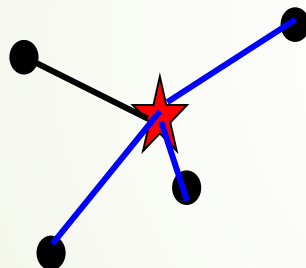
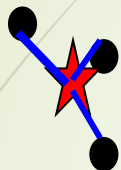


Генеративная модель для кластерного анализа

Д.т.н. Миркин Борис Григорьевич

- **Профессор, ФКН НИУ Высшая школа экономики, Москва РФ**
- **Prof. Emeritus, University of London UK**
- **Member, Academia Europaea**

Самый популярный метод кластер-анализа К-средних (K-Means)



Критерий:
минимизировать сумму
расстояний между
точками/объектами и
центрами их кластеров $D(S, c)$

S - кластеры

C – центры (звездочки)

$D(S, c)$ - суммарное расстояние

K-Means criterion:

5

Find partition S and centers c to minimize:

$$D(S, c) = \sum_{k=1}^K \sum_{i \in S_k} d(i, c_k)$$

Criterion: Sum of distances between entities and centers of their clusters

Distance $d(.,.)$ (squared Euclidean):

$$X = [1, 2, -2]$$

$$Y = [1, -1, -1]$$

$$X - Y = [1 - 1, 2 - (-1), -2 - (-1)] = [0, 3, -1]$$

$$d(X, Y) = \langle X - Y, X - Y \rangle = 0^2 + 3^2 + (-1)^2 = 10$$

A generative model for K-means:

K-means criterion

$$D(S, c) = \sum_{k=1}^K \sum_{i \in S_k} \sum_{v=1}^V (y_{iv} - c_{kv})^2$$

S_k is cluster, v - feature, y_{iv} - v -th coordinate of object i , $c_k = (c_{kv})$ - S_k ' center

is the least-squares criterion for model

$$y_{iv} = c_{kv} + e_{iv} \quad \text{for } i \in S_k, k=1, 2, \dots, K$$

or

$$y_{iv} = \sum_{k=1}^K z_{ik} c_{kv} + e_{iv} \quad (*)$$

where $z_{ik} = 1$ for $i \in S_k$, and $z_{ik} = 0$ for $i \notin S_k$

Это модель матричной факторизации

8

$$Y \approx ZC^T$$

$$Y = ZC^T + E, \quad ||E||^2 \Rightarrow \min$$

- ➡ Y - $N \times V$ data matrix (observed, pre-processed)
- ➡ Z - $N \times K$ cluster incidence binary matrix to find
- ➡ C - $V \times K$ cluster center matrix to find

Подобная таким же моделям для

- ➡ PCA Principal Component Analysis
 - ➡ Z : factor scores, C : feature loadings
 - ➡ mutual orthogonality, maximum contribution
- ➡ Hierarchical clustering (Z - triple divergence base)

Три приложения

- **1. Сегментация данных**
- **2. Консенсус кластеринг**
- **3. Определение структурной сложности таблицы данных (на основе 1. и 2.)**

Сегментация данных

