# Контрольные и тестовые вопросы по ПР2

# «Анализ данных с использованием кластерного анализа в RapidMiner» по вариантам с ответами

**Вариант 1**

1. Какой тип алгоритма является основой метода K-Means?
   A) Иерархический
   B) Разделяющий (partitioning)
   C) Эволюционный
   D) Графовый

**Ответ: B**

2. Что произойдет при выборе слишком большого значения k в алгоритме K-Means?
   A) Увеличится точность на тестовой выборке
   B) Кластеры станут разреженными и переобученными
   C) Алгоритм станет иерархическим
   D) Улучшится интерпретируемость кластеров

**Ответ: B**

3. Почему важно нормализовать данные перед кластерным анализом?
   A) Для визуализации
   B) Чтобы уменьшить число итераций
   C) Чтобы все признаки вносили равный вклад в метрику расстояния
   D) Для бинаризации переменных

**Ответ: C**

4. Какая метрика расстояния чаще всего используется в K-Means?
   A) Манхэттенское
   B) Евклидово
   C) Косинусное
   D) Джаккарда

**Ответ: B**

5. Что может служить критерием остановки в K-Means?
   A) Превышение порога времени
   B) Сходимость центроидов или достижение максимального числа итераций
   C) Рост качества классификации
   D) Размер обучающей выборки

**Ответ: B**

6. В каком случае K-Means может не сработать корректно?
   A) При наличии категориальных признаков без кодирования
   B) При использовании данных без выбросов
   C) При одинаковом числе признаков и наблюдений
   D) При применении PCA перед кластеризацией

**Ответ: A**

7. Какое преимущество дает построение графиков распределения по кластерам после кластеризации?
   A) Снижает дисперсию модели
   B) Обеспечивает визуальную проверку корректности кластеров
   C) Ускоряет расчеты
   D) Автоматически определяет количество кластеров

**Ответ: B**

8. Почему K-Means чувствителен к начальному положению центроидов?
   **Ответ:** Инициализация влияет на траекторию схождения алгоритма, что может привести к разным результатам кластеризации при разных запусках.

9. Как можно оценить качество кластеризации без учёта истинных меток?
   **Ответ:** С помощью метрик внутренней оценки, таких как силуэтный коэффициент, индекс Калински-Харабаза, внутрикластерная и межкластерная дисперсия.

10. Назовите два типа визуализаций, полезных для интерпретации кластеров в RapidMiner.
    **Ответ:** Boxplot (для анализа распределения признаков по кластерам), Heatmap (для визуального сравнения средних значений).

---

**Вариант 2**

1. Какая характеристика является основной целью применения кластерного анализа?
   A) Предсказание целевого признака
   B) Максимизация плотности
   C) Группировка объектов на основе схожести
   D) Снижение размерности

**Ответ: C**

2. Какой оператор RapidMiner используется для автоматической замены пропущенных значений?
   A) Filter Examples
   B) Normalize
   C) Replace Missing Values
   D) Select Attributes

**Ответ: C**

3. Что обозначает «центроид» в алгоритме K-Means?
   A) Центр распределения всех признаков
   B) Среднее значение признаков в кластере
   C) Первый элемент кластера
   D) Медиана всех наблюдений

**Ответ: B**

4. Почему выбор метрики расстояния критичен в кластеризации?
   A) Она определяет число кластеров
   B) Она влияет на структуру кластеров и границы между ними
   C) Она нормализует признаки
   D) Она кодирует категориальные признаки

**Ответ: B**

5. Как влияет наличие выбросов на результат кластеризации методом K-Means?
   A) Алгоритм становится устойчивее
   B) Центроиды адаптируются
   C) Центроиды смещаются, ухудшая разделение кластеров
   D) Выбросы игнорируются автоматически

**Ответ: C**

6. Что является типичным недостатком K-Means?
   A) Не умеет работать с числовыми признаками
   B) Требует нормализации и фиксированного числа кластеров
   C) Требует разметки данных
   D) Ограничен только двумерными задачами

**Ответ: B**

7. Что означает плотность внутри кластера?
   A) Количество объектов в кластере
   B) Среднее расстояние до центроида
   C) Степень схожести объектов между собой
   D) Отношение размера кластера к его среднему

**Ответ: C**

8. Почему алгоритм K-Means не рекомендуется для кластеризации категориальных признаков?
   **Ответ:** Он основан на вычислении расстояний, которые неприменимы к категориальным значениям без предварительного кодирования.

9. Назовите одну метрику, которая может быть визуализирована через Heatmap для оценки качества кластеров.
   **Ответ:** Среднее значение признака внутри каждого кластера (mean attribute value).

10. Какие типы кластеров плохо выявляет K-Means?
    **Ответ:** Нелинейные, не сферические, разной плотности или сильно перекрывающиеся кластеры.

---

**Вариант 3**

1. Какая предпосылка лежит в основе работы алгоритма K-Means?
   A) Кластеры должны быть плотными и равномерно распределёнными
   B) Все признаки категориальные
   C) Кластеры имеют одинаковую дисперсию и форму
   D) Центроиды равны нулю

**Ответ: C**

2. Что такое «инерция» в контексте кластерного анализа?
   A) Скорость схождения центроидов

B) Количество итераций
C) Сумма квадратов расстояний точек до ближайшего центроида
D) Влияние выбросов на кластеризацию

**Ответ: C**

3. Что будет, если не нормализовать данные перед кластеризацией?
   A) Улучшится интерпретируемость
   B) Все признаки будут обрабатываться одинаково
   C) Признаки с большим масштабом будут доминировать, искажая кластеризацию
   D) Алгоритм остановится

**Ответ: C**

4. Что делает оператор "Clustering Performance" в RapidMiner?
   A) Объединяет кластеры
   B) Определяет количество кластеров
   C) Вычисляет внутрикластерную и межкластерную дисперсии
   D) Удаляет шум

**Ответ: C**

5. Как влияет увеличение числа кластеров на значение внутрикластерной дисперсии?
   A) Оно увеличивается
   B) Оно уменьшается
   C) Оно остаётся постоянным
   D) Оно становится недоступным

**Ответ: B**

6. Какой способ инициализации центроидов минимизирует вероятность попадания в локальные минимумы?
   A) Случайный выбор
   B) K-Means++
   C) Первый элемент
   D) Среднее значение всех данных

**Ответ: B**

7. Когда использование силуэтного коэффициента наиболее оправдано?
   A) При линейной регрессии
   B) Для оценки качества модели классификации

C) Для оценки качества кластеризации без меток классов
D) Только при использовании PCA

**Ответ: C**

8. В чём заключается основная идея K-Means++ инициализации?
   **Ответ:** Центроиды выбираются так, чтобы они были максимально удалены друг от друга, что повышает шансы на успешную кластеризацию.

9. Почему K-Means не работает с отсутствующими значениями?
   **Ответ:** Алгоритм требует численного сравнения расстояний, и пропущенные значения нарушают вычисление расстояния между точками.

10. Какой подход может быть использован в RapidMiner для автоматического определения числа кластеров?
    **Ответ:** Построение графика зависимости внутрикластерной дисперсии от числа кластеров («метод локтя»).

---

**Вариант 4**

1. Что представляет собой кластер в результате работы алгоритма K-Means?
   A) Набор признаков
   B) Совокупность центроидов
   C) Группа объектов, близких по метрике расстояния
   D) Граф взаимосвязей

**Ответ: C**

2. Какая стратегия используется в K-Means для назначения объектов кластерам?
   A) Случайная
   B) Максимизация плотности
   C) Назначение по минимальному расстоянию до центроида
   D) Принудительное деление на равные группы

**Ответ: C**

3. Что такое межкластерная дисперсия?
   A) Вариация внутри одного кластера
   B) Среднее расстояние между объектами одного кластера

C) Расстояние между центроидами кластеров
D) Объём выборки

**Ответ: C**

4. Какой тип данных нужно преобразовать перед применением K-Means?
   A) Признаки с отсутствующими значениями
   B) Нормализованные числовые данные
   C) Категориальные признаки
   D) Класс метки

**Ответ: C**

5. Почему важно учитывать дисбаланс в данных при кластеризации?
   A) Он ускоряет расчёты
   B) Он увеличивает плотность кластеров
   C) Он может привести к формированию односторонних кластеров
   D) Он всегда автоматически устраняется

**Ответ: C**

6. Как визуализировать результат кластеризации в RapidMiner?
   A) С помощью Apply Model
   B) Построить диаграмму рассеяния с цветом по кластеру
   C) Через Cross Validation
   D) С помощью таблицы корреляций

**Ответ: B**

7. Какой недостаток у K-Means при наличии кластеров разной плотности?
   A) Центроиды смещаются в сторону малых кластеров
   B) Алгоритм работает быстрее
   C) Центроиды не обновляются
   D) Он становится чувствительным к нормализации

**Ответ: A**

8. Что происходит, если задать k больше, чем количество уникальных объектов?
   **Ответ:** Некоторые кластеры останутся пустыми, так как не смогут быть заполнены уникальными точками.

9. Назовите два критерия, по которым можно сравнивать альтернативные результаты кластеризации.

**Ответ:** Внутрикластерная дисперсия (inertia), силуэтный коэффициент (silhouette score).

10. Как можно оценить устойчивость кластеризации?
    **Ответ:** Повторный запуск с разной инициализацией центроидов, сравнение результатов по стабильности кластерных распределений.