CS 6362 Machine Learning, Fall 2017: Homework 4

Venkataramana Nagarajan

Question 1:

(a)

$$minS = S_1, S_2, ..., S_k \sum_{k=1}^{K} \sum_{x_i \epsilon S_k} ||x_i - \mu_k||^2{}_2$$

can be written as

$$min\left[S = S_1, S_2, ..., S_{k-1} \sum_{k=1}^{K-1} \sum_{x_i \epsilon S_k} ||x_i - \mu_k||^2{}_2 + \sum_{k=K} \sum_{x_i \epsilon S_k} ||x_i - \mu_k||^2{}_2\right]$$

What we did is just removed cluster k=K outside from the equation.

Now suppose that the cluster k=K is divided into 2 clusters.So we effectively have now k=K and K+1 clusters. Since the new clusters formed contain datapoints from the old cluster the new means of both clusters will be collectively closer to the new datapoints as compared to the distance between the old mean and the old data points.

Say that the new cluster k = K + 1 is created by taking one data point x_o from k = K^{th} cluster.

New mean of K^{th} cluster will be more closer to remaining points.

$$\sum_{x_i \epsilon S_{K'}} ||x_i - \mu_{K'}||^2{}_2 <= \sum_{x_i \epsilon S_K} ||x_i - \mu_K||^2{}_2$$

The cluster k = K + 1 has only 1 datapoint, so $\mu_{K+1} = x_o$

$$||x_o - \mu_{K+1}||^2{}_2 = 0$$

Thus we can say that the new γ formed will be smaller than the previous γ. Hence, γ_k is non-increasing.

(b)

$$\gamma_k = minS = S_1, .., S_k \sum_{K=1}^{K} \sum_{x_i \epsilon S_k} ||\phi(x_i) - \alpha_k||^2{}_2 \quad (1)$$

and

$$\alpha_k = \frac{1}{|S_k|} \sum_{x_i \epsilon S_k} \phi(x_i) \tag{2}$$

Substituting (2) in (1)

$$\gamma_k = minS = S_1, .., S_k \sum_{K=1}^{K} \sum_{x_i \epsilon S_k} ||\phi(x_i) - \frac{1}{|S_k|} \sum_{x_i \epsilon S_k} \phi(x_i)||^2_2 \tag{3}$$

We know that,

$$||x - y||^2_2 = (x - y).(x - y)$$

Changing (3) according to the above expansion

$$\gamma_k = minS = S_1, .., S_k \sum_{K=1}^{K} \sum_{x_i \epsilon S_k} ||\phi(x_i) - \frac{1}{|S_k|} \sum_{x_i \epsilon S_k} \phi(x_i)||.||\phi(x_i) - \frac{1}{|S_k|} \sum_{x_i \epsilon S_k} \phi(x_i)||$$

$$= minS = S_1, .., S_k \sum_{K=1}^{K} \sum_{x_i \epsilon S_k} \phi(x_i).\phi(x_i) + \frac{\sum_{x_i, x_j \epsilon S_k} \phi(x_i).\phi(x_j)}{|S_k|^2} - \frac{2 \sum_{x_i, x_j \epsilon S_k} \phi(x_i).\phi(x_j)}{|S_k|} \tag{4}$$

It is given that,

$$k(x, x_i) = \phi(x).\phi(x_i)$$

Putting it in (4)

$$\gamma_k = minS = S_1, .., S_k \sum_{K=1}^{K} \sum_{x_i \epsilon S_k} k(x_i, x_i) - \frac{2 \sum_{x_i, x_j \epsilon S_k} k(x_i, x_j)}{|S_k|} + \frac{\sum_{x_i, x_j \epsilon S_k} k(x_i, x_j)}{|S_k|^2}$$

Question 2:

(a) According to L1 norm the γ would be given as follows

$$\gamma = argmin_{j=1..k} ||x_i - \mu_j||_1$$

When we consider the L1 norm the centroid value is considered as the median. Hence μ_k is calculated in each single dimension(feature).

As new clusters are formed their μ_ks are updated considering the datapoints in those clusters.

(b) This algorithm is called K-median because the L1 norm uses median to calculate the centroid. Using median minimizes the distance for L1 norm. Since we are calculating median over each single dimension the individual attribute will come from the given dataset as opposed to k-means method.

Question 3:

(a) We have M data points here. Say that the total number of clusters is K. Now since we get 1 data point(x_i) at a time , the E-step would be given as:

$$Pr_{Z_k} = Pr_{(Z_k|x_i)} = \frac{Pr(x_i|Z_{ik} = 1, \mu_k, \sigma_k)\alpha_k}{\sum_{n=1}^{K} Pr(x_i|Z_k = 1, \mu_n, \sigma_n)\alpha_n}$$

We need to update μ_k, σ_k, α_k in the M-Step:

$$\sigma_k = \frac{M\alpha_k\sigma_k + Pr_{Z_k}}{M\alpha_k + 1} + \frac{M\alpha_k\mu_k{\mu_k}^T + Pr_{Z_k}x_i{x_i}^T}{M\alpha_k + Pr_{Z_k}} - \mu_k{\mu_k}^T$$

$$\mu_k = \frac{M\alpha_k\mu_k + Pr_{Z_k}x_i}{M\alpha_k + Pr_{Z_k}}$$

$$\alpha_k = \frac{M\alpha_k + Pr_{Z_k}}{M\alpha_k + 1} \qquad \text{(Since We have only 1 data point coming int)}$$

Note : I used a paper as a reference(Titled : Highly Efficient Incremental Estimation of Gaussian Mixture Models for Online Data Stream Clustering) for this in which they introduce M data points at a time. I basically substituted M with 1 and made some changes to the exiting E and M steps to come up with the above equations.

(b) Total memory requirement for streaming GMM is : O(KIm^2)