CS 6362 Machine Learning, Fall 2017: Homework 4

Jiani Li

1)

(a) Assume at some point we have K clusters and the centroids of those clusters are $\{\mu_1, \dots, \mu_K\}$. If clustering converges to the global minimum, then γ_K is the smallest sum of the distance from every point to its closest cluster's centroid. At this convergence point, assume we introduce one more centroid μ_{K+1}. Now, this μ_{K+1} may replace some $\mu_k, k \in \{1, \dots, K\}$ and becomes the closest centroid to some data points. Thus, now the smallest sum of the distance from every point to its closest cluster's centroid will be $\leq \gamma_K$. And we know the optimum sum of distance (γ_{K+1}) after convergence must be $\leq$ the sum of distance at this point. Therefore, $\gamma_{K+1} \leq \gamma_K$. Thus, we prove γ_K is non-increasing.

(b) At E-step, for each x_i, we find α_k that is closest to $\phi(x_i)$. And at M-step, we update α_k.

For E-step:

$$r_{ik} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\phi(x_i) - \alpha_j\|^2 \\ 0 & \text{otherwise} \end{cases}$$

where k is obtained by the following minimization function:

$$\begin{aligned}
&\min_k \|\phi(x_i) - \alpha_k\|_2^2 \\
&= \min_k \|\phi(x_i) - \frac{1}{|S_k|} \sum_{x_j \in S_k} \phi(x_j)\|_2^2 \\
&= \min_k [\phi(x_i) \cdot \phi(x_i) - \frac{2}{|S_k|} \sum_{x_j \in S_k} \phi(x_i) \cdot \phi(x_j) + \frac{1}{|S_k|^2} \sum_{x_j \in S_k} \sum_{x_l \in S_k} \phi(x_j) \cdot \phi(x_l)] \\
&= \min_k [k(x_i, x_i) - \frac{2}{|S_k|} \sum_{x_j \in S_k} k(x_i, x_j) + \frac{1}{|S_k|^2} \sum_{x_j \in S_k} \sum_{x_l \in S_k} k(x_j, x_l)]
\end{aligned}$$

For M-step:

We update α_k as follows:

$$\alpha_k = \frac{\sum_i r_{ik} \phi(x_i)}{\sum_i r_{ik}}$$

2)

(a) At the 2nd step, we update cluster centroids by solving the following optimization function:

$$\mu_k = \arg\min_{\mu_j} \sum_{x_i \in S_k} \|x_i - \mu_j\|_1$$

The solution to the above problem is the median of $x_i \in S_k$. Therefore, at this step we should find the median of the data in the current cluster. Suppose total N examples, the update formula for the 2nd step of the resulting clustering algorithm is:

$$\alpha_k = \begin{cases} x_{\frac{N+1}{2}}, & \text{if } N \text{ is odd} \\ \frac{1}{2}(x_{\frac{N}{2}} + x_{\frac{N}{2}+1}), & \text{if } N \text{ is even} \end{cases}$$

(b) The algorithm updating K centroids of clusters by computing the mean of the data points in the cluster is called K-means. Therefore, the algorithm updating K centroids of clusters by computing the median of the data points in the cluster is called K-medians.

3)

(a) E-step:

For each Gaussian $k = \{1, \dots, K\}$,

$$\gamma(k) = \frac{\pi_k \mathcal{N}(x_i|\mu_k, \sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i|\mu_j, \sigma_j)}$$

M-step:

$$N_{k,\texttt{old}} = N_k \text{ (Initialization: } N_{k,\texttt{old}} = 0)$$

$$N_k = N_k + \gamma(k)$$

$$\pi_k = \frac{N_k}{N}$$

$$\mu_k = \frac{N_{k,\texttt{old}} \cdot \mu_k + \gamma(k) x_i}{N_k}$$

$$\sigma_k = \frac{N_{k,\texttt{old}} \cdot \sigma_k + \gamma(k)(x_i - \mu_k)(x_i - \mu_k)^T}{N_k}$$

where N is the total number of examples being processed. Here, $N = i$.

(b) We have K Gaussians, for each Gaussian, we store seven parameters:

$N(x_i|\mu_k, \sigma_k), \gamma(k), N_{k,\texttt{old}}, N_k, \pi_k, \mu_k, \sigma_k$.

Total memory for Gaussian parameters is: $O(K)$.