ПЕРМСКИЙ ГОСУДАРСТВЕННЫЙ НАЦИОНАЛЬНЫЙ

ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Фонды оценочных средств по дисциплине «Компьютерные технологии обработки больших массивов данных»

Направление подготовки 01.04.02 Прикладная математика и информатика

1. Формируемые дисциплиной компетенции

ОПК.4.1
Комбинирует и адаптирует современные информационно-коммуникационные технологии для реализации решения математических задач

ОПК.2.3
Реализует математический метод на языке программирования высокого уровня и/или с помощью специализированных пакетов программ

ПК.22.1 Применяет методы объяснимого искусственного интеллекта для построения объяснимой модели интеллектуальной системы

ПК.22.2 Применяет методы объяснимого искусственного интеллекта для построения объясняющего интерфейса интеллектуальной системы

ПК.22.3 Применяет и разрабатывает стандарты в области объяснимого искусственного интеллекта

2. Спецификация теста

Тест по дисциплине «Компьютерные технологии обработки больших массивов данных» состоит из 20 заданий. Рекомендованное время решения теста испытуемым – 90 минут. Максимальный балл за верное выполнение всех заданий теста – 30 баллов. Минимальный проходной балл – 12, что соответствует минимальному порогу для выставления отметки «удовлетворительно».

Схема конвертации баллов в отметки:

0-12 баллов – «неудовлетворительно»

13-18 баллов – «удовлетворительно»

19-24 баллов – «хорошо»

25-30 баллов – «отлично»

Структура теста:

Наименование раздела/темы	Планируемый результат	Количество заданий в тесте
Основы программирования на Python	Знает и способен использовать технологии обработки текстовых и числовых данных большого объема, может выбрать подходящую структуру данных и запрограммировать алгоритм обработки.	6
Основы машинного обучения	Умеет использовать библиотеку машинного обучения Skleaarn, способен работать с документацией и может применять методы машинного обучения для решения задач анализа данных.	10
Решение практических задач обработки данных	Способен решать практическую задачу обработки данных, используя изученные технологии и интерпретировать полученные результаты.	4

Тест по дисциплине «Компьютерные технологии обработки больших массивов данных**», вариант 1.**

1. Вам дан набор из 10.000 писем, отправленных одним и тем же человеком, и требуется сгруппировать их так, чтобы в одной группе оказались письма на схожие темы — например, личная переписка, письма с авиабилетами и т.д. Что это за задача? (1 балл)

а) Регрессия

б) Классификация

в) Кластеризация

2. Какая из этих фраз наиболее точно описывает переобучение? (1 балл)

а) Переобучение — это ситуация, в которой алгоритм показывает одинаково плохое качество и на обучающей выборке, и на новых данных

б) Переобучение — это ситуация, в которой алгоритм часто отказывается от построения прогноза на новых данных.

в) Переобучение — это ситуация, в которой алгоритм выдает недетерминированные ответы на новых данных (то есть при разных запусках на одном и том же объекте можно получить разные предсказания)

г) Переобучение — это ситуация, в которой алгоритм показывает хорошее качество на обучающей выборке, но при этом плохо работает на новых данных

3. Что из этого — корректные названия типов признаков? Правильных ответов может быть несколько. (1 балл)

а) Устойчивые признаки

б) Номинальные (категориальные) признаки

в) Бинарные признаки

г) Нетривиальные признаки

д) Числовые (количественные) признаки

4. Какие типы данных есть в Python? Правильных ответов может быть несколько. (1 балл)

а) int

б) double

в) array

г) str

д) bool

5. Какие коллекции есть в Python? Правильных ответов может быть несколько. (1 балл)

а) dict

б) map

в) tuple

г) matrix

д) list

6. Выберите изменяемые типы Python. Правильных ответов может быть несколько. (1 балл)

а) list

б) set

в) str

г) tuple

д) dict

7. Какие задачи являются задачами классификации? Правильных ответов может быть несколько. (1 балл)

а) Зная характеристики клиента банка, определить, вернет ли клиент кредит.

б) По данным о фильме определить возрастную категорию пользователей (дети, подростки, взрослые, пенсионеры), которым понравится этот фильм.

в) По данным об автомобиле определить максимальную цену, за которую его удастся продать.

г) Выстроить список подписчиков сайта в порядке от тех, кому, скорее всего понравится новость, к тем, кому она, скорее всего не понравится

8. Рассмотрим признак "Число обращений клиента в службу поддержки банка". Он принимает только целые неотрицательные значения. Какой тип имеет данный признак? (1 балл)

а) Вещественный — ведь целые числа тоже являются вещественными.

б) Категориальный — он ведь может принимать значения лишь из конечного множества.

в) Бинарный — почему бы нет.

9. Выберите признаки, которые могут рассматриваться только как категориальные (и не могут рассматриваться как бинарные или вещественные). Правильных ответов может быть несколько. (1 балл)

а) Тип дома — кирпичный, блочный, панельный и т.д.

б) Цвет автомобиля

в) Наличие у клиента банка военного билета

г) Год рождения

д) Тарифный план клиента мобильного оператора

10. Выберите верные утверждения про машинное обучение. Правильных ответов может быть несколько. (1 балл)

а) В задачах обучения с учителем требуется построить алгоритм, который по признаковому описанию объекта предсказывает некоторый ответ.

б) Алгоритм (или модель) — это функция, которая принимает на вход признаковое описание объекта и выдает некоторое предсказание ответа

в) Функционал ошибки позволяет определить, насколько данный алгоритм подходит для решения задачи на конкретной выборке.

г) Алгоритм (или модель) — это функция, которая принимает на вход признаковое описание объекта и его ответ, и выдаёт качество предсказаний для данного объекта.

д) Функционал ошибки определяет уровень шума в признаках.

11. Выберите метрики качества, которые можно использовать при решении задачи классификации. (1 балл)

а) коэффициент детерминации

б) F-мера

в) площадь под ROC-кривой

г) доля правильных ответов

д) площадь по PR-кривой

12. Что вычисляется по формуле $M_i = y_i\langle w, x_i\rangle$ (1 балл)

а) отступ в задачах классификации

б) величина ошибки в задачах регрессии

в) вероятность принадлежности объекта к заданному классу в листе решающего дерева

г) ответ случайного леса на данном объекте

13. Магазин вел статистику своих продаж смартфонов в течение года. В базе имеется информация о количестве работавших сотрудников, средней, максимальной и минимальной стоимости смартфона и количестве единиц продукции, имеющейся в наличии, количестве проданных за день смартфонов и объеме дневной выручки. Требуется спрогнозировать выручку магазина на ближайший месяц. Какую задачу машинного обучения придется решать? (1 балл)

а) задача реализации метода наименьших квадратов

б) задача кластеризации

в) задача многоклассовой классификации

г) задача поиска аномалий

д) задача регрессии

14. Какие функции numpy можно использовать для формирования числовой последовательности. Например, нужно сформировать последовательность всех четных чисел из промежутка [1; 100]. Правильных ответов может быть несколько. (1 балл)

а) linspace

б) arange

в) hstack

г) array

15. Дан кортеж A. Как можно сформировать список B из тех же значений? Правильных ответов может быть несколько. (1 балл)

а) B = list(A)

б) B = tuple(A)

в) B = [elem for elem in A]

г) B = (elem for elem in A)

д) B = [A[i] for i in range(len(A))]

16. Сколько функций будет на графике, построенном командой

import matplotlib.pyplot as plt

plt.plot(x, x, '-', x, 2*x, '--', x, 3*x, ':', x, 4*x, '-.')

(1 балл)

Ответ: _____

17. Дана строка, в которой перечислены целые числа - оценки студентов, разделенные пробелами. Напишите как можно более короткий код для определения среднего балла по данным оценкам. (3 балла)

Ответ: _____

_____

18. Дана строка с произвольным текстом. Напишите код на Python, который позволит построить таблицу с частотой встречаемости каждого из знаков препинания, присутствующего в данной строке. Для решения задачи обязательно использовать словарь. (4 балла)

Ответ: _____

19. Даны две строки с произвольным текстом на русском языке. Напишите код на Python, который позволит определить схожесть текстов на основе метрики косинусного расстояния.

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

Косинусное расстояние – это мера похожести двух векторов. Векторами считать частоты встречаемости слов в текстах. (5 баллов)

Ответ: _____

20. Сформулируйте что такое «переобучение» в машинном обучении и опишите в общем виде, как можно идентифицировать переобученность модели. (2 балла)

Ответ: _____

_____

_____

_____

_____

_____

Тест по дисциплине «Компьютерные технологии обработки больших массивов данных**», вариант 2.**

1. Вам нужно предсказать, каким завтра будет курс доллара. Какая это задача? (1 балл)

а) Регрессия

б) Классификация

в) Кластеризация

2. Какие из этих задач являются задачами классификации? Правильных ответов может быть несколько. (1 балл)

а) Разделение книг, хранящихся в электронной библиотеке, на научные и художественные

б) Поиск групп похожих пользователей интернет-магазина

в) Прогноз оценки студента по пятибалльной шкале на экзамене по машинному обучению в следующей сессии

г) Прогноз температуры на следующий день

3. Выберите верные утверждения. Правильных ответов может быть несколько. (1 балл)

а) Одна из задач машинного обучения — научиться делать прогнозы для объектов

б) Одна из задач машинного обучения — научиться делать прогнозы для признаков

в) Признаки описываются с помощью объектов

г) Объекты описываются с помощью признаков

4. Какие типы данных есть в Python? Правильных ответов может быть несколько. (1 балл)

а) float

б) NoneType

в) char

г) integer

д) bool

5. Какие коллекции есть в Python? Правильных ответов может быть несколько. (1 балл)

а) unordered_set

б) set

в) array

г) matrix

д) tuple

6. Объекты каких типов могут быть ключом в словаре? Правильных ответов может быть несколько. (1 балл)

а) int

б) str

в) bool

г) list

д) tuple

7. Какие задачи относятся к категории "обучение с учителем"? Правильных ответов может быть несколько. (1 балл)

а) классификация

б) кластеризация

в) поиск аномалий

г) регрессия

д) ранжирование

8. Выберите верные утверждения про признаки. Правильных ответов может быть несколько. (1 балл)

а) Признаки не используются в задачах кластеризации и поиска аномалий.

б) Признаки описывают объекты в формате, с которым легко работать на компьютере.

в) Признаки задаются только вещественными числами.

г) Предсказания для объектов делаются на основе значений признаков.

д) Набор значений признаков на объекте представляет собой вектор определённой размерности.

9. Выберите вещественные признаки из списка. Правильных ответов может быть несколько. (1 балл)

а) Количество детей

б) Город, в котором прописан клиент

в) Наличие у клиента банка военного билета

г) Температура воздуха

д) Год рождения

10. В чем смысл регуляризации? Правильных ответов может быть несколько. (1 балл)

а) Штрафовать за большие веса в линейной модели

б) Бороться с переобучением модели

в) Приводить все данные в нормальную форму путем, чтобы они принадлежали промежутку [0; 1]

г) Преобразовывать категориальные данные в несколько бинарных столбцов со значениями 0 и 1

11. Что является недостатком одного решающего дерева. Правильных ответов может быть несколько. (1 балл)

а) Склонность к переобучению

б) Большое время построения

в) Сложность реализации

г) Требует большого объема обучающей выборки

д) Сильная изменчивость при малейшем изменении исходных данных

12. Выберите гиперпараметры модели. Правильных ответов может быть несколько. (1 балл)

а) коэффициент регуляризации

б) размер обучающей выборки

в) количество частей разбиения выборки при кросс-валидации

г) характеристики спрямляющего пространства

д) коэффициенты при неизвестных в аналитическом виде целевой функции

13. Штатный психолог в школе заболел. Для заполнения отчета требуется срочно всех учеников школы разбить на две группы по психологическому типу: интроверты и экстраверты. Для решения задачи будут использованы данные психологического тестирования и профили активностей учеников в социальных сетях. Какая задача машинного обучения здесь будет решаться? (1 балл)

а) задача кластеризации

б) задача классификации

в) Задача ранжирования

г) Задача регрессии

д) Задача визуализации

14. Какие алгоритмы из библиотеки sklearn можно использовать для решения задач восстановления регрессии. Правильных ответов может быть несколько. (1 балл)

а) Ridge

б) Lasso

в) LinearRegression

г) SGDRegressor

д) MakeRegression

15. Дан список A с координатами точек (x, y) в двумерном пространстве. Каждая точка хранится кортежем. Как отсортировать эти точки по ключу

x (убыв.) + y (убыв.)

Правильных ответов может быть несколько. (1 балл)

а) A.sort()

A.reverse()

б) A.sort().reverse()

в) A.sort(key=lambda elem: (-elem[0], -elem[1]))

г) A.sort(key=lambda elem: (elem[0], elem[1]), reverse=True)

д) A.sort(reverse=True)

16. Сколько функций будет на графике, построенном командой

```
import matplotlib.pyplot as plt
plt.plot(x, 2*x, '--', x, 3*x, ':', x, 4*x, '-.')
```

(1 балл)

Ответ: _____

17. Дана строка, в которой перечислены целые числа – температура воздуха, в каждый из дней месяца, разделенные запятыми. Напишите как можно более короткий код для определения минимальной температуры. (3 балла)

Ответ: _____

_____

18. Дана строка с произвольным текстом на русском языке. Напишите код на Python, который позволит построить таблицу с частотой встречаемости гласных букв, присутствующих в данной строке. Для решения задачи обязательно использовать словарь. (4 балла)

Ответ: _____

19. Даны две строки с произвольным текстом на русском языке. Напишите код на Python, который позволит определить схожесть текстов на основе метрики косинусного расстояния.

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

Косинусное расстояние – это мера похожести двух векторов. Векторами считать частоты встречаемости слов в текстах. (5 баллов)

Ответ: _____

20. Сформулируйте что такое «выброс» и опишите в общем виде, как можно идентифицировать выбросы в данных.(2 балла)

Ответ: _____

_____

_____

_____

_____

_____

Ключ к тесту

Вариант 1	Вариант 2
1. в	1. а
2. г	2. а,в
3. б,в,д	3. а, г
4. а, г,д	4. а, б, д
5. а, в, д	5. б, д
6. а, б, д	6. а, б, в, д
7. а, б	7. а, г, д
8. а	8. б, г, д
9. а, б, д	9. а, г, д
10. а, б, в	10. а, б
11. б, в, г, д	11. а, д
12. а	12. а, б, в, г
13. д	13. а
14. а, б	14. а, б, в, г
15. а, в, д	15. а, в, г, д
16. 4	16. 3