решений некоторой системы линейных неравенств:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1d}x_d \le b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2d}x_d \le b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{md}x_d \le b_m, \end{cases}$$

Доказать, что область, в которой все точки имеют одинаковых k ближайших соседей (для евклидова расстояния) есть полиэдр.

31. В дереве решений (для некоторой задачи классификации с K классами) рассмотрим вершину m и соответствующий ящик R_m. Рассмотрим классификатор, который для объектов, попавших в ящик R_m выбирает класс случайно, причем класс k выбирается с вероятностью, равной p_{mk}. Доказать, что математическое ожидание частоты ошибок этого классификатора на объектах обучающей выборки, попавших в R_m, равно индексу Джини.

32. Пусть при построении дерева решений в задаче классификации с двумя классами в текущую вершину попало 400 объектов из первого класса и столько же из второго. Пусть необходимо сделать выбор между разбиением на две ветви (300,100) и (100,300) и разбиением на две ветви (100,400), (200,0). Какое из этих разбиений кажется предпочтительнее (объясните)? Какое разбиение выберет критерий на основе минимизации ошибки, энтропийный критерий и критерий Джини? Приведите свой пример, когда все три критерия дают разные разбиения.

33. Пусть в задаче классификации на 2 класса {0,1} некоторый классификатор (например, наивный байесовский) определяет следующие оценки g(x) апостериорной вероятности принадлежности объекта x к классу 1:

i	1	2	3	4	5	6	7	8	9
$y^{(i)}$	0	0	0	0	0	1	1	1	1
$g(x^{(i)})$	0.10	0.50	0.66	0.23	0.82	0.75	0.11	0.09	0.15

Постройте ROC-кривую. Вычислите AUC. Для классификатора $f(x) = I(f(x) \geq 0.5)$ выпишите матрицу рассогласования и найдите FPR, FNR, TNR, TPR, PPV, accuracy, error, F1.

Приложение 4

к рабочей программе

Федеральное государственное бюджетное образовательное учреждение высшего образования

УРАЛЬСКИЙ ГОСУДАРСТВЕННЫЙ ЭКОНОМИЧЕСКИЙ УНИВЕРСИТЕТ

УТВЕРЖДЕНЫ

на заседании кафедры информационных технологий и статистики

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ДЛЯ ПРОВЕДЕНИЯ

ТЕКУЩЕГО КОНТРОЛЯ

по дисциплине

Инструменты обработки и анализа корпоративных данных

Тест №1

1 ***Эконометрика занимается изучением***

1 экономической природы изучаемого объекта

2 числовых характеристик случайных величин

3 статистических таблиц распределений случайных величин

4 ▼ качественного и количественного влияния разных факторов на экономические объекты

2 ***Математической моделью в эконометрических задачах является***

1 ▼ уравнение регрессии или система уравнений регрессии

2 коэффициент корреляции

3 график зависимости переменных

4 показатели качества полученного уравнения регрессии

3 ***Метод наименьших квадратов состоит***

1 ▼ в минимизации суммы квадратов отклонений реальных значений y от расчетных

2 в поиске наименьшего квадрата коэффициента корреляции

3 в нахождении квадратов всех значений исследуемых переменных

4 в минимизации разностей попарных значений переменных

4 ***Для чего применяется МНК?***

1 для перехода от нелинейной формы зависимости переменных к линейной

2 для расчета Мат. ожидания, Нелинейности и Коэффициента вариации

3 ▼ для оценки параметров регрессии

4 для определения достоверности статистики

5 ***Можно ли на основании решения регрессионной задачи прогнозировать изменение Y в зависимости от изменения X?***

1 можно, только если регрессия - линейная

2 прогноз вообще невозможен на основании построения модели

3 ▼ можно, только если построенная регрессионная модель является качественной

4 всегда можно

6 ***Уравнение регрессии оценивает***

1 тесноту связи исследуемых переменных

2 причину наличия случайной составляющей

3 сумму квадратов отклонений реальных значений от расчетных

4 ▼ форму зависимости исследуемых переменных

7 ***Эндогенная переменная уравнения регрессии - это***

1 независимая переменная

2 ▼ зависимая переменная

3 бинарная переменная

4 свободный член уравнения

8 ***Экзогенная переменная уравнения регрессии - это***

1 фиктивная переменная

2 свободный член уравнения

3 зависимая переменная

4 ▼ независимая переменная

9 ***Чему будет равен Y в парной линейной регрессии, если Y-пересечение = 2, b = 5, x = 4?***

1 2

2 ▼ 22

3 5

4 4

10 ***Как в уравнении регрессии интерпретируется коэффициент перед переменной x?***

1 показывает статистическую значимость переменной x

2 ▼ показывает величину изменения y при единичном изменении x

3 показывает среднее значение x

4 показывает тесноту связи в уравнении регрессии

11 ***Нуль-гипотеза для какого-либо параметра состоит в предположении, что***

1 ▼ этот параметр является нулевым

2 этот параметр не учитывался в регрессионном анализе

3 этот параметр не нуждается в корректировке

4 этот параметр имеет нулевое отклонение от среднего значения

12 ***В уравнении y = a + bx незначимость Y-пересечения означает, что***

1 a + bx равно нулю

2 влияние переменной x на переменную y отсутствует

3 ▼ в уравнении регрессии отсутствует константа

4 статистические данные по y недостоверны

13 ***Коэффициент корреляции оценивает***

1 статистическую значимость коэффициентов регрессии

2 ▼ тесноту связи в уравнении регрессии

3 величину общей дисперсии независимой переменной

4 95-% интервал, в который попадают оценки истинных параметров регрессии

14 ***Как рассчитывается коэффициент детерминации?***

1 как доля остаточной дисперсии в общей дисперсии зависимой переменной

2 как доля значимых переменных в общем количестве переменных модели

3 как доля статистических выбросов в общем количестве наблюдений

4 ▼ как доля объясненной регрессией дисперсии в общей дисперсии зависимой переменной

15 ***Что показывает величина нормированного коэффициента детерминации?***

1 норму, рассчитанную для коэффициента детерминации конкретной модели

2 прямая или обратная связь существует между зависимой и независимой переменными

3 ▼ какая доля общей дисперсии объясняется включенными в регрессионную модель факторами

4 какая доля общей дисперсии объясняется выбросами

16 ***Какой должна быть сумма квадратов остатков при использовании МНК?***

1 отрицательной

2 положительной

3 максимальной

4 ▼ минимальной

17 ***Какое наблюдение считается статистическим выбросом?***

1 наблюдение, величина стандартного остатка которого по больше 0,05 (5%)

2 ▼ наблюдение, величина стандартного остатка которого по модулю больше 2

3 наблюдение, не вошедшее в выборку, по которой производится регрессионный анализ

4 наблюдение, порядковый номер которого больше 40

18 ***Переход к нелинейной модели регрессии осуществляется***

1 обязательно в любом процессе регрессионного исследования

2 ▼ в случае невозможности построения адекватной и качественной линейной модели

3 при наличии статистических выбросов

4 по желанию исследователя

19 ***Уравнение регрессии y = a + b ln(x) является***

1 ▼ моделью, нелинейной относительно независимой переменной

2 моделью, нелинейной относительно независимой переменной и параметров модели

3 моделью, нелинейной относительно зависимой переменной

4 линейной моделью

20 ***Что описывает функция Кобба-Дугласа?***

1 ▼ связь между национальным доходом и рабочей силой и производственными фондами

2 зависимость спроса на товары различных групп от дохода населения

3 парную нелинейную регрессионную зависимость

4 зависимость на основе бинарных переменных

21 ***Если в производственной функции Кобба-Дугласа b1 + b2 < 1, то имеет место***

1 постоянная отдача от масштаба

2 возрастающая отдача от масштаба

3 ▼ убывающая отдача от масштаба

4 наличие статистических выбросов

22 ***Если в производственной функции Кобба-Дугласа b1 + b2 > 1, то имеет место***

1 убывающая отдача от масштаба

2 ▼ возрастающая отдача от масштаба

3 наличие статистических выбросов

4 постоянная отдача от масштаба

23 ***Зачем в регрессионном анализе используются фиктивные переменные?***

1 для определения стандартных ошибок параметров регрессии

2 чтобы выявить связи независимых переменных между собой

3 ▼ чтобы учесть в модели факторы, выражающиеся не количественными значениями

4 для фиксации свойств найденной регрессии

24 ***Фиктивная переменная является***

1 показателем качества регрессионной модели

2 второстепенной переменной регрессионной модели, которую можно не учитывать в анализе

3 ▼ равноправной переменной регрессионной модели

4 константой

25 ***Может ли коэффициент при бинарной переменной быть отрицательным?***

1 да, если есть статистические выбросы

2 да, если бинарная переменная принимает больше двух значений

3 ▼ да

4 нет

26 ***Статистическая значимость бинарной переменной означает***

1 наличие этой переменной

2 наличие тесной связи между зависимой и объясняющими переменными модели

3 ▼ подтвержденное влияние данного качественного признака на зависимую переменную

4 достаточность статистических наблюдений для достоверных выводов

27 ***Временной ряд - это***

1 ▼ ряд значений одного показателя за несколько последовательных моментов времени

2 временное обозначение каких-либо регрессионных показателей

3 ряд значений какого-либо показателя за конкретный момент времени

4 рядовые данные о последовательных моментах времени

28 ***Что такое исходный уровень временного ряда?***

1 ▼ наблюдение (значение) временного ряда в один момент времени

2 минимальное из значений временного ряда

3 среднее значение временного ряда

4 дисперсия временного ряда

29 ***Какие модели существуют для описания временных рядов?***

1 модель последовательно-временных приближений

2 модель Торнквиста

3 трендовая, циклическая

4 ▼ аддитивная, мультипликативная

30 ***Факторы, формирующие временной ряд, делятся на группы:***

1 положительная и отрицательная составляющая

2 линейная и нелинейная компоненты

3 ▼ трендовая, циклическая и случайная компоненты

4 временная и пространственная компоненты

31 ***Тренд временного ряда - это***

1 ▼ аналитическая функция, характеризующая зависимость уровней ряда от времени

2 значение коэффициента детерминации, рассчитанного по данным временного ряда

3 суммарное значение периодов времени по всему ряду

4 временное преобразование уровней ряда

32 ***Метод последовательных разностей заключается***

1 ▼ в замене исходных уровней ряда первыми или вторыми разностями

2 в последовательном вычитании стандартных остатков из остатков

3 в последовательном удалении статистических выбросов

4 в нахождении разности коэффициентов корреляции и детерминации

33 ***Если тенденция временного ряда представляет собой параболу 2-го порядка, то для моделирования тренда***

1 можно применить метод первых разностей

2 ▼ можно применить метод вторых разностей

3 можно применять только аналитическое выравнивание

4 нет подходящих условий

34 ***Нелинейный тренд временного ряда находится***

1 после исключения линейной трендовой компоненты

2 только для мультипликативной модели

3 ▼ путем линеаризации

4 исключением сезонной компоненты

35 ***Может ли циклическая компонента временного ряда быть нелинейной?***

1 ▼ да, более того, она не может быть линейной

2 нет, она может быть только линейной

3 да - она, в принципе, может быть как линейной, так и нелинейной

4 да, при условии введения бинарных переменных

36 ***В каких случаях при моделировании циклической компоненты временного ряда используется аддитивная модель?***

1 только при наличии трендовой компоненты

2 только при отсутствии трендовой компоненты

3 ▼ если амплитуда колебаний временного ряда приблизительно постоянна

4 если амплитуда колебаний временного ряда возрастает или уменьшается

37 ***В каких случаях при моделировании циклической компоненты временного ряда используется мультипликативная модель?***

1 если амплитуда колебаний временного ряда приблизительно постоянна

2 ▼ если амплитуда колебаний временного ряда возрастает или уменьшается

3 только при наличии трендовой компоненты

4 только при отсутствии трендовой компоненты

38 ***Что такое условия Гаусса-Маркова?***

1 условия соглашения между Гауссом и Марковым

2 описание процедуры регрессионного анализа

3 обоснование метода наименьших квадратов

4 ▼ предпосылки для применения метода наименьших квадратов

39 ***Какими свойствами должны обладать оценки параметров регрессии, проведенной с помощью МНК?***

1 парность и/или множественность

2 коррелированность

3 ▼ несмещенность, состоятельность и эффективность

4 понятность, логичность и корректность

40 В каком случае может проявляться мультиколлинеарность?

1 в случае парной регрессии

2 ▼ в случае множественной регрессии

3 только в случае решения задач нелинейной регрессии

4 в случае наличия большого количества статистических выбросов

41 Что такое мультиколлинеарность?

1 мультипликативная модель регрессии

2 независимость эндогенной и экзогенных переменных

3 нелинейность одной или нескольких объясняющих переменных

4 ▼ линейная взаимосвязь двух или нескольких объясняющих переменных

42 Может ли зависимая переменная быть мультиколлинеарной?

1 да

2 нет, если только она не бинарная

3 да, если рассматривается модель временных рядов

4 ▼ нет, вообще мультиколлинеарность - свойство модели, а не переменной

43 Следствием наличия мультиколлинеарности в регрессионной модели является

1 ▼ невозможность однозначного и достоверного определения коэффициентов регрессии

2 невозможность определения тесноты связи искомой регрессионной зависимости

3 появление статистических выбросов

4 изменение средних значений величин исходных данных проводимой регрессии

44 Чем характеризуется гетероскедастичность?

1 ▼ зависимостью дисперсии ошибок от номера наблюдения

2 постоянством дисперсии ошибок

3 равенством нулю дисперсии ошибок

4 отсутствием ошибок

45 ***Чем характеризуется гомоскедастичность?***

1 равенством нулю дисперсии ошибок

2 отсутствием ошибок

3 ограниченным объемом анализируемых данных

4 ▼ постоянством дисперсии ошибок

46 ***В чем проявляются отрицательные последствия гетероскедастичности?***

1 в невозможности линейного преобразования уравнения регрессии

2 в немотивированном появлении статистических выбросов

3 ▼ в неэффективности оценок МНК

4 в невозможности расчета стандартных ошибок параметров регрессии

47 ***Что такое автокорреляция?***

1 следствие использования меода последовательных разностей

2 наличие сильной корреляции между переменными модели

3 ▼ корреляция между показателями, упорядоченными во времени или в пространстве

4 равенство остатков (отклонений) во всех наблюдениях

48 ***Какой вид автокорреляции более характерен для экономических процессов?***

1 линейная

2 нелинейная

3 ▼ положительная

4 отрицательная

49 ***Каковы последствия автокорреляции?***

1 ▼ выводы о значимости коэффициентов регрессии и корреляции могут быть неверны

2 невозможность рассчитать доверительные интервалы

3 нарушается структура исходных данных

4 появляются статистические выбросы, которые не определяются обычными способами

50 ***В каком случае в эконометрике применяются системы одновременных уравнений?***

1 когда обычный МНК неприменим

2 когда одновременно рассматривается гомоскедастичные и гетероскедастичные модели

3 ▼ когда изучения изолированных уравнений недостаточно

4 по желанию исследователя

Тест №2

Вопрос		Ответы
Методы кластерного анализа предназначены для	1	сокращения размерности признакового пространства
	2	**разделения совокупности объектов на однородные группы**
	3	построения эконометрических моделей
	4	определения собственных значений объектов
Факторный анализ дает возможность	1	разделить совокупность объектов на однородные группы
	2	выполнить классификацию объектов
	3	**лаконичного и более простого объяснения многомерных структур**
	4	построить систему эконометрических уравнений
Кластерный анализ временных рядов позволяет	1	выявлять автокорреляцию уровней временного ряда
	2	исследовать временные ряды на наличие тренда
	3	кластерный анализ временных рядов дает практических результатов
	4	**выделять периоды, когда значения показателей были достаточно близкими**

Радиус кластера представляет собой	1	максимальное расстояние между несколькими кластерами
	2	**максимальное расстояние точек от центра кластера**
	3	минимальное расстояние точек от центра кластера
	4	минимальное расстояние между несколькими кластерами
Стандартизация переменных в кластерном анализе используется	1	**для устранения неоднородности единиц измерения признаков**
	2	не используется в кластерном анализе
	3	только при анализе временных рядов
	4	для сокращения размеров кластера

Иерархические методы кластерного анализа	1	**последовательном объединении меньших кластеров в большие или разделении больших кластеров на меньшие**
	2	предполагают задание некоторых начальных условий
	3	проводят классификацию с объектами, имеющими не менее 15 характеристик
	4	проводят классификацию с объектами, имеющими не более 15 характеристик
Методы и модели факторного анализа предназначены	1	для упорядочивания исходного признакового пространства по возрастанию признаков
	2	**для снижения размерности исходного признакового пространства**
	3	для упорядочивания исходного признакового пространства по убываанию признаков
	4	для исключения коррелированных исходных признаков
В чем заключается принципиальное отличие модели факторного анализа от регрессионных схем	1	**в моделях факторного анализа переменные не являются непосредственно наблюдаемыми**
	2	в моделях факторного анализа переменные д.б. коррелированными между собой
	3	в моделях факторного анализа переменные д.б. непосредственно наблюдаемыми
	4	в моделях факторного анализа переменные измеряются на статистически обследованных объектах

Общие факторы в моделях факторного анализа - это	1	в моделях факторного анализа все факторы являются общими
	2	факторы, которые одновременно влияют только на одну переменную
	3	факторы, к которым относится погрешность в наблюдениях
	4	**факторы, которые влияют на несколько переменных одновременно**
Целью вращениие факторов является	1	переход к линейной структуре данных
	2	**получение структуры, при которой большинство наблюдений находится вблизи осей координат**
	3	переход к нелинейной структуре данных
	4	переход к однородной структуре данных

Задание №	Варианты ответа
	Укажите не менее двух вариантов ответа
На сколько ед. и каким образом можно увеличить для худшего кластера значение Y за счет фактора x15	☐ 12
$Y = 27 - 3x3 + 12x7 - 7x11 + 24x15 - 5x17$	☐ 120
	☐ 24
	☐ увеличивая влияние фактора
	☐ 17
	☐ снижая влияние фактора

Показатели регрессионной	Средние значения	
	кластер 1	кластер 2
нерегулируемые факторы		
x7	41	38
x11	27	31
регулируемые факторы		
x3	23	28
x15	17	12
x17	54	57

Задание №

Варианты ответа

Определить модельное значение Y для худшего кластера

$$Y = 27 - 3x3 + 12x7 - 7x11 + 24x15 - 5x17$$

Показатели регрессионной	Средние значения	
	кластер 1	кластер 2
нерегулируемые факторы		
x7	41	38
x11	27	31
регулируемые факторы		
x3	23	28
x15	17	12
x17	54	57

- 399
- 158
- 214
- 372
- 185

Задание №

Варианты ответа

Найти для худшего кластера процент роста Y за счет изученных нерегулируемых факторов

$$Y = 15 - 3x2 - 2x3 + 0,5x5 + 30x7 + 5x12$$

Показатели регрессионной	Средние значения	
	кластер 1	кластер 2
1	2	3
нерегулируемые факторы		
x3	20	30
x5	24	18
регулируемые факторы		
x2	15,0	19
x7	12	10
x12	95	91

- 4,9%
- 13,9%
- 17,4%
- 6,5%
- 3,5%

Задание №

Варианты ответа

Укажите не менее двух вариантов ответа

6-е предприятие относится к худшему кластеру за счет

лучший кластер

	$Y_{расч}$	$Y_{кластер-2} - Y_{среднее}$	$Y_{расч} - Y_{кластер-2}$	$Y_{кластер-1} - Y_{кластер-2}$	$Y_{расч} - Y_{кластер-1}$
1	77,12	8,24	-7,02	13,73	1,30
2	67,05	3,14	-1,32	7,10	-1,22
7	62,78	12,78	-14,77	-15,62	4,85
8	74,74	-2,12	21,05	-17,93	3,12
9	81,23	-4,13	8,53	11,32	-2,80

худший кластер

	$Y_{расч}$	$Y_{кластер-2} - Y_{среднее}$	$Y_{расч} - Y_{кластер-2}$	$Y_{кластер-1} - Y_{кластер-2}$	$Y_{расч} - Y_{кластер-1}$
3	51,51	-17,95	15,89	16,06	-3,17
4	57,75	-18,76	-7,94	10,05	2,11
5	56,79	7,99	-0,62	-0,87	0,25
6	53,56	5,00	18,35	-11,72	-6,63
12	59,36	-2,17	-6,71	5,49	1,21

- низкого нормативного уровня выработки
- случайно отнесено к худшему кластеру
- негативного влияния неучтенных факторов
- негативного влияния нерегулируемых факторов
- неэффективного использования изученных регулируемых факторов

Задание №

Варианты ответа

Укажите не менее двух вариантов ответа

1-е предприятие относится к лучшему кластеру за счет

лучший кластер

	$Y_{расч}$	$Y_{кластер-2} - Y_{среднее}$	$Y_{расч} - Y_{кластер-2}$	$Y_{кластер-1} - Y_{кластер-2}$	$Y_{расч} - Y_{кластер-1}$
1	77,12	-8,24	-7,02	13,73	1,30
2	67,05	3,14	-1,32	7,10	-1,22
7	62,78	12,78	-14,77	-15,62	4,85
8	74,74	-2,12	21,05	-17,93	3,12
9	81,23	-4,13	8,53	11,32	-2,80

худший кластер

	$Y_{расч}$	$Y_{кластер-2} - Y_{среднее}$	$Y_{расч} - Y_{кластер-2}$	$Y_{кластер-1} - Y_{кластер-2}$	$Y_{расч} - Y_{кластер-1}$
3	51,51	-17,95	15,89	16,06	-3,17
4	57,75	-18,76	-7,94	10,05	2,11
5	56,79	7,99	-0,62	-0,87	0,25
6	53,56	5,00	18,35	-11,72	-6,63
12	59,36	-2,17	-6,71	5,49	1,21

- позитивного влияния нерегулируемых факторов
- эффективного использования изученных регулируемых факторов
- случайно отнесено к лучшему кластеру
- позитивного влияния неучтенных факторов
- высокого нормативного уровня выработки

Задание №

5-е предприятие относится к худшему кластеру за счет

лучший кластер

	$Y_{факт}$	$Y_{кластер\text{-}2} - Y_{среднее}$	$Y_{факт} - Y_{кластер\text{-}2}$	$Y_{кластер\text{-}1} - Y_{кластер\text{-}2}$	$Y_{факт} - Y_{кластер\text{-}1}$
1	77,12	-8,24	-7,02	13,73	1,30
2	67,05	3,14	-1,32	7,10	-1,22
7	62,78	12,78	-14,77	-15,62	4,85
8	74,74	-2,12	21,05	-17,93	3,12
9	81,23	-4,13	8,53	11,32	-2,80

худший кластер

	$Y_{факт}$	$Y_{кластер\text{-}2} - Y_{среднее}$	$Y_{факт} - Y_{кластер\text{-}2}$	$Y_{кластер\text{-}1} - Y_{кластер\text{-}2}$	$Y_{факт} - Y_{кластер\text{-}1}$
3	51,51	-17,95	15,89	16,06	-3,17
4	57,75	-18,76	-7,94	10,05	2,11
5	56,79	7,99	-0,62	-0,87	0,25
6	53,56	5,00	18,35	-11,72	-6,63
12	59,36	-2,17	-6,71	5,49	1,21

Варианты ответа

Укажите не менее двух вариантов ответа

- [ ] неэффективного использования изученных регулируемых факторов
- [ ] негативного влияния нерегулируемых факторов
- [ ] случайно отнесено к худшему кластеру
- [ ] негативного влияния неучтенных факторов
- [ ] низкого нормативного уровня выработки

Реферат

Объем реферата – 25-35 страниц. Реферат оформляется по всем требования «Положения об оформлении…». Реферат должен содержать следующие структурные элементы: титульный лист, содержание, введение, основная часть (две главы – теоретическая и практическая, разбитые на параграфы), заключение, список использованных источников, приложения (если имеются). Реферат должен содержать не менее 51% оригинального текста (по системе antiplagiat.ru). Реферат сдается в электронном виде на портал ЭОР (в раздел «Задания»), а после его проверки преподавателем распечатывается и сдается преподавателю лично.

Реферат должен быть написан студентом самостоятельно с использованием источников (учебники, книги, монографии, научные статьи, официальные сайты). Студент обязан в полном объеме владеть материалом темы, представленной в реферате. Преподаватель имеет право во время зачетного занятия задать дополнительные вопросы по теме реферата.

Темы рефератов:

1. Методы многомерного статистического анализа как сформировавшаяся самостоятельная область теоретической статистики, их особенности и отличия от методов классической статистики.
2. Классификация многомерных статистических методов (МСМ).
3. Параметрические и непараметрические методы: особенности, перечень, необходимость использования при моделировании социально-экономических явлений и процессов.
4. Системность и комплексность как базовые принципы изучения сложных социально-экономических явлений и процессов.
5. Примеры применения многомерных статистических методов в социально-экономических исследованиях.
6. Понятие признакового пространства. Примеры одномерного, двумерного и многомерного признакового пространства.
7. Особенности статистического анализа многомерных данных.
8. Двумерные и многомерные случайные величины. Дискретные и непрерывные многомерные случайные величины.
9. Распределения многомерных случайных величин: определения и примеры.
10. Особенности практического приложения и взаимные связи разнообразных распределений при решении задач в экономике, социологии, психологии и т.д.
Грубые ошибки и причины их появления в статистической совокупности. Методы выявления грубых ошибок в статистической совокупности данных: критерий Смирнова-Граббса, критерии Титьена-Мура.
11. Устойчивое оценивание: итеративная процедура Хубера, алгоритм последовательного улучшения данных по Винзору. Критериальная или логическая проверка устойчивого оценивания.
12. Многомерный случай «засорения» совокупности данных. Интерпретация результатов устойчивого оценивания. Рассмотрение практических примеров анализа экономических данных, поиск и устранение ошибок.
13. Робастное статистическое оценивании в моделировании временных рядов.
14. Цели и задачи фундаментального анализа. Понятия и определения в фундаментальном анализе: общий фактор, латентный фактор, элементарный признак и т.д.
15. Основная концепция фундаментального анализа. Геометрическая интерпретация наблюдаемых объектов в фундаментальном анализе.

16. Сущность методов фундаментального анализа и их классификация: простые методы, аппроксимирующие методы.
17. Существование модели фундаментального анализа. Теорема Терстоуна.
18. Вычислительная схема фундаментального анализа.
19. Определение структуры редуцированной корреляционной матрицы: методом наибольшей корреляции, методом Барта, методом триад, методом малого центроида.
20. Состав дисперсии элементарного признака в фундаментальном анализе: полная дисперсия, общность, характерность, специфичность, ненадежность, надежность.
21. Методы уточнения моделей фундаментального анализа.
22. Вращение факторного пространства. Типы вращения: ортогональное и косоугольное.
23. Построение матриц преобразования для факторов при вращении. Основные подходы и рекомендации к выбору угла вращения системы факторов.
24. Дисперсия факторных нагрузок как мера сложности структуры факторов. Критерии качества структуры общих факторов: квартимакс, варимакс, облимакс, квартимин, облимин.
25. Статистические критерии проверки надежности решений методами фундаментального анализа. Использование критериев Уилкса и Лоули для проверки значимости матрицы парных корреляций и в оценке достаточности общих факторов.
26. Использование коэффициента конгруэнтности и подхода Хармана в оценке качества факторных решений.
27. Применение фундаментального анализа в маркетинговых исследованиях (выявление потребительских мотиваций, предпочтений при выборе товара).
28. Цели и задачи многомерного шкалирования. История появления и развития методов многомерного шкалирования, их отличительные особенности от других многомерных статистических методов.
29. Понятия и определения, применяемые в многомерном шкалировании: событие-стимул, стимульное пространство, шкала, эксперт, предпочтение, профиль, стресс-формула и др. Представление и первичная обработка данных в многомерном шкалировании.
30. Использование матрицы условных и совместных вероятностей, матрицы перехода для анализа данных. Нормирование разнотипных данных и построение матрицы мер различия профилей.

Приложение 5

к рабочей программе

Федеральное государственное бюджетное образовательное учреждение высшего образования

УРАЛЬСКИЙ ГОСУДАРСТВЕННЫЙ ЭКОНОМИЧЕСКИЙ УНИВЕРСИТЕТ

УТВЕРЖДЕНЫ

на заседании кафедры информационных технологий и статистики

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ДЛЯ ПРОВЕДЕНИЯ

ПРОМЕЖУТОЧНОГО КОНТРОЛЯ

по дисциплине

Инструменты обработки и анализа корпоративных данных

Экзаменационный билет №1

1. Вероятностная постановка задачи обучения с учителем. Функция потерь. Средний (ожидаемый) риск. Эмпирический риск. Регрессионная функция и байесов классификатор. Неустранимая (байесовская) ошибка.

2. Понятие о глубоком обучении. Сверточные нейронные сети.

3. Пусть g: $R^n \to R^m$, x ∈ Rn. Матрицей Якоби называется матрица

$$\frac{\partial g}{\partial x} = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} & \cdots & \frac{\partial g_2}{\partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial g_m}{\partial x_1} & \frac{\partial g_m}{\partial x_2} & \cdots & \frac{\partial g_m}{\partial x_n} \end{pmatrix}.$$

В частности, если m = 1 (т. е. g(x) — скалярная функция векторного аргумент x), то $\partial g/\partial x$ — градиент функции g.

Доказать, что

1) если $a \in R^n$, $x \in R^n$, то $\partial(a^\top x)/\partial x = a$;

2) если $A \in R^{m\times n}$, $x \in R^n$, то $\partial(Ax)/\partial x = A$;

3) если $A \in R^{n\times n}$, $x \in R^n$, òî $\partial(x^\top Ax)/\partial x = (A + A^\top)x$; в частности, если $A^\top = A$, то $\partial(x^\top Ax)/\partial x = 2Ax$;

4) если x ∈ Rn , то $\partial|x|^2/\partial x = 2x$;

5) если g — скалярная функция и под g(x) понимается применение функции g к каждой компоненте вектора x ∈ R^n, то

$\partial g(x)/\partial x = \mathrm{diag}(g'(x))$,

где diag(a) — диагональная матрица с диагональю a;

6) если $h : R^n \to R^m$, $g : R^m \to R^p$, $x \in R^n$, то

$\partial g(h(x))/\partial x = (\partial g(h(x))/\partial h)(\partial h(x)/\partial x)$.

Экзаменационный билет №2

1. Ошибки 1-го и 2-го рода. Чувствительность, специфичность, точность, полнота. ROC-кривая. Площадь под ROC-кривой.

2. Машина опорных векторов. Формулировка задачи в виде задачи математического программирования. Двойственная задача. (Случай линейно разделимых и неразделимых классов)

3. Пользуясь №1, найдите градиент $\partial g(\beta)/\partial\beta$ и гессиан $(\partial^2 g(\beta))/(\partial\beta^\top\partial\beta)$

функции $g(\beta) = |X\beta - y|^2$. Выведите отсюда, что решение линейной задачи наименьших квадратов $\hat{\beta} = \mathrm{argmin}\, |X\beta - y|^2$ является решением нормальной системы линейных уравнений $X^\top X\beta = X^\top y$.

Экзаменационный билет №3

1. Принцип минимизации эмпирического риска. Минимизация отступа. Регуляризация.

2. Машина опорных векторов. Ядра и спрямляющие пространства.

3. Дана обучающая выборка

x	1	1	0	0	-1
y	4	4	0	2	6

1) изобразить точки;

2) методом наименьших квадратов построить модель вида $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$; построить график этой функции;

3) построить модель того же вида методом ридж-регрессии с параметром регуляции $\lambda = 1$; построить график этой функции.

Замечание: при ручных вычислениях по методу наименьших квадратов квадратов рекомендуется составить систему $X^\intercal X\beta = X^\intercal y$ и решить ее. Регулязированная система $(X^\intercal X + \lambda I)\beta = X^\intercal y$, где I — единичная матрица.

Экзаменационный билет №4

1. Экспериментальная (эмпирическая) оценка качества обучения. Обучающая, проверочная и тестовая выборки. Метод перекрестного (скользящего) контроля.

2. Деревья решений. Алгоритм CART.

3. Рассмотрим задачу восстановления регрессии, в которой y распределен согласно нормальному закону $N(X\beta, \sigma^2 I)$, а β имеет априорное распределение $N(0, \tau I)$. Найти апостериорное распределение для β. Доказать, что β^{ridge} есть его математическое ожидание. Найти связь между параметром регуляризации λ и дисперсиями τ , σ^2.

Экзаменационный билет №5

1. Метод k ближайших соседей в задачах классификации и восстановления регрессии. Теорема об оценке риска в методе ближайшего соседа. Идея доказательства.

2. Случайный лес. Экстремально случайные деревья.

3. Показать, что процедура гребневой регрессии эквивалентна обычному методу наименьших квадратов, примененному к расширенным данным: к центрированной матрице X дописывается матрица $(\sqrt{\lambda})I$, к вектору y приписывается d нулей.

Экзаменационный билет №6

1. Метод наименьших квадратов. Система нормальных уравнений. Псевдорешение.

2. Алгоритм XGBoost.

3. Показать, как (и объяснить почему) задачу квадратичного программирования в методе лассо

$$\widehat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{d} x_{ij}\beta_j \right)^2 \right\},$$

при условии

$$\sum_{j=1}^{d} |\beta_j| \leq s$$

можно свести к задаче квадратичного программирования с 2d+1 неизвестными и 2d+1 линейными ограничениями.

Экзаменационный билет №7

1. Борьба с переобучением в методе наименьших квадратов. Сокращение числа параметров. Полный перебор всех подмножеств признаков. Жадный (Forward stepwise) алгоритм.

2. Градиентный бустинг деревьев решений.

3. Метод использования линейной регрессии в задаче классификации заключается в следующем. Сопоставим каждому классу k вектор $(y_1,y_2,...y_K)$, в котором $y_k = 1$? а $y_i = 0$ при $i \neq k$. Собрав вместе индикаторные векторы объектов обучающей выборки, получим матрицу Y размера $N \times K$. Пусть X — матрица размера $N \times (d + 1)$, первый столбец которой состоит из единиц, а последующие представляют собой векторы из обучающей выборки. Применяя метод наименьших квадратов одновременно к каждому столбцу матрицы Y, получаем значения

$\hat{Y} = X(X^TX)^{-1}X^TY.$

Для каждого столбца y_k матрицы Y получим свой столбец коэффициентов $\hat{\beta}_k$. Соберем их в матрицу $\hat{B}$ размера $(d + 1) \times K$. Имеем

$\hat{B} = (X^TX)^{-1}X^TY.$

Объект x будем классифицировать согласно следующему правилу: Вычислим вектор-строку длину K

$g(x) = (1, x)\,\hat{B}.$

Отнесем x к классу

$f(x) = \underset{k}{\operatorname{argmax}}\, g_k(x)$

Доказать, что

$$\sum_{k=1}^{K} g_k(x) = 1.$$

Доказать, что в случае K = 2 данный метод эквиваленте решению одной задачи восстановления регрессии. Какой?

Экзаменационный билет №8

1. Борьба с переобучением в методе наименьших квадратов. Ридж-регрессия (регуляризация). Метод "Лассо".

2. Дилемма "Смещение-разброс". Кривая обучения.

3. Дана обучающая выборка

x_1	0	1	0	2	2	2	4	3
x_2	-1	0	0	0	1	0	1	2
y	0	0	0	0	0	1	1	1

1) Методом линейного дискриминантного анализа для каждого класса построить дискриминатную функцию и записать уравнение разделяющей поверхности.

2) Методом квадратичного дискриминантного анализа построить дискриминантные функции.

Изобразить точки и разделяющие поверхности (кривые)

Экзаменационный билет №9

1. Наивный байесовский классификатор. Сглаживание Лапласа. Использование наивного байесовского классификатора для количественых признаков.

2. Задача понижения размерности. Метод главных компонент.

3. Задача Фишера сводится к максимизации отношения Рэлея

$\max_a \ (a^T Ba)/(a^T Wa)$.

Показать, как эта задача сводится к обощенной задаче на собственные значения

$Ba = \lambda W$.

Экзаменационный билет №10

1. Линейный дискриминантный анализ. Квадратичный дискриминантный анализ.

2. Задача кластеризации. Метод центров тяжести. Метод медоидов.

3. Показать, что оптимальная гиперплоскость, разделяющая два множества, является плоскостью, проходящей через середину отрезка, соединяющего пару ближайших точек из выпуклой оболочки

каждого из классов, и перпендикулярно ему. Указание: рассмотреть задачу, двойственную к задаче определения оптимальной гиперплоскости.

Экзаменационный билет №11

1. Логистическая регрессия. Логистическая функция и softmax

2. Алгоритм "Ожидание-максимизация".

3. Показать, что в алгоритме SVM задача

$$\min_{\beta,\beta_0,\xi_i} \frac{1}{2}\|\beta\|^2 + \gamma\sum_{i=1}^{N}\xi_i,$$

при ограничениях

$y_i(x_i^\top\beta + \beta_0) \geq 1 - \xi_i$, $\xi_i \geq 0$ (i = 1, 2, . . . , N),

эквивалентна задаче

$$\min_{\beta,\beta_0} \sum_{i=1}^{N}\left[1 - y_i\left(x_i^\top\beta + \beta_0\right)\right]_+ + \alpha\|\beta\|^2,$$

где $[\cdot]_+$ означает положительную часть, и $\alpha = 1/(2\gamma)$.

Экзаменационный билет №12

1. Математические переменные в экономических исследованиях; анализ связи переменных; множественный R; его экономический смысл; расчет; оценка генеральной совокупности; характеристика вида и тесноты связи переменных

2. Виды расстояний между кластерами.

3. SVM и задача восстановления регрессии. Для восстановления β , β_0 в модели $f(x) = x^\top\beta + \beta_0$. Рассмотрим задачу минимизации функции

$$H(\beta,\beta_0) = \sum_{i=1}^{N} V\left(y_i - f(x_i)\right) + \frac{\alpha}{2}\|\beta\|^2,$$

где

$$V(t) = V_\varepsilon(t) = \begin{cases} 0, & \text{если } |t| < \varepsilon, \\ |t| - \varepsilon & \text{в противном случае.} \end{cases}$$

Доказать, что решение β^, $\beta^_0$, минимизирующее функцию þ H(β, β_0), можно представить в виде

$$\widehat{\beta} = \sum_{i=1}^{N} (\widehat{\alpha}_i^* - \widehat{\alpha}_i) x_i, \quad \widehat{f}(x) = \sum_{i=1}^{N} (\widehat{\alpha}_i^* - \widehat{\alpha}_i) \langle x, x_i \rangle + \beta_0,$$

где $a^_i$ и $a^_i$* являются решением следующей задачи квадратичного программирования:

$$\min_{\alpha_i, \alpha_i^*} \left(\varepsilon \sum_{i=1}^{N} (\alpha_i^* + \alpha_i) - \sum_{i=1}^{N} y_i (\alpha_i^* - \alpha_i) + \frac{1}{2} \sum_{i,j=1}^{N} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) \langle x_i, x_j \rangle \right)$$

при ограничениях

$$0 \le \alpha_i, \alpha_i^* \le \frac{1}{\lambda}, \quad \sum_{i=1}^{N} (\alpha_i^* - \alpha_i) = 0, \quad \alpha_i \alpha_i^* = 0.$$

Экзаменационный билет №13

1. Понятие о глубоком обучении. Сверточные нейронные сети.

2. Алгоритм DBSCAN.

3. Дана обучающая выборка

x_1	-1	-1	1	1
x_2	-1	1	-1	1
y	1	-1	-1	1

Подобрать ядро и указать соответствующий SVM-классификатор, для которого ошибка на обучающей выборке равна 0.

Экзаменационный билет №14

1. Машина опорных векторов. Формулировка задачи в виде задачи математического программирования. Двойственная задача. (Случай линейно разделимых и неразделимых классов)

2. Алгоритмы иерархической кластеризации.

Дана обучающая выборка

x_1	0	0	1	1	0	0	1	1	1	0
x_2	0	1	0	1	1	1	1	1	1	1
y	0	0	0	0	0	1	1	1	1	1

С помощью наивного байесова классификатора оценить вероятность $Pr(Y = 0|X_1 = 1, X_2 = 1)$; $Pr(Y = 1|X_1 = 1, X_2 = 1)$.

Экзаменационный билет №15

1. Машина опорных векторов. Ядра и спрямляющие пространства.

2. Вероятностная постановка задачи обучения с учителем. Функция потерь. Средний (ожидаемый) риск. Эмпирический риск. Регрессионная функция и байесов классификатор. Неустранимая (байесовская) ошибка.

3. Дана выборка

x_1	4	0	-2	2
x_2	3	1	-3	-1

Найти главные направления и дисперсии по главным компонентам.

Изобразить точки и главные направления.

Экзаменационный билет №16

1. Ошибки 1-го и 2-го рода. Чувствительность, специфичность, точность, полнота. ROC-кривая. Площадь под ROC-кривой.

2. Деревья решений. Алгоритм CART.

3. Дана выборка

x_1	4	0	-1	3	4
x_2	2	-3	-2	1	2
x_3	3	2	2	1	-3

Найти главные направления и дисперсии по главным компонентам.

Экзаменационный билет №17

1. Случайный лес. Экстремально случайные деревья.

2. Принцип минимизации эмпирического риска. Минимизация отступа. Регуляризация.

3. Пусть $\sigma(z) = 1/(1 + e^{-z})$ (сигмоидальная функция). Проверить, что $\sigma' = \sigma(1 - \sigma)$.

Экзаменационный билет №18

1. Алгоритм XGBoost.

2. Экспериментальная (эмпирическая) оценка качества обучения. Обучающая, проверочная и тестовая выборки. Метод перекрестного (скользящего) контроля.

3. Пусть в задаче классификации на K классов $\{1,2,\ldots,K\}$ последний слой нейронной сети вычисляет softmax-функцию:

$$g_k(s_1, s_2, \ldots, s_K) = e^{s_k} \Big/ \sum_{\ell=1}^{K} e^{s_\ell} .$$

В качестве штрафа используется кросс-энтропия (logross-функция):

$$R^{(i)} = -\sum_{k=1}^{K} I(y^{(i)} = k) \ln g_k(s_1, s_2, \ldots, s_K),$$

где $g_k(s_1,s_2,\ldots,s_K)$ — softmax-функция. Доказать, что

1) $\partial g_k/\partial s_\ell = g_k \cdot I(k = \ell) - g_\ell$;

2) $\partial R^{(i)}/\partial gk = - I(y^{(i)})/g_k$;

3) $\partial R^{(i)}\, \partial s_\ell = g_\ell - I(\ell = y^{(i)})$.

Экзаменационный билет №19

1. Градиентный бустинг деревьев решений.

2. Метод k ближайших соседей в задачах классификации и восстановления регрессии. Теорема об оценке риска в методе ближайшего соседа. Идея доказательства.

3. Предположим, что рассматривается K задач двуклассовой классификации, в каждой из которых по $x \in X$ требуется предсказать $y_K \in \{0,1\}$ (k = 1, 2, …, K) (например, требуется определить, присутствует или отсутствует на изображении x объект k). Обучающая выборка составлена из пар $(x^{(i)},y^{(i)})$, где $y^{(i)} = (y^{(i)}_1, y^{(i)}_2, \ldots, y^{(i)}_K) \in \{0, 1\}^K$ (i=1, 2, …, N). Для решения такой задачи можно использовать нейронную сеть, последний слой которой вычисляет K сигмоидальных функций

$g_k(s_k) = 1/(1 + e^{-sk})$ (k = 1, 2, . . . , K).

Пусть в качестве штрафа используется

$$R^{(i)} = -\sum_{k=1}^{K} \left(y_k^{(i)} \ln g_k + (1 - y_k^{(i)}) \ln(1 - g_k) \right).$$

Доказать, что

$\partial R^{(i)}/\partial sk = g_k - y^{(i)}{}_k$.

Экзаменационный билет №20

1. Нейронная сеть с двумя нелинейными слоями вычмсляет функцию saftmax(B(σ(Ax))). Пусть в качестве функции потерь используется logloss. Таким образом, на объекте $x^{(i)}$ потери равны

logloss(softmax(B(σ(Ax)))). Пользуясь результатом задачи №1, выпишите матричные формулы для алгоритма backpropagation.

2. Метод наименьших квадратов. Система нормальных уравнений. Псевдорешение.

3. Нейронная сеть с двумя нелинейными слоями вычмсляет функцию saftmax(B(σ(Ax))). Пусть в качестве функции потерь используется logloss. Таким образом, на объекте $x^{(i)}$ потери равны logloss(softmax(B(σ(Ax)))). Пользуясь результатом задачи №1, выпишите матричные формулы для алгоритма backpropagation.

Экзаменационный билет №21

1. Задача понижения размерности. Метод главных компонент.

2. Борьба с переобучением в методе наименьших квадратов. Сокращение числа параметров. Полный перебор всех подмножеств признаков. Жадный (Forward stepwise) алгоритм.

3. Пусть рассматривается задача 2-классовой классификации. Доказать, что если известно сколько в выборке представителей каждого из двух классов, то по любым двум показателям из списка TPR, TNR, PPV, NPV определяются остальные два.

Экзаменационный билет №22

1. Логистическая регрессия. Логистическая функция и softmax

2. Задача кластеризации. Метод центров тяжести. Метод медоидов.

3. Пусть рассматривается задача 2-х классовой классификации. Верно ли, что

1) если у двух классификаторов на одной и той же выборке совпадают PPV (Precision) и совпадают TPR (Recall или Sensitivity), то будут совпадать TNR (Specificity) и NPV;

2) если у двух классификаторов на одной и той же выборке совпадают TNR (Specificity) и совпадают NPV, то будут совпадать PPV(Precision) и TPR (Recall или Sensitivity);

3) совпадение ROC кривых(для двух классификаторов на одной и той же выборке) влечет совпадение Precision-Recall кривых и наоборот?

Экзаменационный билет №23

1. Алгоритм "Ожидание-максимизация".

2. Нейронные сети. Алгоритм обучения Back-Propagation.

3. Рассмотрим задачу восстановления регрессии с квадратичной функцией потерь $L(y',y)== (y' - y)^2$ Доказать, что если $f^*(x) = E(Y|X = x)$ (регрессионная функция). Чему тогда равен средний риск $R(f^*)$?

Экзаменационный билет №24

1. Алгоритм DBSCAN.

2. Понятие о глубоком обучении. Сверточные нейронные сети.

3. Рассмотрим задачу восстановления регрессии с функцией потерь $L(y', y) = |y' - y|$. Доказать, что минимум среднему риску доставляет при этом условная медиана $f(x) = \text{median}(Y|X = x)$.

Экзаменационный билет №25

1. Алгоритмы иерархической кластеризации.

2. Машина опорных векторов. Формулировка задачи в виде задачи математического программирования. Двойственная задача. (Случай линейно разделимых и неразделимых классов)

3. Пусть в задаче классификации с двумя классами $\{0,1\}$ используется функция потерь $L(y', y)$, такая, что $L(0,0) = L(1,1)=0$, $L(1,0) = l_1$, $L(0,1) = l_0$. Докажите, что в этом случае байесов классификатор $f^*(x)$ удовлетворяет условию

Экзаменационный билет №26

1. Деревья решений. Алгоритм CART.

2. Вероятностная постановка задачи обучения с учителем. Функция потерь. Средний (ожидаемый) риск. Эмпирический риск. Регрессионная функция и байесов классификатор. Неустранимая (байесовская) ошибка.

3. Выразить байесов классификатор $f^*(x)$ для классификации с K классами, если функция потерь равна $L(y', y) = \ell_{y'y}$ $(y', y = 1, 2, \ldots, K)$.

Экзаменационный билет №27

1. Случайный лес. Экстремально случайные деревья.

2. Ошибки 1-го и 2-го рода. Чувствительность, специфичность, точность, полнота. ROC-кривая. Площадь под ROC-кривой.

3. Пусть N точек распределены случайно равномерно в единичной d-мерной гиперсфере. Доказать, что медианное расстояние от центра сферы до ближайшей точки равно

$$\rho(d, N) = \sqrt[d]{1 - \sqrt[N]{1/2}}.$$

Найти предел $\rho(d, N)$ при $d \to \infty$, $N = O(d)$. Какой вывод из этого можно сделать применительно к методу ближайшего соседа при больших d?

Экзаменационный билет №28

1. Принцип минимизации эмпирического риска. Минимизация отступа. Регуляризация.

2. Алгоритм XGBoost.

3. Bias-variane trade-off. Рассмотрим задачу восстановления зависимости $Y = f*(X)$, где X - случайная величина, а f* - неизвестная детерминированная функция. Пусть $x^{(1)}, x^{(2)}, \ldots, x^{(N)}$ суть независимые реализации величины X. В качестве модельной зависимости возьмем функцию f(x, D), где D = x (1), x(2), . . . , x(N). Разложить $E_D f(x, D) - f^*(x)^2$ в сумму квадрата математического ожидания смещения (bias) и дисперсии (variance).

Экзаменационный билет №29

1. Принцип минимизации эмпирического риска. Минимизация отступа. Регуляризация.

2. Градиентный бустинг деревьев решений.

3. Может ли использование коррелированных переменных улучшить качество предсказания? Рассмотрим задачу классификации с двумя классами. Пусть пространство признаков двумерное. Объекты каждого класса имеют нормальное распределение с математическим ожиданием (-1,-1) и (1,1) соответственно и единичной матрицей коварации каждый. Априорные вероятности классов равны ½.

1) Вычислите коэффициент корреляции для переменных x_1, x_2.

2) Найти байесов классификатор и вычислить байесову ошибку для усеченной задачи, рассматривая только одну переменную x_1.

3) Найти байесов классификатор и вычислить байесову ошибку для исходной задачи.

4)Приводит ли использование второй переменной к уменьшению ошибки?

Экзаменационный билет №30

1. Экспериментальная (эмпирическая) оценка качества обучения. Обучающая, проверочная и тестовая выборки. Метод перекрестного (скользящего) контроля.

2. Дилемма "Смещение-разброс". Кривая обучения.

3. Рассмотрим задачу классификации с двумя классами 0 и 1. Пусть пространство признаков двумерное. Объекты каждого класса имеют нормальное распределение с математическим ожиданием (0,0) и (1,1) соответственно и матрицей ковариации

$$\mathbf{\Sigma} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Априорные вероятности классов равны $\Pr\{Y = 0\} = 1/3$ и $\Pr\{Y = 1\} = 2/3$.

1) Найти уравнение разделяющей поверхности байесова классификатора.

2) Найти собственное разложение матрицы Σ.

3) Перейти к новым координатам, оси которых совпадают с собственными векторами матрицы Σ.

4) Выписать уравнение разделяющей поверхности байесова классификатора в новых координатах.