CS534 — Homework Assignment 4 *Solution*

Please submit your solution in a single pdf file via TEACH.

1. Prove that the kmeans objective

$$J = \sum_{c=1}^{k} \sum_{x_i \in C_c} |x_i - \mu_c|^2$$

monotonically decreases with each iteration of the Kmeans algorithm.

Solution: *See slide 26 titled "does Kmeans converge?"*

2. **Picking k for Kmeans with J?**. Prove that the minimum of the kmeans objective J is a decreasing function of k (the number of clusters) for $k = 1, ..., n$, where n is the number of points in the dataset. Argue that it is a bad idea to choose the number of clusters by minimizing J.

Solution: *We just need to show that the minimum of $J(k+1)$ is smaller than the minimum of $J(k)$. Consider the solution for minimum $J(k)$. Now let's take an arbitrary point that is not the cluster center of its cluster to be the $k+1$-th cluster center. Such a point will always exist unless $k = n$. Since this new configuration with $k+1$ clusters has not yet converged, we know that the minimum of $J(k+1) \leq$ the current J value. Observe that at least for the point that is now the new cluster center, the term in the current J will be 0 which is decreased from its previous contribution to $J(k)$. This means that the current J value $\leq$ the previous minimum $J(k)$, hence the minimum of $J(k+1)$ must be less than the minimum of $J(k)$.*

If we were to pick the k that minimized J, we would end up picking $k = n$ since this makes $J = 0$. One possible strategy for selecting k is to use a hold-out validation set or cross-validation and choose the k value that minimizes J on the validation set. One could also select k to be the elbow point on the SSE curve, i.e., the k value where the decreasing rate of SSE decreases abruptly.

3. **Gaussian Mixture Models**. Let our data be generated from a mixture of two univariate gaussian distributions, where $f(x|\theta_1)$ is a Gaussian with mean $\mu_1 = 0$ and $\sigma^2 = 1$, and $f(x|\theta_2)$ is a Gaussian with mean $\mu_2 = 0$ and $\sigma^2 = 0.5$. The only unknown parameter is the mixing parameter α (which specifies the prior probability of θ_1.). Now we observe a single sample x_1, please write out the likelihood function of x_1 as a function of α, and determine the maximum likelihood estimation of α.

Solution:

Thus we can write the likelihood function $L(\alpha) = p(x_1|\alpha)$ as:

$$p(x_1|\alpha) = \frac{\alpha}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} + \frac{1-\alpha}{\sqrt{\pi}} e^{-x^2}$$

Consider that a single sample x_1 has been observed. Determine the maximum likelihood estimate of α.

We can write the likelihood as follows:

$$p(x_1|\alpha) = \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_1^2} - \frac{1}{\sqrt{\pi}} e^{-x_1^2} \right) \alpha + \frac{1}{\sqrt{\pi}} e^{-x_1^2}$$

Thus, we see that the likelihood is simply a linear function of alpha where the sign of the slope is determined by which Gaussian produces the larger response. Since we know that $0 \leq \alpha \leq 1$, this tells us that if the slope is positive that we should choose $\alpha = 1$ and otherwise if the slope is negative we should choose $\alpha = 0$. Using straightforward algebra one can show that the slope is positive whenever $x_1^2 \geq \log 2$ and we should set $\alpha = 1$ otherwise set $\alpha = 0$. Alternatively, one could also apply Expectation maximization for this problem (not an efficient solution). Starting with $\alpha = 0.5$ and applying EM, you would observe that in each iteration, your estimate of α will strictly increase or decrease depends on which of the two Gaussians fit x_1 better, eventually lead to 1 or 0 accordingly.

4. **Expectation Maximization for Mixture of Multinomials**
Consider a categorical random variable x with M possible values $1, \cdots, M$. We now represent x as a vector $\mathbf{x}$ such that for $j = 1, ..., M$, $\mathbf{x}(j) = 1$ iff $x = j$. The distribution of $\mathbf{x}$ is described by a mixture of K discrete Multinomial distributions such that:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}|\mu_k)$$

and

$$p(\mathbf{x}|\mu_k) = \prod_{j=1}^{M} \mu_k(j)^{\mathbf{x}(j)}$$

where π_k denotes the prior probability of cluster k, and μ_k specifies the parameters of the kth component. Specifically, $\mu_k(j)$ represents the probabilities $p(\mathbf{x}(j) = 1|z = k)$, and satisfies that $\sum_j \mu_k(j) = 1$. Given an observed data set $\{\mathbf{x}_i\}, i = 1, \cdots, N$, derive the E step and M step for the EM algorithm.

Solution: E-step: *In E-step, we compute the posterior probability of the cluster labels given the current parameters: $\mu_k, \pi_k, k = 1, ..., K$*

$$\begin{aligned} p(z_i = k|\mathbf{x}_i; \theta) &= \frac{p(\mathbf{x}_i|z_i = k; \theta)p(z_i = k|\theta)}{p(\mathbf{x}_i|\theta)} \\ &= \frac{\pi_k p(\mathbf{x}_i|\mu_k)}{\sum_{j=1}^{K} \pi_j p(\mathbf{x}_i|\mu_j)} \\ &= \frac{\pi_k p(\mathbf{x}_i|\mu_k)}{\sum_{j=1}^{K} \pi_j \prod_{l=1}^{M} \mu_j(l)^{\mathbf{x}_i(l)}} \\ &= \frac{\pi_k \prod_{j=1}^{M} \mu_k(j)^{x_i(j)}}{\sum_{j=1}^{K} \pi_j \prod_{l=1}^{M} \mu_j(l)^{\mathbf{x}_i(l)}} \end{aligned}$$

M-step: *Now we can view each example i as K weighted examples, one assigned to each cluster k with weight $P(z_i = k|x_i; \theta)$. We can then re-estimate the parameters $\mu_k, \pi_k, k = 1, ..., K$ for each cluster using weighted MLE esitmation. Note that in the following equation the $P(z_i = k|x_i; \theta)$ is computed in the E-step using the old θ parameters.*

$$\mu_k(l) = \frac{\sum_{i=1}^{N} P(z_i = k|x_i; \theta)\mathbf{x}_i(l)}{\sum_{i=1}^{N} P(z_i = k|x_i; \theta)}$$

Here the numerator tallies up the weights of all of the cluster k examples whose l-element is 1 and the denominator sums up the weights of all cluster k examples.

This can be simply interpreted as among the total mass that was deemed to belong to cluster k (denominator), the portion that had $x(l) = 1$(numerator).

The prior probability of each cluster can be estimated as:

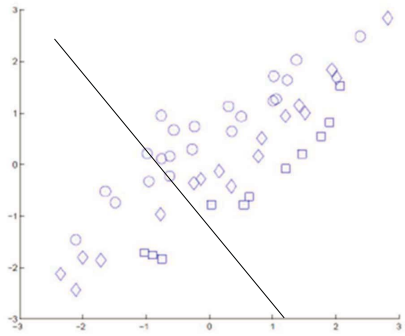
$$\pi_k = \frac{\sum_{i=1}^{N} P(z_i = k|x_i; \theta)}{N}$$

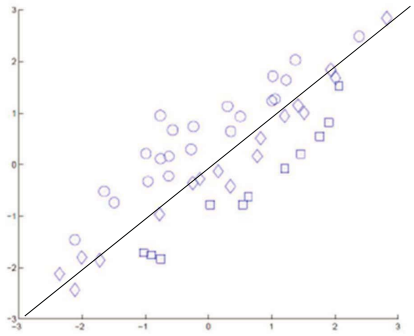
This again can be simply interpreted as among a total of N examples, what proportion is deemed to belong to cluster k.

5. **Dimension reduction.**

   a. Consider the following data set, please draw on the picture the the 1st Principal component direction, and the direction for LDA respectively. Note for PCA, please ignore the markers, and for LDA, we treat the circles as one class and the rest as the other class.

   Answer: *The projection lines by LDA and PCA are shown in the figure below:*


LDA


PCA

b. Given three data points, (0,0), (1,2), (-1, -2) in a 2-d space. What is the first principal component direction (please write down the actual vector)? If you use this vector to project the data points, what are their new coordinates in the new 1-d space? What is the variance of the projected data?
Answer: *One could simply noting that the three points lie on a straight line, which must be the principal component direction. As such, we can simply write out the line* $2x - y = 0$. *The direction vector for this line is simply:* $(\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}})$. *One could also follow the recipe given in class. Begin by forming the covariance matrix:*

$$S = \frac{1}{3}\left(\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} + \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}\right) = \frac{2}{3}\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$$

We now solve for the eigenvectors by solving $S - I\lambda = 0$. *Using the determinant:*

$$\begin{aligned} \begin{vmatrix} 2/3-\lambda & 4/3 \\ 4/3 & 8/3-\lambda \end{vmatrix} &= 0 \\ (2/3-\lambda)(8/3-\lambda) - 16/3 &= 0 \\ \lambda^2 - 10/3\lambda &= 0 \\ \lambda(\lambda - 10/3) &= 0 \\ \Rightarrow \lambda_1 = 0, \lambda_2 &= 10/3 \end{aligned}$$

To recover the first eigenvector we solve the system

$$\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 5x_1 \\ 5x_2 \end{pmatrix}$$

which gives

$$u_1 = \begin{pmatrix} \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{pmatrix}$$

If we project the data by this vector, we get

$$\begin{pmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{pmatrix}\begin{pmatrix} 0 & 1 & -1 \\ 0 & 2 & -2 \end{pmatrix} = \begin{pmatrix} 0 & \sqrt{5} & -\sqrt{5} \end{pmatrix}$$

The variance of this data is

$$\frac{1}{3}\left(0^2 + \sqrt{5}^2 + (-\sqrt{5})^2\right) = \frac{10}{3}$$

Note that above we are using MLE for the variance. If one choose to use an unbiased estimator of the variance (i.e., replacing N with N-1 in the normalization term), the solution for PC does not change but the (co)variance before and after projection will change proportionally, and everything will work out in a similar way.