31. Задачи многомерной классификации объектов. Методы многомерных классификаций и их основные характеристики (задачи какого вида решают). Кластерный анализ. Основные типы задач и алгоритмов кластерного анализа. Расстояние между объектами.

1. Задачи многомерной классификации объектов

Многомерная классификация направлена на разделение совокупности объектов, каждый из которых характеризуется набором из p признаков ($X_1, X_2, \ldots, X_p$), на относительно однородные группы или классы.

Основные задачи:

- **Выявление структуры:** Обнаружение «естественного» разбиения совокупности на группы, когда априорная информация о количестве и составе классов отсутствует.
- **Дискриминация (различение):** Отнесение нового объекта к одному из уже известных (заданных) классов на основе его характеристик.
- **Сжатие информации:** Описание множества объектов через характеристики нескольких типичных групп (кластеров), что упрощает дальнейший анализ.
- **Построение таксономии:** Создание классификационных схем для ранее не изученных явлений.

2. Методы многомерных классификаций и их характеристики

В зависимости от наличия обучающей информации выделяют два подхода:

1. **Дискриминантный анализ («Классификация с обучением»):**
    - **Вид задачи:** Решает задачу идентификации объектов в ситуации, когда **классы известны заранее**.
    - **Характеристика:** На основе обучающей выборки строится решающее правило (дискриминантная функция), позволяющее предсказать принадлежность нового объекта к группе.
2. **Кластерный анализ («Классификация без обучения»):**
    - **Вид задачи:** Используется, когда **структура групп изначально неизвестна**.
    - **Характеристика:** Объекты объединяются в группы (кластеры) исключительно на основе их сходства в многомерном пространстве признаков.

Кластерный анализ представляет собой совокупность методов многомерной классификации, предназначенных для объединения объектов, описываемых набором признаков, в относительно однородные группы (кластеры). В отличие от дискриминантного анализа, кластерный анализ является **классификацией «без учителя»**, так как структура разбиения на группы и их количество изначально неизвестны.

3. Основные цели и задачи кластерного анализа

Главная цель — разделение множества объектов на кластеры таким образом, чтобы **объекты внутри одного кластера были максимально схожи**, а объекты из разных кластеров — максимально различны.

В источниках выделяют следующие типы задач кластерного анализа:

- **Построение таксономии:** разработка классификационных схем для ранее не изученных совокупностей (например, классификация регионов по уровню инвестиционной привлекательности).
- **Исследование концептуальных схем:** выявление скрытых закономерностей и «естественных» скоплений в данных для понимания их внутренней структуры.
- **Сжатие информации:** описание больших массивов данных через характеристики типичных групп (профили кластеров), что существенно упрощает анализ.
- **Проверка гипотез:** подтверждение или опровержение предположений о существовании определенных типов или классов в исследуемой совокупности.

- **Выделение аномалий:** обнаружение объектов, которые не вписываются ни в одну из групп (так называемых «выбросов»), что полезно для контроля качества данных.

4. Алгоритмы кластерного анализа

Выбор алгоритма зависит от объема данных и требуемого результата классификации. Основные методы делятся на две категории:

Иерархические алгоритмы

Эти методы строят систему вложенных кластеров, которая визуализируется в виде древовидной схемы — **дендрограммы**.

- **Агломеративные (объединительные):** процесс начинается с того, что каждый объект считается отдельным кластером. На каждом шаге два наиболее близких кластера объединяются, пока вся выборка не сольется в одну группу.
- **Дивизимные (разделяющие):** работают в обратном порядке — изначально вся совокупность рассматривается как один кластер, который последовательно дробится на более мелкие части.

Неиерархические (итеративные) алгоритмы

Наиболее популярным среди них является **метод K-средних (K-means)**.

- **Принцип работы:** исследователь заранее задает количество кластеров (k). Алгоритм итеративно перераспределяет объекты между группами таким образом, чтобы минимизировать внутрикластерную изменчивость (сумму квадратов отклонений объектов от центров их кластеров).
- **Особенность:** такие методы более эффективны для обработки больших массивов данных, чем иерархические.

5. Расстояние между объектами

Близость объектов в многомерном пространстве определяется с помощью **метрики (расстояния)**. Если объект i описывается вектором $X_i = (x_{i1}, x_{i2}, \dots, x_{ip})$, а объект j — вектором $X_j = (x_{j1}, x_{j2}, \dots, x_{jp})$, то используются следующие формулы,:

1. **Евклидово расстояние (Euclidean distance):**
   Наиболее распространенная мера геометрической близости объектов.
$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$
2. **Евклидово взвешенное расстояния:**
   Используется для усиления влияния больших различий между объектами.
$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 w_k}$$

где $\sum_1^p w_k = 1$. w_k - степень важности переменной x_{ik}

3. **Расстояние городских кварталов / Хэмингово:**
   Сумма модулей разностей координат; менее чувствительно к отдельным выбросам.
$$d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$
4. **Расстояние Махаланобиса:**
$$d_{ik} = \sqrt{(x_i - x_k)^T \Lambda^T \Sigma^{-1} \Lambda (x_i - x_k)}$$

где Λ - диагональная матрица весовых коэффициентов исходных показателей, Σ - ковариационна матрица.

5. **Расстояние Минковского**
$$d_{ij} = \sqrt[p]{\sum_{k=1}^{p} |x_{ik} - x_{jk}|^p}$$