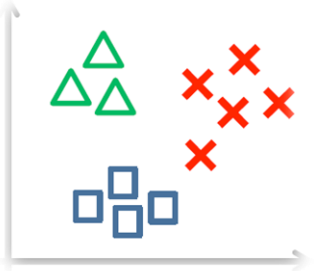
Кластеризация



Задание.

Разработать программу обработки неразмеченных данных, позволяющую проводить разбиение выборки на кластеры методом k-средних.

- Ч.1. Разработать программу, реализующую алгоритм метода k-средних, для разбиения выборки на заданное количество K кластеров за заранее заданное количество итераций.
- Ч.2. Модифицировать алгоритм и программу кластеризации, учитывающую возможность автоматического определения количества итераций.
- Ч.3 Модифицировать алгоритм и программу кластеризации, учитывающую возможность автоматического выбора начальных центроидов.

Кластеризация

Ч.1. Разработать программу, реализующую алгоритм метода k-средних, для разбиения выборки на заданное количество K кластеров за заранее заданное количество итераций.

1. Загрузить данные из файла «data.csv» с помощью функции в виде DataFrame размером [n×m], где n – кол-во объектов, m – кол-во признаков;
2. Задать количество кластеров K=3;
3. Рандомизировать выборку и выбрать K центроидов $C^{(k)}=\{C_i^{(k)}\}_{[K\times m]}$;
4. Разработать функцию [M]=Object(X,C), позволяющую распределить объекты по кластерам, используя принцип наименьшего расстояния до кластера:

$$X_i \in C^{(k)}, if\ \rho_{ik} = \min_{p=1,..K}\left(\left\|x_{ij} - C_j^{(p)}\right\|^2\right),$$

$$i = \overline{1,N}, k - \text{номер кластера}, j = \overline{1,m} - \text{метрность пространства}$$

примечание: удобнее запоминать не координаты объектов, а номера кластеров в виде массива M=[Mi] размером [n×1], где строка i соответствует номеру объекта, n – кол-во объектов.

Кластеризация

Ч.1. Разработать программу, реализующую алгоритм метода k-средних, для разбиения выборки на заданное количество K кластеров за заранее заданное количество итераций.

5. Разработать функцию [C]=Centriods(X,M,K), позволяющую переопределять координаты центроидов как среднее значение соответствующих координат:

$$C_j^{(k)} = \frac{1}{N_k} \sum_{\substack{i=1 \\ (M_i==k)}}^{N_k} x_{ij}$$

6. Повторять пункты 4 – 5 заданное количество итераций.

7. Представить результат работы программы графически в виде трехмерного графика (фиксируя три произвольных признака), на котором отобразить объекты разными маркерами в соответствии с кластером.

Кластеризация

Ч.2. Модифицировать алгоритм и программу кластеризации, учитывающую возможность автоматического определения количества итераций.

2.1 На каждой итерации считать ошибку в виде суммарного расстояния от каждого объекта до соответствующего центроида.
2.2 Условием остановки итерационного процесса считать незначительное отклонение *e* ошибки на текущем и предыдущем шаге:

$$\left|J^{iter} - J^{iter-1}\right| < e$$

Ч.3. Модифицировать алгоритм и программу кластеризации, учитывающую возможность автоматического выбора начальных центроидов.

3.1 Задать количество прогонов p=2,3…P;
3.2 Для каждого прогона начальные центроиды инициализируются случайным образом:

$$C_p^{(k)} = \{C_{pi}^{(k)}\}_{[K \times m]}$$

3.3 Модифицировать программу так, чтобы кластеризация проводилась для каждого значения p и записывалась ошибка J для каждого p;
3.4 Выбрать лучшее решение из условия:

$$J^* = \min_p J^p$$