Ответы на контрольные вопросы

1. Какие существуют алгоритмы кластерного анализа данных? Назовите не менее 3-х и опишите их суть с математической точки зрения и расскажите чем они отличаются друг от друга.

Алгоритмы иерархической кластеризации:

Среди алгоритмов иерархической кластеризации выделяются два основных типа: восходящие и нисходящие алгоритмы. Нисходящие алгоритмы работают по принципу «сверху-вниз»: в начале все объекты помещаются в один кластер, который затем разбивается на все более мелкие кластеры. Более распространены восходящие алгоритмы, которые в начале работы помещают каждый объект в отдельный кластер, а затем объединяют кластеры во все более крупные, пока все объекты выборки не будут содержаться в одном кластере. Таким образом строится система вложенных разбиений. Результаты таких алгоритмов обычно представляют в виде дерева – дендрограммы. Классический пример такого дерева – классификация животных и растений.

Для вычисления расстояний между кластерами чаще все пользуются двумя расстояниями: одиночной связью или полной связью.

К недостатку иерархических алгоритмов можно отнести систему полных разбиений, которая может являться излишней в контексте решаемой задачи.

Алгоритмы квадратичной ошибки:

Задачу кластеризации можно рассматривать как построение оптимального разбиения объектов на группы. При этом оптимальность может быть определена как требование минимизации среднеквадратической ошибки разбиения:

$$e^2(X,L) = \sum_{j=1}^{K} \sum_{i=1}^{n_j} \| x_i^{(j)} - c_j \|^2$$

где cj — «центр масс» кластера j (точка со средними значениями характеристик для данного кластера).

Алгоритмы квадратичной ошибки относятся к типу плоских алгоритмов. Самым распространенным алгоритмом этой категории является метод k-средних. Этот алгоритм строит заданное число кластеров, расположенных как можно дальше друг от друга. Работа алгоритма делится на несколько этапов:

- Случайно выбрать k точек, являющихся начальными «центрами масс» кластеров.
- Отнести каждый объект к кластеру с ближайшим «центром масс».
- Пересчитать «центры масс» кластеров согласно их текущему составу.
- Если критерий остановки алгоритма не удовлетворен, вернуться к п. 2.

В качестве критерия остановки работы алгоритма обычно выбирают минимальное изменение среднеквадратической ошибки. Так же возможно останавливать работу алгоритма, если на шаге 2 не было объектов, переместившихся из кластера в кластер.

К недостаткам данного алгоритма можно отнести необходимость задавать количество кластеров для разбиения.

Нечеткие алгоритмы:

Наиболее популярным алгоритмом нечеткой кластеризации является алгоритм c-средних (c-means). Он представляет собой модификацию метода k-средних. Шаги работы алгоритма:

- Выбрать начальное нечеткое разбиение n объектов на k кластеров путем выбора матрицы принадлежности U размера n x k.
- Используя матрицу U, найти значение критерия нечеткой ошибки:

$$E^2(X,U) = \sum_{i=1}^{N} \sum_{k=1}^{K} U_{ik} \| x_i^{(k)} - c_k \|^2$$

где c_k — «центр масс» нечеткого кластера k:

$$c_k = \sum_{i=1}^{N} U_k x_i$$

- Перегруппировать объекты с целью уменьшения этого значения критерия нечеткой ошибки.
- Возвращаться в п. 2 до тех пор, пока изменения матрицы U не станут незначительными.

Этот алгоритм может не подойти, если заранее неизвестно число кластеров, либо необходимо однозначно отнести каждый объект к одному кластеру.

Алгоритм выделения связных компонент:

В алгоритме выделения связных компонент задается входной параметр R и в графе удаляются все ребра, для которых «расстояния» больше R. Соединенными остаются только наиболее близкие пары объектов. Смысл алгоритма заключается в том, чтобы подобрать такое значение R, лежащее в диапазон всех «расстояний», при котором граф «развалится» на несколько связных компонент. Полученные компоненты и есть кластеры.

Для подбора параметра R обычно строится гистограмма распределений попарных расстояний. В задачах с хорошо выраженной кластерной структурой данных на гистограмме будет два пика – один соответствует внутрикластерным расстояниям, второй – межкластерным расстояния. Параметр R подбирается из зоны минимума между этими пиками. При этом управлять количеством кластеров при помощи порога расстояния довольно затруднительно.

Сравнение алгоритмов:

Вычислительная сложность алгоритмов:

Алгоритм кластеризации	Вычислительная сложность
Иерархический	O(n2)
k-средних c-средних	O(nkl), где k – число кластеров, l – число итераций
Выделение связных компонент	зависит от алгоритма

Сравнительная таблица алгоритмов:

Алгоритм кластеризации	Форма кластеров	Входные данные	Результаты
Иерархический	Произвольная	Число кластеров или порог расстояния для усечения иерархии	Бинарное дерево кластеров
k-средних	Гиперсфера	Число кластеров	Центры кластеров
c-средних	Гиперсфера	Число кластеров, степень нечеткости	Центры кластеров, матрица принадлежности
Выделение связных компонент	Произвольная	Порог расстояния R	Древовидная структура кластеров