CS 6362 Machine Learning, Fall 2017: Homework 4

Venkataramana Nagarajan

Question 1:

(a) Let us take the assumption that k = m sets. This implies that there are m number of centroids in the data set. By inference if all m samples in the data set are being treated as centroids then the minimum for γ will be equal to 0. This implies that globally γ will always be decreasing.

However the above stated does not imply that the succesive addition of a centroid or a new cluster will always be decreasing value for γ. Hence for that the below equation can be used:

All x_i belong to the set of clusters.

$$\sum_{k=1}^{K}(x_i - \mu_k)^2 <= \sum_{k=1}^{K} +1(x_i - \mu_k)^2 \tag{1}$$

Since the addition of a new cluster is the formation of a centroid which is closer to the other data points. The only possible inference is that at every addition of a new cluster the γ will always decrease and hence it is also non-increasing.

(b) Applying the kernel trick to the K-means algorithm.

$\gamma_k = 1/S_k$

Replacing the given α_k equation in the K-means cost function we get,

Assume both the given sum limits are present for all the below equations as defined in the problem.

$$||\phi(x) - \sum_{j=1}^{M} \gamma_{ij}\phi(x_j)||_2^2 \tag{2}$$

Rewriting the equation using the property $||x - y||_2^2 = (x - y).(x - y)$

$$(\phi(x) - \sum_{j=1}^{M} \gamma_{kj}\phi(x_j)).(\phi(x) - \sum_{j=1}^{M} \gamma_{kj}\phi(x_j)) \tag{3}$$

Recall that $k(x, x') = \phi(x).\phi(x')$

Hence,

$$k(x,x) - 2\sum_{j=1}^{M} \gamma_{kj} k(x,x_j) + \sum_{i,j=1}^{M} \gamma_{ki}\gamma_{kj} k(x_i,x_j) \tag{4}$$

The derived equation can be used for the K-means algorithm using the kernel trick. However, this does significantly increase complexity and also makes it difficult to deal with data that has high dimensionality which is a known problem with kernel inspired methods.

Question 2:

(a)

(b)

Question 3:

(a) E-step and M-step updates for Online Gaussian Mixture Models

Estimation step:

Calculation of weights,

$$w_{ik} = \frac{\mathcal{N}(x_i|z_k = 1, \mu_k, \sigma_k)\alpha_k}{\sum_{n=1}^{K} \mathcal{N}(x_i|z_k = 1, \mu_n, \sigma_n)\alpha_n} \tag{5}$$

Maximization step:

Please note that for easier reading the Normal distribution sample data point is simply being represented by N, its sign without the parameters.

$$\mathcal{N} = \mathcal{N}(x_i, \mu, \sigma) \tag{6}$$

The historic information is being preserved using the previous mean while only adding the new value re-calculating it. Same holds true for co-variance and class weights.

This approach will be evaluated m times or dependent upon the number of samples in the dataset. The k variable denotes the Gaussian class.

$$\mu = \frac{\mathcal{N}\alpha_j\mu_j + w_{ik}\mu_k}{\mathcal{N}\alpha_j + w_{ik}} \tag{7}$$

$$\sigma = \frac{\mathcal{N}\alpha_j\sigma_j + w_{ik}\sigma_k}{\mathcal{N}\alpha_j + w_{ik}} + \frac{\mathcal{N}\alpha_j\mu_j\mu_j^t + w_{ik}\mu_k\mu_k^t}{\mathcal{N}\alpha_j + w_{ik}} - \mu\mu^t \tag{8}$$

$$\alpha = \frac{\mathcal{N}\alpha_j + w_{ik}}{\mathcal{N}\alpha_j} \tag{9}$$

(b) Space complexity required is : O(I K m^2)