CS4269/6362 Machine Learning, Spring 2016: Homework 4

Zejian Zhan

Instructions Please fill in your name above.

Question 1:

(a) There are two main steps in K-means algorithm: data assignment and new centroid process. For the data assignment process, each observation chooses the minimum distance(Euclidean distance) to the designated cluster center. So in this case, all observations choose the new nearest center, which apparently means the within cluster sum of square γ_k decreases. On the other hand, we mathematically compute the arithmetic mean for the new center, which serves as a way of Least-squares estimation. So the within cluster sum of square γ_k must decrease. Based on the two steps, we can clearly say that γ_k won't increase in K.

(b) We can first compute the two norm from $\phi(x_i)$ to α_k:

$$||\phi(x_i) - \alpha_k||^2 = \phi(x_i)^2 - \frac{2}{|S_k|}\sum_{x_j \in S_k}\phi(x_j)\phi(x_i) + \frac{1}{|S_k|^2}\sum_{x_j \in S_k, x_h \in S_k}\phi(x_j)\phi(x_h) \tag{1}$$

$$= k(x_i, x_i) - \frac{2}{|S_k|}\sum_{x_j \in S_k} k(x_i, x_j) + \frac{1}{|S_k|^2}\sum_{x_j \in S_k, x_h \in S_k} k(x_j, x_h) \tag{2}$$

So we can rewrite the formulate for the first step:

$$\arg\min_k ||\phi(x_i) - \alpha_k||^2 \tag{3}$$

$$= \arg\min_k \left\{ k(x_i, x_i) - \frac{2}{|S_k|}\sum_{x_j \in S_k} k(x_i, x_j) + \frac{1}{|S_k|^2}\sum_{x_j \in S_k, x_h \in S_k} k(x_j, x_h) \right\} \tag{4}$$

and the second step of updating the mean vector for each cluster:

$$\alpha_k = \frac{\sum_{x_i \in S_k}\phi(x_i)}{|S_k|} \tag{5}$$

Question 2:

(a) We can write the second step of updating cluster center by the following formula:

$$\arg\min_{x'} \sum_{x \in S_k} ||x - x'||_1 = \arg\min_{x'} \sum_{x \in S_k} \sum_{i=1}^{d} |x_i - x'_i| \tag{6}$$

$$= \arg\min_{x'} \sum_{i=1}^{d} \sum_{x \in S_k} |x_i - x'_i| \tag{7}$$

From the given useful fact and d n the above formula denotes the feature dimension, we can see that the updated cluster center is derived as the median of the ith dimensional features of data in S_k. And clearly that's why it's also called $K - median$

(b) The reason why it is also called $K - medians$ is that when we update the centroid, actually we always find the median of x_is in the set S_k. And by using $Kmedians$ we can make the clusters formed by them are the most compact.

Question 3:

(a) For E-step, we can write it like:

$$\gamma(z_k) = \frac{\pi_k \mathcal{N}(x|\mu_k, \sum_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x|\mu_j, \sum_j)} \tag{8}$$

For F-step, we can derive the formula like:

$$\mathbb{E}_Z[\ln p(x, z|\mu, \sum, \pi)] = \sum_{k=1} K\gamma(z_k) \left\{ \ln \pi_k + \ln \mathcal{N}(x|\mu_k, \sum_k) \right\} \tag{9}$$

Basically we maximize the likelihood with regards to μ,$\sum$ and π

(b) The total memory space will be $O(Km^2)$, where K denotes the dimension of π and m denotes the dimension of data.