CSCI3320: Homework 6, Spring 2020 Deadline: March 29, 2020, 11:59 pm
Teacher: John C.S. Lui

1. *After we apply the k-mean clustering and obtain the clusters, should we partition the training instances as different clusters, and then compute the covariance matrix* $\boldsymbol{S}_i$ *for each cluster (assuming data from each cluster came from a multivariate normal distribution) so that we can do the classification job later? Explain why.*

Answer: This may not be a good idea for two reasons:

(a) First, k-means does hard partitioning but it is always better to do a soft partitioning (using $h_i^t \in (0, 1)$ instead of $b_i^t \in \{0, 1\}$) so that instances (in between two clusters) can contribute to the parameters (the covariance matrix in this case) of more than one cluster allowing a smooth transition between clusters.

(b) k-means uses the Euclidean distance and this implies that Euclidean distance implies features that have the same scale and are independent. Using $\boldsymbol{S}_i$ implies the use of Mahalanobis distance and hence taking care of differences in scale and dependencies.

2. *What are similarities and differences between the average-link clustering and k-means?*

Answer: They both measure similarity by looking at the average of instances that fall in a cluster. Note however that in a hierarchical scheme, there are clusters at different resolutions.

3. *How can we make k-means more robust to outliers ?*

Answer: An outlier is an instance that is very far from all centers. We would not want outliers to affect the solution. One possibility is to not take such instances into account when calculating the parameters—for ex- ample, means and covariances. Note that to detect an outlier, we can use the Mahalanobis distance or the likelihood, but we cannot use the posterior.