Приложение 2 к РПД
Анализ текстовых данных
01.03.02 Прикладная математика и информатика
Направленность (профиль)
Управление данными и машинное обучение
Форма обучения – очная
Год набора – 2021

МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ (МОДУЛЯ)

1.	Кафедра	Математики, физики и информационных технологий
2.	Направление подготовки	01.03.02 Прикладная математика и информатика
3.	Направленность (профиль)	Управление данными и машинное обучение
4.	Дисциплина (модуль)	Б1.В.01.05 Анализ текстовых данных
5.	Форма обучения	Очная
6.	Год набора	2021

1. Перечень компетенций

- **ПК-1:** Способен собирать, обрабатывать и интерпретировать данные современных научных исследований, необходимые для формирования выводов по соответствующим прикладным исследованиям
- **ПК-2:** Способен работать в составе научно-исследовательского и производственного коллектива и решать задачи профессиональной деятельности
- **ПК-3:** Способен к разработке и применению алгоритмических и программных решений в области системного и прикладного программного обеспечения

2. Критерии и показатели оценивания компетенций на различных этапах их формирования

Этапы формирования компетенций (разделы, темы дисциплины)	Формируемая компетенция	Критерии и показатели оценивания компетенций			Формы контроля сформированности компетенций
		Знать:	**Уметь:**	**Владеть:**	
Раздел 1. Введение в анализ текстовых данных	ПК-1, ПК-2, ПК-3	• основные способы получения и обработки информации, необходимой для профессиональной деятельности; • основные парадигмы машинного обучения	• применять методы машинного обучения для решения задач анализа текстовых данных; • оценивать качество моделей машинного обучения	• навыком исследования и моделирования предметной области; • владеть терминологией машинного обучения; • владеть инструментальными средствами для построения моделей машинного обучения с учителем	Выполнение лабораторных работ 1-3 Тестирование по темам дисциплины
Раздел 2. Задачи текстового анализа	ПК-1, ПК-2, ПК-3	• модели и методы предобработки текстовых данных; • методы оценки качества моделей машинного обучения	• обрабатывать и анализировать результаты эксперимента, проводить расчеты по экспериментальным данным с использованием компьютерных программ	• навыками работы с наиболее распространенными прикладными пакетами для математической обработки данных; • основными методами, способами и средствами получения, хранения, переработки информации	Выполнение лабораторных работ 4-6 Тестирование по темам дисциплины

Шкала оценивания в рамках балльно-рейтинговой системы

«неудовлетворительно» – 60 баллов и менее; «удовлетворительно» – 61-80 баллов; «хорошо» – 81-90 баллов; «отлично» – 91-100 баллов

3. Критерии и шкалы оценивания

4.1. Критерии оценки выполнения лабораторных работы

1. 8 баллов выставляется, если студент вовремя и полностью выполнил задание на лабораторную работу, правильно и полностью описал и изложил необходимые результаты в отчете, аргументировав их на защите лабораторной работы.
2. 6 балла выставляется, если студент выполнил задание на лабораторную работу, правильно описал и изложил необходимые результаты в отчете, аргументировав их на защите лабораторной работы, но задержал сдачу работы на одну неделю.
3. 4 балла выставляется, если студент выполнил задание на лабораторную работу, правильно описал и изложил необходимые результаты в отчете, аргументировав их на защите лабораторной работы, но задержал сдачу работы на две недели.
4. 2 балла выставляется, если студент выполнил задание на лабораторную работу, описал и изложил необходимые результаты в отчете, аргументировав их на защите лабораторной работы, но задержал сдачу работы более чем три недели.
5. 0 баллов - если студент не выполнил задания и/или предоставил отчет.

4.2. Тестирование по темам дисциплины

Процент правильных ответов	До 60	61-80	81-100
Количество баллов за решенный тест	0	3	6

4.3. Критерии оценки выступление с презентацией (доклад, реферат)

Характеристика выступления с презентацией	количество баллов
Содержание	
Сформулирована цель работы	0,5
Понятны задачи и ход работы	0,5
Информация изложена полно и четко	0,5
Иллюстрации усиливают эффект восприятия текстовой части информации	0,5
Сделаны выводы	0,5
Оформление презентации	
Единый стиль оформления	0,5
Текст легко читается, фон сочетается с текстом и графикой	0,5
Все параметры шрифта хорошо подобраны, размер шрифта оптимальный и одинаковый на всех слайдах	0,5
Ключевые слова в тексте выделены	0,5
Эффект презентации	
Общее впечатление от просмотра презентации	0,5
Max количество баллов	5

4.4. Критерии оценки разработки и защиты проекта

Характеристики работы студента	количество баллов
- студент глубоко и всесторонне усвоил проблему; - уверенно, логично, последовательно и грамотно его излагает; - опираясь на знания основной и дополнительной литературы, тесно привязывает усвоенные научные положения с лабораторной деятельностью; - умело обосновывает и аргументирует выдвигаемые им идеи;	10

- делает выводы и обобщения; - свободно владеет понятиями	
- студент твердо усвоил тему, грамотно и по существу излагает ее, опираясь на знания основной литературы; - не допускает существенных неточностей; - увязывает усвоенные знания с лабораторной деятельностью; - аргументирует научные положения; - делает выводы и обобщения; - владеет системой основных понятий	7
- тема раскрыта недостаточно четко и полно, то есть студент освоил проблему, по существу излагает ее, опираясь на знания только основной литературы; - допускает несущественные ошибки и неточности; - испытывает затруднения в лабораторном применении знаний; - слабо аргументирует научные положения; - затрудняется в формулировании выводов и обобщений; - частично владеет системой понятий	3
- студент не усвоил значительной части проблемы; - допускает существенные ошибки и неточности при рассмотрении ее; - испытывает трудности в лабораторном применении знаний; - не может аргументировать научные положения; - не формулирует выводов и обобщений; - не владеет понятийным аппаратом	0

5. Типовые контрольные задания и методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций в процессе освоения образовательной программы

5.1 Типовое контрольное тестовое задание:

1) Необработанный материал, предоставляемый источником и используемый потребителями для формирования на его основе полезного результата:

a) информация,
b) данные,
c) знания.

2) Совокупность фактов, закономерностей и эвристических правил, с помощью которых решается поставленная задача

a) информация,
b) данные,
c) знания.

3) Свойство информации, характеризующее степень ее соответствия настоящему моменту времени

a) полнота
b) достоверность
c) ценность
d) адекватность
e) актуальность
f) доступность

4) Для многомерных данных, предполагающих учёт пространственных координат, признаковых переменных и времени часто используется термин «… данных»:

a) гиперквадрат,
b) гиперкуб,
c) гиперсфера,
d) гипербола.

5) Что из перечисленного не является машинным обучением?

a) обучение по прецедентам,
b) обучение с учителем,
c) обучение без учителя,
d) глубокое обучение,
e) обучение с подкреплением,
f) обучение по контрпримерам.

6) Наиболее распространённым плотностным методом кластеризации является…

a) метод К-средних,
b) метод EM,
c) метод DBSCAN,
d) метод Кохонена.

7) Какой вид анализа многомерных данных не позволяет перейти к пространству меньшей размерности?

a) корреляционный анализ,
b) анализ главных компонент,
c) факторный анализ,
d) анализ с использованием карт Кохонена.

8) Какой вид анализа временного ряда использует понятие «мгновенная частота»?

a) Фурье-анализ,
b) вейвлет-анализ,
c) сингулярный спектральный анализ,
d) анализ эмпирических мод Хуанга.

9) Какая из перечисленных логических функций двух переменных не может быть смоделирована нейроном МакКаллока-Питтса?

a) OR,
b) AND,
c) XOR,
d) все перечисленные могут быть смоделированы.

10) Какие математические объекты могут быть найдены с помощью аффинитивного анализа?

a) ассоциативные правила,
b) кластеры наибольшей мощности,
c) главные компоненты,
d) непериодическая последовательность максимальной длины.

Ключ: 1 - b 2 - c 3 - e 4 –b 5 - f 6 - c 7 - a 8 - d 9 - c 10 - a

5.2 Примерные темы докладов:

1. Методы сбора данных.
2. Способы хранения данных.
3. Технология OLAP.
4. Данные, представление данных.
5. Информация и знание.
6. Виды и способы измерений.
7. Виды и способы использования шкал.
8. Свойства информации.
9. Свойства знаний.
10. Примеры использования аффинитивного анализа.
11. Примеры использования кластерного анализа.
12. Примеры использования методов классификации.
13. Прикладные задачи классификации.
14. Примеры использования метода построения деревьев решений.
15. Области применения двухслойных персептронов.
16. Глубокое обучение в задачах компьютерного зрения.
17. Глубокое обучение в задачах распознавания речи.
18. Глубокое обучение в задачах анализа текста на естественном языке.
19. Глубокое обучение в задачах прогнозирования валютных и социально-экономических показателей.
20. Задачи анализа регулярных пространственно распределённых данных и методы их решения.

5.3 Вопросы к зачету:

1. Анализ данных: понятия «информация», «данные», «знания».
2. Виды шкал данных. Примеры процедур шкалирования данных предметной области.
3. Представление пространственно-временных данных: гиперкуб данных; суть многомерности данных; временные ряды, карты пространственных распределений, векторы признаков и состояний.
4. Свойства временных рядов: аксиоматические и проверяемые. Многомерный временной ряд гридированных данных в прямоугольной области: формальное описание, примеры.
5. Задачи исследования временных рядов.
6. Аддитивная модель временного ряда. Декомпозиция одномерного временного ряда на аддитивные составляющие: Фурье-анализ и вейвлет-анализ.
7. Аддитивная модель временного ряда. Декомпозиция временного ряда (одномерный и многомерный случаи) на аддитивные составляющие: сингулярный спектральный анализ и декомпозиция на эмпирические моды Хуанга.
8. Прогнозирование временных рядов по результатам сингулярного спектрального анализа (метод «Гусеница»).
9. Статистические модели одномерных временных рядов: AR, MA, ARIMA, ARX, TARX, GARCH.
10. Представление данных векторами линейного пространства признаков. Основные понятия Data Science: open data, big data, data mining, machine learning, supervised learning, unsupervised learning, pattern recognition, text mining.
11. Задача классификации: дерево решений.
12. Задача кластеризации: формулировка, основные понятия.
13. Метод кластеризации К-средних и его модификации.
14. Метод кластеризации Expectation-Maximization. Формула Байеса. Расчёт вероятности для различных законов распределения (Бернулли, биномиального, нормального).
15. Использование нейронных сетей Кохонена для распознавания образов.
16. Иерархические агломеративные и дивизивные методы кластеризации. Дендрограмма.
17. Методы кластеризации на основе нейронных сетей Кохонена: слоя, простой прямоугольной карты, растущей иерархической карты. Способы инициализации весов.

18. Метод кластеризации DBSCAN.
19. Краткая сравнительная характеристика методов и моделей, используемых для кластеризации: К-средние, Expectation-Maximization, карты Кохонена, DBSCAN.
20. Анализ главных компонент: эквивалентные формулировки задачи, процедура формирования векторов нового базиса, приложения. Подготовка матрицы данных: центрирование, нормирование, стандартизация. Матрицы нагрузок и счётов.
21. Связь метода главных компонент с корреляционным анализом, задачей о собственных числах и собственных векторах матрицы, сингулярным спектральным анализом, картами Кохонена и многослойным персептроном.
22. Общие черты и отличия корреляционного, регрессионного, дискриминантного, факторного и дисперсионного анализов.
23. Методы анализа текстовой информации – Text Mining.