

## **Вопросы итогового экзамена по курсу «Интеллектуальный анализ данных»**

1. Определение интеллектуального анализа данных. Опишите, что такое интеллектуальный анализ данных (Data Mining) и в чём заключается его отличие от традиционной статистической обработки.
2. Приведите конкретные примеры применения методов Data Mining в различных отраслях (маркетинг, медицина, финансы и др.) и поясните выгоду от их использования.
3. Классификация типов данных. Какие существуют основные типы данных (номинальные, порядковые, интервальные, относительные) и в чём их различия при анализе?
4. В чём различие между двумерными и многомерными наборами данных? Какие дополнительные сложности возникают при работе с многомерными наборами?
5. Основные критерии качества данных. Какие аспекты (точность, полнота, согласованность и т. д.) определяют качество данных и почему это важно?
6. Какие метрики или методы позволяют оценить качество данных в реальном проекте (например, процент пропусков, дубликаты и т. д.)?
7. Какие задачи (очистка, интеграция, преобразование, сокращение) решаются на этапе предварительной обработки и почему именно они важны?
8. Методы очистки данных. Опишите несколько подходов к обработке пропусков и дубликатов данных (удаление, заполнение средними значениями, использование алгоритмов восстановления).
9. Объясните разницу между нормализацией (Min-Max, Z-Score) и стандартизацией данных. Когда и зачем их применяют?
10. Что такое дискретизация и как она помогает при использовании алгоритмов классификации и ассоциации?
11. Объясните, в каких случаях требуется бинаризация признаков, и приведите примеры реальных ситуаций, когда это уместно.
12. Выбор оптимальных методов предварительной обработки. Какие факторы влияют на выбор методов предварительной обработки в конкретном проекте (например, тип данных, алгоритмы, метрики качества)?
13. Опишите цель и основные задачи исследовательского анализа данных (EDA - Exploratory Data Analysis) и его отличие от финальных этапов моделирования.
14. Графические методы визуализации. Назовите 3–4 распространённых графических метода (гистограммы, boxplot, scatter plot и т. д.) и опишите, как они помогают понять структуру данных.
15. Как визуализация и статистические методы (например, boxplot) помогают обнаружить выбросы и аномалии в данных?
16. Корреляционный анализ. Как можно использовать корреляционные матрицы и диаграммы рассеяния для понимания взаимосвязей между признаками?
17. Выводы на основе EDA (Exploratory Data Analysis). Как результат исследовательского анализа данных (EDA) влияет на выбор дальнейших методов моделирования и на корректировку исходных гипотез?

18. Понятие контролируемого обучения. Объясните, что понимается под контролируемым обучением (supervised learning). Какова ключевая цель этого подхода?
19. Типичные задачи классификации. Приведите примеры типичных задач классификации и объясните, в чём заключается их актуальность.
20. Перечислите основные классические алгоритмы классификации (деревья решений, k-NN (k-ближайших соседей) и т. д.) и кратко охарактеризуйте их особенности.
21. Важность разбиения на обучение и тест. Как правильно делить данные на обучающую и тестовую выборки при создании моделей классификации?
22. Как работает алгоритм построения дерева решений? В чём ключевой смысл пошагового разбиения признаков?
23. Критерии разделения. Объясните, как рассчитываются индекс Джини и энтропия. В чём принципиальная разница в их подходах к оценке «неоднородности»?
24. Обрезка дерева (pruning). Что такое обрезка дерева (пререзка и постпререзка) и как она влияет на переобучение?
25. Базовые метрики производительности. Что означают точность (accuracy), прецизионность (precision), отзыв (recall) и F1-мера? Как они вычисляются?
26. Как строится ROC-кривая и что означает площадь под ROC-кривой (AUC)?
27. Методы перекрёстной проверки (cross-validation). В чём суть k-fold cross-validation (перекрёстной проверки) и почему перекрёстная проверка важна для более точной оценки модели?
28. Баланс между разными метриками. Как и почему выбор метрики зависит от специфики задачи (например, медицинская диагностика и фильтрация спама)?
29. Какие проблемы могут возникнуть, если выбирать только одну метрику (например, только точность) и игнорировать остальные?
30. Алгоритмы формирования правил. Как в общих чертах формируются правила для классификаторов (поиск шаблонов, жадные алгоритмы и т. д.)?
31. Как можно преобразовать дерево решений в набор правил и зачем это делать?
32. Байесовская формула и условная вероятность. Как используется формула Байеса в классификаторе Наив Байес (Naïve Bayes)?
33. В чём заключается «наивное» допущение алгоритма Наив Байес (Naïve Bayes) и почему оно иногда оказывается весьма эффективным?
34. Различные модели Наив Байес (Naïve Bayes). Какие существуют варианты наивных байесовских классификаторов (гауссовский, мультиномиальный и т. д.) и где они применяются?
35. Практические применения Наив Байес (Naïve Bayes). Приведите реальный пример задачи (например, фильтрация спама), где метод Наив Байес даёт хорошие результаты.
36. Ограничения и недостатки Наив Байес (Naïve Bayes). В каких случаях Naïve Bayes может дать неверные результаты и почему?
37. Основная идея k-NN. Каков принцип работы k-ближайших соседей (k-NN) в задаче классификации?
38. Выбор метрики расстояния. Какие метрики расстояния (евклидово, манхэттенское и т. д.) наиболее часто применяются в k-ближайших соседей (k-NN) и в чём их особенности?

39. Выбор оптимального  $k$  в  $k$ -ближайших соседей ( $k$ -NN). Как влияет выбор параметра  $k$  на качество классификации и как его подбирать?
40. Почему важно нормализовать признаки перед применением  $k$ -ближайших соседей ( $k$ -NN) и как это влияет на результат?
41. Преимущества и недостатки  $k$ -ближайших соседей ( $k$ -NN). Объясните, в каких ситуациях  $k$ -NN эффективен, а в каких – не очень. Учитывайте как вычислительную сложность, так и точность.
42. Модель персептрона. Как устроен персептрон в простейшем виде и как выполняется обучение путём корректировки весов?
43. Многослойная нейронная сеть. Что такое многослойный персептрон (MLP) и какую роль играют скрытые слои при обучении?
44. Процесс обратного распространения ошибки (backpropagation). Объясните, как реализуется алгоритм backpropagation и почему он эффективен для обучения глубоких сетей.
45. Роль функции активации в искусственных нейронных сетях. Зачем нужны функции активации (Sigmoid, ReLU, Tanh и др.) и как их выбор влияет на обучение?
46. Тенденции глубокого обучения. Назовите современные направления развития глубоких нейронных сетей (ResNet, Transformers и т. д.) и кратко опишите одну из них.
47. Основная идея машины опорных векторов (SVM). Каков принцип классификации в машинах опорных векторов с точки зрения максимизации «зазора» (margin)?
48. Методы ядра (kernel trick). Что такое «ядро» в машины опорных векторов (SVM) и как оно позволяет решать нелинейные задачи классификации?
49. Типы ядер в SVM (машина опорных векторов). Перечислите распространённые типы ядер (линейное, полиномиальное, RBF и т. д.) и поясните, как выбрать подходящее в конкретной задаче.
50. Объясните линейные и нелинейные алгоритмы SVM (машина опорных векторов) и их различия. Приведите пример их применения в реальных жизненных задачах.
51. Что такое ансамблевые методы и почему объединение нескольких слабых моделей может привести к более высокой точности?
52. Как работает метод бэггинга (Bagging) и почему случайный лес (Random Forest) считается одним из лучших его примеров?
53. В чём принципиальное отличие бустинга (Boosting) от бэггинга (Bagging)? Как изменяются «веса» примеров в процессе обучения?
54. Опишите основные идеи случайного леса (Random Forest). Какие параметры особенно важны при обучении (число деревьев, глубина деревьев и др.)?
55. Как формируется итоговая композиция в AdaBoost, и в чём особенность обновления весов объектов классификации?
56. Что понимается под несбалансированностью классов и почему это усложняет задачу классификации?
57. Перечислите несколько подходов к работе с несбалансированными данными (oversampling, undersampling, SMOTE и т. д.) и кратко охарактеризуйте их.
58. Приведите пример реальной области (например, обнаружение мошенничества), где имеет место сильный дисбаланс классов. Как там решается проблема?
59. Обучение без учителя (unsupervised learning). Чем кластеризация отличается от задач классификации (контролируемого обучения)? Приведите концептуальные примеры.

60. Задачи и цели кластеризации. Зачем нужно разбивать объекты на кластеры? Назовите минимум три практические задачи, где это полезно.
61. Каковы основные типы кластеризации (разбиение на группы (на основе центроидов), иерархическая, плотностная)? В чём их различие на концептуальном уровне?
62. Критерии «качества» кластера. Какие критерии и метрики (внутренние, внешние, силуэт и т. д.) используются, чтобы понять, хорошо ли объекты разделены на кластеры?
63. Алгоритм К-средних (k-means). Как пошагово работает алгоритм k-means, начиная с инициализации центроидов и заканчивая остановкой алгоритма?
64. Выбор числа кластеров k в алгоритме k-means (К-средних). Какие методы (elbow method, silhouette score) помогают определить оптимальное количество кластеров?
65. Какова вычислительная сложность k-means (К-средних) и какие факторы могут препятствовать быстрой сходимости?
66. Меры расстояния в k-means (К-средних). Почему часто используется евклидово расстояние? Можно ли применять другую метрику и как это повлияет на алгоритм?
67. Ограничения k-means (К-средних). В каких случаях k-means плохо работает (пример: не сферические кластеры) и какие есть альтернативы?
68. Опишите различия между агломеративной и иерархически-разделяющей (дивизивной) кластеризацией. Как выбирается стратегия слияния/разделения?
69. Методы связи (linkage) в иерархической кластеризации. В чём разница между одиночной (single), полной (complete) и средней (average) связью? Как это влияет на форму кластеров?
70. Построение дендрограммы в иерархической кластеризации. Как интерпретировать дендрограмму и на каком этапе принимается решение о количестве кластеров?
71. Как иерархические методы кластеризации реагируют на шумовые точки? Есть ли встроенные механизмы «защиты»?
72. В каких ситуациях иерархическая кластеризация может дать лучшее решение, чем k-means (К-средних)?
73. Как работает алгоритм DBSCAN? Объясните понятия «eps» (эпсилон), «minPts» (минимальные точки), «core points» (основные точки), «border points» (граничные точки) и «noise points» шумовые точки.
74. Выявление шума и выбросов. Почему DBSCAN способен выделять выбросы как отдельные объекты и почему это преимущество по сравнению с k-means?
75. Как формализуется идея достижимости и связности по плотности (density reachability и density connectivity) в методах кластеризация на основе плотности.?
76. Выбор параметров DBSCAN. Как выбрать параметры eps (эпсилон) и minPts (минимальные точки), на практике? Назовите хотя бы два метода или эвристики.
77. Плюсы и минусы DBSCAN. В каких случаях DBSCAN оказывается более эффективным, чем другие методы, и в чём его ограничения?
78. Что такое часто встречающиеся (frequent) наборы элементов? Приведите пример из розничной торговли (market basket analysis).
79. Как работает алгоритм Apriori: каково основное свойство для уменьшения пространства поиска?
80. Метрики оценки правил. Поясните смысл метрик «поддержка» (support), «уверенность» (confidence) и «подъём» (lift) в анализе ассоциаций.

81. Назовите примеры использования ассоциативных правил в индустрии (рекомендательные системы, cross-selling и т. д.).
82. Какие есть ограничения у классического алгоритма Apriori и какие способы улучшения (FP-Growth, Eclat) существуют?
83. Какие ключевые факторы (размер данных, тип задачи, доступные ресурсы и т. п.) нужно учитывать при выборе алгоритма интеллектуального анализа данных?
84. Сравнение контролируемого (supervised learning) и неконтролируемого обучения (unsupervised learning). Приведите примеры ситуаций, когда лучше применять контролируемые методы, и когда – неконтролируемые.
85. Возможно ли комбинировать методы классификации и кластеризации в одном проекте? Приведите пример такого подхода и объясните выгоды.
86. **Практическое задание:** Спроектируйте модель дерева решений на основе следующего набора данных и объясните метод построения модели.

**Набор данных:**

| ID | Возраст | Доход    | Студент | Кредитный рейтинг | Купить компьютер |
|----|---------|----------|---------|-------------------|------------------|
| 1  | <= 30   | Высокий  | Нет     | Справедливый      | Нет              |
| 2  | <= 30   | Высокий  | Нет     | Отличный          | Нет              |
| 3  | 31–40   | Высокий  | Нет     | Справедливый      | Да               |
| 4  | > 40    | Середина | Нет     | Справедливый      | Да               |
| 5  | > 40    | Низкий   | Да      | Справедливый      | Да               |

87. **Практическое задание:** Выполнить иерархическую кластеризацию (MIN, MAX, Average, Centroid) на основе следующего набора данных (в графической форме) и объяснить метод построения модели на основе иерархической кластеризации.

**Набор данных:**

| ID | X   | Y   |
|----|-----|-----|
| 1  | 1.0 | 1,5 |
| 2  | 1,5 | 1,8 |
| 3  | 5.0 | 8.0 |
| 4  | 8.0 | 8.0 |
| 5  | 1.0 | 0,6 |

88. **Практическое задание:** Разработайте структуру модели случайного леса на основе следующего набора данных и объясните метод построения модели.

**Набор данных:**

| ID | Feature1 | Feature2 | Label |
|----|----------|----------|-------|
| 1  | 2,5      | 0,5      | 0     |
| 2  | 3.0      | 1,5      | 1     |
| 3  | 2.0      | 2.0      | 0     |
| 4  | 3,5      | 3.0      | 1     |
| 5  | 1,5      | 0,5      | 0     |

89. **Практическое задание:** Спроектировать дерево решений (Decision Tree) на основе следующего набора данных и объяснить метод построения модели на основе дерева решений.

**Набор данных:**

| № | Пол     | Возраст | Доход    | Купит это |
|---|---------|---------|----------|-----------|
| 1 | Мужской | 25      | Высокий  | Да        |
| 2 | Женский | 30      | Середина | Нет       |
| 3 | Мужской | 35      | Низкий   | Нет       |
| 4 | Женский | 45      | Высокий  | Да        |
| 5 | Мужской | 40      | Середина | Да        |

90. **Практическое задание:** Выполните иерархическую кластеризацию (MIN, MAX, Average, Centroid) на основе следующего набора данных и объясните, как построить модель.

**Набор данных:**

| № | X   | Да  |
|---|-----|-----|
| 1 | 1.0 | 2.0 |
| 2 | 2,5 | 4,5 |
| 3 | 3.0 | 6.0 |
| 4 | 8.0 | 8,5 |
| 5 | 1,5 | 0,8 |

91. **Практическое задание:** Разработайте структуру модели для построения алгоритма случайного леса на основе следующего набора данных и объясните метод построения модели.

**Набор данных:**

| № | Признак1 | Признк2 | Класс |
|---|----------|---------|-------|
| 1 | 1.0      | 3,5     | 0     |
| 2 | 2.0      | 2.0     | 1     |
| 3 | 3,5      | 1,5     | 1     |
| 4 | 4.0      | 3.0     | 0     |
| 5 | 5.0      | 2,5     | 1     |

92. **Практическое задание:** Создайте дерево решений на основе следующего набора данных и объясните, как построить модель.

**Набор данных:**

| № | Тип работы | Опыт | Зарплата | Повышение |
|---|------------|------|----------|-----------|
| 1 | Менеджер   | 5    | Высокий  | Да        |
| 2 | Инженер    | 3    | Середина | Нет       |
| 3 | Техник     | 2    | Низкий   | Нет       |
| 4 | Инженер    | 8    | Высокий  | Да        |

|   |          |    |         |    |
|---|----------|----|---------|----|
| 5 | Менеджер | 10 | Высокий | Да |
|---|----------|----|---------|----|

93. **Практическое задание:** выполнить иерархическую кластеризацию на основе следующего набора данных и объяснить метод построения модели.

**Набор данных:**

| № | X   | Y   |
|---|-----|-----|
| 1 | 1.2 | 0,9 |
| 2 | 2,8 | 3.4 |
| 3 | 3.0 | 5.0 |
| 4 | 6.0 | 7,5 |
| 5 | 0,5 | 1.0 |

94. **Практическое задание:** Постройте модель дерева решений на основе следующего набора данных и объясните метод построения модели.

**Набор данных:**

| № | Экзамен1 | Экзамен2 | Прошел |
|---|----------|----------|--------|
| 1 | 25       | 20000    | Нет    |
| 2 | 34       | 45000    | Да     |
| 3 | 29       | 32000    | Нет    |
| 4 | 45       | 50000    | Да     |
| 5 | 39       | 30000    | Нет    |

95. **Практическое задание:** разработать алгоритм случайного леса на основе следующего набора данных и объяснить метод построения модели.

**Набор данных:**

| № | Экзамен1 | Экзамен2 | Прошел |
|---|----------|----------|--------|
| 1 | 85       | 78       | Да     |
| 2 | 52       | 61       | Нет    |
| 3 | 89       | 92       | Да     |
| 4 | 55       | 45       | Нет    |
| 5 | 72       | 81       | Да     |

96. **Практическое задание:** Выполните иерархическую кластеризацию (MIN, MAX, Average, Centroid) на основе следующего набора данных и объясните, как построить модель.

**Набор данных:**

| № | X   | Y   |
|---|-----|-----|
| 1 | 1.2 | 2.3 |
| 2 | 1,5 | 2,7 |
| 3 | 5.0 | 7.0 |
| 4 | 7,8 | 8.1 |

|   |     |     |
|---|-----|-----|
| 5 | 2.0 | 2,9 |
|---|-----|-----|

97. **Практическое задание:** Разработайте алгоритм случайного леса на основе следующего набора данных и объясните, как построить модель.

**Набор данных:**

| ID | Feature1 | Feature1 | Feature1 | Target |
|----|----------|----------|----------|--------|
| 1  | 3.0      | 2,5      | 1,8      | 0      |
| 2  | 2,8      | 3,5      | 1,5      | 1      |
| 3  | 4.0      | 4.0      | 2,5      | 1      |
| 4  | 3.2      | 2,8      | 1.0      | 0      |
| 5  | 4,5      | 4.1      | 2,8      | 1      |

98. **Практическое задание:** Постройте матрицу расстояний для иерархической кластеризации на основе следующего набора данных и укажите метод связи (единичный, полный) и объясните метод построения модели.

**Набор данных:**

| ID | X | Y |
|----|---|---|
| 1  | 2 | 4 |
| 2  | 6 | 8 |
| 3  | 4 | 2 |
| 4  | 8 | 6 |
| 5  | 5 | 7 |

99. **Практическое задание:** Разработайте модель для построения алгоритма случайного леса на основе следующего набора данных и объясните метод построения модели.

**Набор данных:**

| ID | Feature1 | Feature2 | Label |
|----|----------|----------|-------|
| 1  | 1,5      | 0,5      | 0     |
| 2  | 3.0      | 1,5      | 1     |
| 3  | 2.0      | 2.0      | 0     |
| 4  | 4.0      | 3.0      | 1     |
| 5  | 5.0      | 4,5      | 1     |

100. **Практическое задание:** выполнить иерархическую кластеризацию на основе следующего набора данных и объяснить метод построения модели.

**Набор данных:**

| ID | X   | Y   |
|----|-----|-----|
| 1  | 1.2 | 0,9 |
| 2  | 2,8 | 3.4 |
| 3  | 3.0 | 5.0 |
| 4  | 6.0 | 7,5 |
| 5  | 0,5 | 1.0 |



## **Внимание студентам!**

Использование мобильных телефонов, планшетов, ноутбуков, «шпаргалок» и других вспомогательных устройств во время итоговой проверки строго запрещено. Студенты, попытавшиеся воспользоваться подобными устройствами, будут немедленно отстранены от экзамена.

Просим вас тщательно подготовиться к экзаменационным вопросам заранее.

Консультация по всем неясным вопросам контрольных заданий состоится **13 января 2025 года в 13:00** в аудитории **Ф-704**. Принять участие может любой желающий.