K-means Clustering

1 Problem Definition

Given a data set $x_1, x_2, ..., x_n$, we want to partition the data set into k sets $S_1, S_2, ..., S_k$ so as to minimize

$$\sum_{i=1}^{k} \sum_{x \in S_k} (x - c_k)^2$$

where c_k is the center of points in S_k.

2 Sub-problem

2.1 Question

Given $x_1, x_2, ..., x_n$, find c to minimize $\sum_{i=1}^{n} (x_i - c)^2$.

2.2 Answer

We use first derivative.

$$\frac{d \sum_{i=1}^{n} (x_i - c)^2}{dc} = -2 \sum_{i=1}^{n} (x_i - c) = 0.$$

$$\therefore c = \frac{\sum_{i=1}^{n} x_i}{n}.$$

3 Algorithm

1. Initialize cluster centers $c_1, c_2, ..., c_k$ randomly.

2. Repeat until convergence

2-1. Assign cluster label to each data point.

2-2. Recalculate the centers.

4 Cosine Similarity

4.1 Question

Given $x_1, x_2, ..., x_n$, find c to **maximize** $\sum_{i=1}^{n} cos\theta(x_i, c)$ where $\theta(x_i, c)$ is the angle between x_i and c.

4.2 Reformulation

Observe that $\sum_{i=1}^{n} cos\theta(x_i, c) = \sum_{i=1}^{n} \frac{x_i \cdot c}{|x_i||c|}$.

When optimizing the above function, we face with the difficulty of dealing with the denominator. Fortunately cosine similarity does not be affected by the length of vectors. Thus we change the formulation of the problem.

4.3 Sub-problem : Cosine Case

Given $x_1, x_2, ..., x_n$, find c to maximize $\sum_{i=1}^{n} cos\theta(x_i, c)$ where $\theta(x_i, c)$ is the angle between x_i and c and **all $\mathbf{x_i}$ and c are unit vectors**.

4.4 Answer

$$\mathbf{argmax_c} \sum_{i=1}^{n} cos\theta(x_i, c)$$

$$= \mathbf{argmax_c} \sum_{i=1}^{n} \frac{x_i \cdot c}{|x_i||c|}$$

$$= \mathbf{argmax_c} \sum_{i=1}^{n} (x_i \cdot c) = (\sum_{i=1}^{n} x_i) \cdot c.$$

The inner product of two vectors are maximized when the two have the same direction. Therefore,

$$c = \frac{\sum_{i=1}^{n} x_i}{|\sum_{i=1}^{n} x_i|}.$$