

Программа вступительного испытания по предмету
«Искусственный интеллект и наука о данных»
для поступающих на обучение по образовательной программе магистратуры
«Искусственный интеллект и наука о данных»
направления подготовки
09.04.03 «Прикладная информатика»

Раздел 1. Организационно-методический раздел

Длительность вступительного испытания

180 минут (3 астрономических часа).

Форма вступительного испытания

Вступительный экзамен проводится в письменной форме без использования каких-либо информационных источников во время экзамена.

В зависимости от порядка проведения испытания, задание выполняется в электронной форме, либо в простой письменной (от руки) форме.

В случае выполнения в электронной форме с использованием программных средств, не предоставляющих возможности ввода формул и форматированного текста, математические формулы рекомендуется вводить в формате (La)TeX.

В случае выполнения экзамена в простой письменной форме, ответы должны быть написаны аккуратно и разборчиво, затем отсканированы или сфотографированы для передачи проверяющим в качестве, не допускающем неоднозначного прочтения написанного.

Структура и содержание вступительного испытания

Экзаменационное задание включает два раздела.

1. **Научно-исследовательская часть.** Абитуриент представляет в обезличенной (не допускающей установление его/её личности) форме аннотацию своей выпускной работы бакалавра или специалиста с раскрытием актуальности решаемой проблемы, указанием применяемых технологий, перечислением полученных результатов работы.
2. **Комбинированный тест.** Абитуриент получает 10 заданий, содержащие 5 заданий в тестовой форме и 5 заданий, требующих развернутого ответа, аналогичным приведённым примерам в разделе 3, сформированным на основе тем для подготовки, представленных в разделе 2 (см. Разделы 2, 3). Во время решения допускается использовать чистые листы бумаги и письменные принадлежности и/или пустым.

Рекомендуемый объём: по 1 части — 300–600 слов.

Критерии оценивания вступительного испытания

Подсчет итоговой оценки за вступительное испытание осуществляется путем суммирования баллов, выставленных за ответы по каждому разделу. Вступительное испытание оценивается по шкале от 0 до 100 баллов.

Научно-исследовательская часть

Критерий	Максимальный балл
Есть четко сформулированная решаемая проблема, указана её актуальность	15
Указаны используемые методы и технологии, релевантные решению обозначенной проблемы	15
Представлены результаты работы, способствующие решению обозначенной проблемы	15
Соответствие тематики и результатов НИР тематике выбранной образовательной программы	5

Итого: от 0 до 50 баллов по данной части.

Задания комбинированного теста

Ответом на тестовые задания является выбор одного или нескольких вариантов из предложенных или число, или слово. Задания, подразумевающие развернутое решение, требуют более детального ответа, который может включать написание программного кода, расчеты или разъяснение хода мыслей. Задания оцениваются отдельно следующим образом:

Качество ответа	Количество баллов
Корректный ответ тестового задания	4
Корректный ответ задания, требующего развернутого ответа	6
Частично корректный ответ задания, требующего развернутого ответа (оценивается пропорционально степени полноты решения)	от 2 до 6
Неправильный ответ	0

Итого: от 0 до 50 баллов по данной части.

Раздел 2. Перечень и содержание тем для подготовки

Тема 1. Линейная алгебра и математический анализ

1. Векторы и матрицы: операции над векторами и матрицами, умножение матриц, транспонирование и обратные матрицы.
2. Линейные операции и преобразования: системы линейных уравнений, определители и их свойства, собственные значения и собственные векторы.
3. Пространства и подпространства: линейные комбинации и линейная зависимость, ранг матриц.
4. Линейные операторы: определение и свойства линейных операторов, матричное представление линейных операторов.
5. Дифференцирование: производные и их свойства, правила дифференцирования, частные производные и градиент.

Тема 2. Теория вероятностей и математическая статистика

1. Основные понятия вероятности: элементарные события и пространство элементарных событий, вероятностные меры.
2. Условная вероятность и независимость событий.
3. Случайные величины: определение случайных величин, функции распределения и плотности вероятности.
4. Числовые характеристики случайных величин. Начальный и центральный моменты. Математическое ожидание, его свойства. Дисперсия случайной величины, ее свойства.
5. Элементы корреляционного анализа. Парная линейная регрессия.
6. Описательные статистики. Проверка статистических гипотез.

Тема 3. Основы программирования на Python

1. Основы языка Python: синтаксис Python, переменные и типы данных, операторы и выражения.
2. Структуры данных: списки, массивы, кортежи, множества, словари. Работа с коллекциями данных.
3. Условные операторы if, elif, else. Циклы for и while. Функции и их определение. Обработка исключений.
4. Работа с файлами и использование модулей.
5. Основы объектно-ориентированного программирования.
6. Назначение и области применения некоторых библиотек: NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn, TensorFlow и PyTorch.

Тема 4. Базы данных

1. Понятие БД и СУБД. Основные функции СУБД. Основные понятия реляционной и объектной моделей данных.
2. Принципы, методы и подходы к проектированию баз данных. Различные архитектуры баз данных и их применение. Основные шаги в процессе проектирования.
3. Язык SQL и его возможности. Средства определения и изменения схемы БД. Определение ограничений целостности. Контроль доступа и средства манипулирования данными.

Тема 5. Вопросы разработки ПО

1. Процесс разработки ПО и организация работы команды. Водопадная модель, V-модель, инкрементальная, спиральная модели.
2. Гибкие методологии разработки. Agile-манифест. Обзор современных гибких методологий.
3. Управление проектом, понятие проекта. Сбор требований, создание плана проекта.
4. Тестирование и обеспечение качества. Типы тестирования: модульное, интеграционное, системное и др. Автоматизированное тестирование. Практики обеспечения качества кода.
5. Системы управления версиями. Централизованное и распределённое управление версиями. Git Workflow, GitHub Workflow.

Тема 6. Основы машинного обучения

1. Типы задач машинного обучения: классификация, регрессия, кластеризация, снижение размерности.
2. Основные алгоритмы машинного обучения. Принципы их работы. Линейная регрессия. Метод k ближайших соседей (k-NN). Метод опорных векторов (SVM). Решающие деревья и случайный лес. Алгоритм кластеризации k-средних.
3. Метрики оценки качества моделей: Accuracy, F1-мера, MSE, RMSE, AUC-ROC.

Раздел 4. Примеры заданий

Задание № 1 (требуется развернутого ответа)

Найдите собственное число матрицы:

$$A = \begin{bmatrix} 9 & 8 \\ -6 & -5 \end{bmatrix}, \quad \vec{x} = \begin{bmatrix} -4 \\ 3 \end{bmatrix}, \quad \lambda = ?$$

Задание № 2 (требуется развернутого ответа)

Если у вас есть два независимых события A и B, каждое из которых имеет вероятность 0.2, какова вероятность, что одно из этих событий произойдет первым?

Задание № 3 (тестовое)

Какое из следующих утверждений неверно относительно коэффициента корреляции?

1. Коэффициент корреляции Пирсона всегда находится в интервале от -1 до 1.
2. Чем ближе коэффициент корреляции к 1, тем сильнее линейная зависимость между переменными.
3. Корреляция всегда указывает на причинно-следственную связь между переменными.
4. Если коэффициент корреляции равен 0, это означает, что нет никакой связи между переменными.

Задание № 4 (требуется развернутого ответа)

Вам даны два набора данных:

Набор 1: $X = [10, 15, 20, 25, 30]$

Набор 2: $Y = [8, 18, 28, 38, 48]$

1. Постройте линейную регрессионную модель, используя X в качестве независимой переменной и Y в качестве зависимой. Укажите уравнение полученной регрессионной прямой.
2. Рассчитайте коэффициент детерминации R^2 и объясните, что он показывает в контексте данной задачи.
3. Проанализируйте остатки модели и сделайте вывод о её качестве.

Задание № 5 (требуется развернутого ответа)

Напишите программу на Python, которая:

1. Считывает с клавиатуры текстовый файл.

2. Подсчитывает количество уникальных слов в тексте, игнорируя регистр и знаки препинания.
3. Создает словарь, где ключом является слово, а значением — количество его вхождений в текст.
4. Выводит 10 наиболее часто встречающихся слов с их частотой в порядке убывания.
5. Анализирует текст и выводит распределение слов по длине (например, сколько слов длиной 1 символ, 2 символа и т.д.).

Пример вывода:

Уникальные слова: 150

Частотный словарь (top-10): {'Python': 25, 'data': 20, ...}

Распределение по длине: {1: 5, 2: 20, 3: 50, ...}

Задание № 6 (требуется развернутого ответа)

У вас есть набор данных для задачи классификации, содержащий 1000 записей с двумя классами (0 и 1). Данные разделены на обучающую (80%) и тестовую (20%) выборки.

- Опишите процесс построения классификационной модели с использованием метода "решающее дерево".
- Укажите, какие гиперпараметры могут влиять на качество модели, и как их можно подбирать.
- Опишите метрики для оценки качества классификатора и предложите подходящий метод для выбора лучшей модели.
- Предложите способ предотвращения переобучения данной модели.

Задание № 7 (требуется развернутого ответа)

Вам даны два фактических значения: $Y_{true} = [10, 15]$ и два предсказанных $Y_{pred} = [8, 12]$. Чему будет равна среднеквадратичная ошибка (MSE)?

Задание № 8 (требуется развернутого ответа)

Какие из перечисленных Python-библиотек содержат алгоритмы машинного обучения и обработки данных? Выберите все подходящие варианты:

- TensorFlow
- BeautifulSoup
- XGBoost
- Seaborn
- Scikit-learn
- NumPy

Обоснуйте выбор для каждой библиотеки, указав ключевые функции, которые она предоставляет в контексте машинного обучения.

Задание № 9 (тестовое)

Что представляет собой инкрементальная модель разработки ПО?

- Последовательное выполнение всех этапов разработки.
- Поэтапное улучшение и добавление новых функций к уже существующей системе.
- Разработка отдельных компонентов параллельно с интеграцией их в общую систему.
- Жесткие этапы, чередующиеся между разработкой и тестированием.

Задание № 10 (требуется развернутого ответа)

Реализация параллельного алгоритма сортировки (Quicksort).

1. Напишите программу на Python, которая:
 - Использует алгоритм Quicksort для сортировки массива чисел.
 - Реализует параллельное выполнение сортировки для улучшения производительности на больших наборах данных.
2. Требования к реализации:
 - Использовать библиотеку multiprocessing для распараллеливания.
 - Определить оптимальное количество процессов на основе количества ядер процессора.
3. Пример работы программы:

Ввод: [45, 12, 78, 34, 89, 23, 56, 9, 1, 67]

Вывод: [1, 9, 12, 23, 34, 45, 56, 67, 78, 89]

Раздел 4. Информационные источники

Список обязательной литературы

1. Фихтенгольц Г.М. Основы математического анализа. Т.1–2. М: Физматлит, 2002–2004.
2. Ильин В.А, Позняк Э.Г. Линейная алгебра, Изд. Физматлит, 2020
3. Фаддеев Д.К. Лекции по алгебре, СПб, 2002.
4. Боровков А. А. Теория вероятностей. – 4-е изд.- М.: Едиториал УРСС, 2003
5. Бородин А.Н. Элементарный курс теории вероятностей и математической статистики. – СПб: Лань, 1998
6. Ивченко Г.И., Медведев Ю.И. Введение в математическую статистику. - М.: ЛКИ, 2010. – 599 с.
7. Златопольский, Д.М. Основы программирования на языке Python. Москва: ДМК Пресс, 2017. 284 с.
8. Марк Лутц. Изучаем Python Т.1// М.: ИД «Вильямс». – 2019. – С. 832.
9. Марк Лутц. Изучаем Python Т.2// М.: ИД «Вильямс». – 2020. – С. 720.
10. Серебряков, В.А. Теория и реализация языков программирования. Учеб. пособие. Москва: Физматлит, 2012. 236 с.
11. Кузнецов С.Д. Базы данных: языки и модели. Учебник. М.: Бином-Пресс, 2008.
12. К. Дж. Дейт Введение в системы баз данных: учебное пособие / К. Дж. Дейт; пер. с англ. и ред.: К. А. Птицын. – 8-е изд. – М.; СПб.; Киев: Издательский дом "Вильямс", 2005. – 1328 с.
13. Вигерс К.И. Разработка требований к программному обеспечению. Русская редакция, 2004.
14. Липаев В.В. Программная инженерия. Методологические основы. М.: Государственный Университет – Высшая школа экономики, 2006.
15. Фредерик Брукс. Мифический человеко-месяц или как создаются программные системы. СПб.: «Символ-Плюс», 2000.
16. Управление проектами: учебник / под ред. Н.М. Филимоновой, Н.В. Моргуновой, Н.В. Родионовой. — Москва: ИНФРА-М, 2022. — 349 с.

17. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / пер. с англ. А.А. Слинкина — М.: ДМК Пресс, 2015 — 400с.
18. Душкин, Р. В. Искусственный интеллект / Р. В. Душкин. — Москва: ДМК Пресс, 2019. 280 с.
19. Силен Д., А. Мейсман А., Али М. Основы Data Science и Big Data. Python и наука о данных: учебное пособие. — СПб.: Питер, 2017.
20. Грас Д. Data Science. Наука о данных с нуля / СПб.: БХВ-Петербург, 2021. 416 с.

Список дополнительной литературы

1. Феллер В. Введение в теорию вероятностей и ее приложения. Т. 1, 2. — М.: Изд. Мир, 1984.
2. Севастьянов Б.А. Курс теории вероятностей и математической статистики. — М.: Изд. Наука, 1982.
3. Комбинаторика для программистов: Пер. с польского / В. Липский; пер. В. А. Евстигнеев, О. А. Логинова; ред. пер. А. П. Ершов. - М.: Мир, 1988. - 213 с.
4. Дейтел П., Дейтел Х. Python: Искусственный интеллект, большие данные и облачные вычисления. СПб.: Питер, 2020. 864 с.
5. Элбон Крис Машинное обучение с использованием Python. Сборник рецептов / Пер. с англ. СПб.: БХВ-Петербург, 2019. 384 с

Перечень иных информационных источников

1. Документация Python 3.10.0: <https://docs.python.org/3/>
2. Pro Git book, интернет-ресурс. <https://git-scm.com/book/ru/v2>
3. The Scrum Guide, интернет-ресурс. <https://www.scrum.org/resources/scrum-guide>