CS-534

Homework Assignment-4

AKASH AGARWAL
OSU ID: 933-471-097

Ques1 Prove that the k-means objective:

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

monotonically decreases with each iteration of the K-means algo.

Solution: For any data point x_i in the dataset, the K-means algorithm reduces the sum of squared distance of the data point from its corresponding center.

⇒ In the first iteration:

Step1: The objective is to fix the value of mean(μ) and optimize the value of class label.

i.e. $$\min_{\mu, C} \sum_{c=1}^{k} \sum_{x \in C} |x - \mu_c|^2$$

Since, μ is fix, the function tries to minimize the distance of each point to the closest centre.
Thus, it can be observed that in the first step of K-means, the objective i.e. J decreases.

Step2: In this step, K-means algorithm fixes class label C_i & tries to optimize mean μ_i:

$$\Rightarrow \quad \min_{\mu} \sum_{c=1}^{K} \sum_{x \in C} |x - \mu_c|^2 \quad \text{—(1)}$$

Taking partial differential w.r.t. μ

$$\Rightarrow \quad 2 \sum_{x \in C} (x - \mu_c)$$

Equating the above equation to 0

$$\Rightarrow 2 \sum_{x \in C} (x - \mu_c) = 0$$

$$\Rightarrow \mu_c = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

Thus, the algorithm ~~which~~ optimizes eq^n (1) to get mean $= \mu_c$

Since, μ_c lies at the minima

$\Rightarrow$ The objective also decreases in this step.

Hence, for each iteration, the objective function is guaranteed to decrease & since there are finite possible C's

$\therefore$ The algorithm is also guaranteed to converge.

Ques 2. Picking k for Kmeans with J? Prove that the minimum of the kmeans objective J is a decreasing function of k (the no. of clusters) for $k = 1, \ldots, n$ where n is the no. of points in the dataset. Argue that it is a bad idea to choose the no. of clusters by minimizing J.

Solution:- (A) $$J = \sum_{c=1}^{k} \sum_{x_i \in C_i} |x_i - \mu_c|^2$$

$$J_{new} = \sum_{c=1}^{k+1} \sum_{x_i \in C_i} |x_i - \mu_c|^2$$

$$= \sum_{x_i \in C_{k+1}} |x_i - \mu_{c_{k+1}}|^2 + \sum_{c=1}^{k} \sum_{x_i \in C_k} |x_i - \mu_c|^2$$

The new cluster is $\mu_{c_{k+1}}$. There will always be one point which belongs to cluster C_{k+1}. The distance will become zero, becomes the new data point will become new cluster.

$$\Rightarrow \quad x - \mu_{k+1} = 0$$

$$\therefore \quad \sum_{x_i \in C_{k+1}} |x_i - \mu_{k+1}|^2 = 0$$

If there are more than one points that belong to the new cluster, the distance b/w those points & new cluster centre will be less than the distance between the points & the cluster centres in case of k clusters.

$$\Rightarrow \quad J = \sum_{c=1}^{k+1} \sum_{x_i \in C_i} |x_i - \mu_c|^2 < \sum_{c=1}^{k} \sum_{x_i \in C_i} |x_i - \mu_c|^2$$

Thus, as the no. of clusters increase, the loss will decrease.

2(B) It is a bad idea to choose no. of clusters based on mini-
-mizing J. This is because, the value of loss
will be zero if each point becomes a cluster in itself.
This way, if we choose the ~~no.~~ no. of clusters based
on this concept, the model will overfit.

i.e if $J = \sum_{c=1}^{k} \sum_{x_i \in C_i}^{M} |x_i - \mu_c|^2 = 0$

$k = m$, $\forall$ m is the total no. of datapoints.

$\downarrow$

no. of clusters.

Ques3. GMM: let our data be generated from a mixture of two univariate gaussian distributions where $f(x|\theta_1)$ is a Gaussian with mean $\mu_1 = 0$ & $\sigma^2 = 1$, & $f(x|\theta_2)$ is a Gaussian with mean $\mu_2 = 0$ & $\sigma^2 = 0.5$. 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 func. of x_1 as a function of α & determine the MLE of α.

Solution: Acc. to question, the data is generated by two gaussian distributions:

$f(x|\theta_1) \Rightarrow$ Gaussian with $\mu_1 = 0$ & $\sigma_1^2 = 1$

$f(x|\theta_2) \Rightarrow$ Gaussian with $\mu_2 = 0$ & $\sigma_2^2 = 0.5$

If a single datapoint x_i is observed, we can find the likelihood of its belongingness to either of the distribution can be calculated as follows:-

$$L = \prod_{i=1}^{N} P(x_i)$$

Now, $P(x) = P(x|y=0)(P(y=0)) + P(x|y=1)P(y=1)$

$\therefore L = P(x_1|\theta_1)P(\theta_1) + P(x_1|\theta_2)P(\theta_2)$

A.T.Q $P(\theta_1) = \alpha$

$\therefore P(\theta_2) = 1 - P(\theta_1) = 1 - \alpha$

$$\therefore \boxed{L = P(x_1|\theta_1)\alpha + P(x_1|\theta_2)(1-\alpha)}$$

$= [P(x_1|\theta_1) - P(x_1|\theta_2)]\alpha + P(x_1|\theta_2)$

It can be seen that

Since $\alpha = P(\theta_1) \Rightarrow 0 \leq \alpha \leq 1$

If $P(x_1|\theta_1) - P(x_1|\theta_2) \leq 0$

$$\Rightarrow \alpha = 0$$

If $P(x_1|\theta_1) - P(x_1|\theta_2) \geq 0$

$$\Rightarrow \alpha = 1$$

$\therefore P(x_1|\theta_1) \geq P(x_1|\theta_2)$ $\because$ There is a mixing of two gaussians & hence $\alpha \neq 0$.

$$\Rightarrow \frac{1}{\sqrt{2\pi}\,\sigma_1} e^{-\left(\frac{(x_1-\mu_1)^2}{2\sigma_1^2}\right)} \geq \frac{1}{\sqrt{2\pi}\,\sigma_2} e^{-\left(\frac{(x_1-\mu_2)^2}{2\sigma_2^2}\right)}$$

$$\Rightarrow \frac{1}{\sqrt{2\pi}} e^{-\frac{x_1^2}{2}} \geq \frac{1}{\sqrt{2\pi \times 0.5}} e^{\frac{-x_1^2}{2\times 0.5}}$$

$$\Rightarrow e^{-\frac{x_1^2}{2}} \geq \sqrt{2}\, e^{-x_1^2}$$

Taking log likelihood of the above equation:-

$$-\frac{x_1^2}{2} \geq \log\sqrt{2} - x_1^2$$

$$\Rightarrow \frac{x_1^2}{2} \geq \log\sqrt{2}$$

$\therefore$ for $[P(x_1|\theta_1) - P(x_1|\theta_2)] \geq 0$

$\boxed{x_1^2 \geq 2\log\sqrt{2}}$; when $\alpha = 1$

& $[P(x_1|\theta_1) - P(x_1|\theta_2)] < 0$, otherwise ; when $\alpha = 0$

Ques 4. Consider a categorical random variable x with M possible values $1, \ldots, M$. We now represent x as a vector x such that for $j = 1, \ldots M$; $x(j) = 1$ iff $x = j$. The distribution of x is described by a mixture of K discrete Multinomial distributions such that:

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

$$\& \quad p(x \mid \mu_k) = \prod_{j=1}^{M} \mu_k(j)^{x(j)}$$

where π_k denotes the prior probability of cluster k, & μ_k specifies the parameters of the k^{th} component. Specifically, $\mu_k(j)$ respectively represents the probabilities $p(x(j) = 1 \mid z = k)$ & satisfies that $\sum_j \mu_k(j) = 1$. Given an observed data-set $\{x_i\}, i = 1, \ldots, N$ derive the E step & M step for the EM Algorithm.

Let the output z be y. $\therefore$ $y = \underbrace{1, 2 \ldots k}_{k \text{ classes}}$.

Solution

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

$$\& \; p(x \mid \mu_k) = \prod_{j=1}^{M} \mu_k(j)^{x(j)}$$

$M \to 1 \ldots M$ $\quad$ $j \to 1, \ldots M$ $\quad$ $x(j) = 1$ iff $x = j$

Here, $\pi_k \Rightarrow$ prior probability of cluster k

$\mu_k \to$ params of k^{th} component

The observed data is given as $\{x_i\}$ $\forall$ $i = 1, \ldots N$

E-Step: $$P(y_i = k \mid x_i) = \frac{P(x_i \mid y_i = k) \, P(y_i = k)}{P(x_i)}$$

$$P(y_i = k \mid x_i) = \frac{P(x_i \mid y_i = k) \, P(y_i = k)}{P(x_i)}$$

$$\Rightarrow \frac{\pi_k P(x_i|\mu_k)}{\sum_{j=1}^{k} \pi_j P(x_i|\mu_j)} = \frac{\pi_k \prod_{j=1}^{M} \mu_k(j)^{x_i(j)}}{\sum_{j=1}^{k} \pi_j \prod_{l=1}^{M} \mu_j(l)^{x_i(l)}}$$

M-Step :

Generally, $\underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N}\sum_{j=1}^{K} P(y_i=j|x_i) \log \frac{P(x_i|y=j)}{P(y_i=j|x_i)}$

$$= \sum_{i=1}^{N}\sum_{j=1}^{K} P(y_i=j|x_i) \log P(x_i|y_i=j)P(y_i=j)$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{K} P(y_i=j|x_i) \log \prod_{k=1}^{M} \mu_j(k)^{x_i(k)} \pi_j$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{K} \{P(y_i=j|x_i)\left(\log \pi_i + \sum_{k=1}^{M} x_i(k)\log \mu_j(k)\right) \quad \text{— (1)}$$

Keeping π_t terms in eq^n (1)

$$\sum_{i=1}^{N} P(y_i=t|x_i)\log \pi_t$$

Since it is a constraint eq^n, with constraint

$$\sum_{j=1}^{K} \pi_j = 1$$

$$L(\pi_t) = \sum_{i=1}^{N} P(y_i=t|x_i)\log \pi_t + \alpha\left(\sum_{j=1}^{K}\pi_j - 1\right)$$

$$\frac{\partial L(\pi_t)}{\partial \pi_t} = \sum_{i=1}^{N} P(y_i=t|x_i)\cdot\frac{1}{\pi_t} + \alpha = 0$$

$$\pi_t = -\frac{1}{\alpha}\sum_{i=1}^{N} P(y_i = t \mid x_i) \quad \text{--- (II)}$$

Now we know:

$$\sum_{j=1}^{K} \pi_j = 1$$

$$\sum_{j=1}^{K}\left(\frac{-1}{\alpha}\sum_{i=1}^{N} P(y_i - j \mid x)\right) = 1$$

$$\Rightarrow -\alpha = \sum_{j=1}^{K}\sum_{i=1}^{N} P(y_i = j \mid x) = N$$

On solving for eq.^{n} (II)

$$\hat{\pi}_t = \sum_{i=1}^{N} \frac{P(y_i = t \mid x_i)}{N}$$

Keeping μ_t terms in eq.^{n} (I)

$$\sum_{i=1}^{N} \{P(y_i = t_k) \sum_{k=1}^{m} x_i(k) \log \mu_t(k)$$

Using Lagrangian

$$\sum_{j=1}^{n} \mu_t(j) = 1$$

$$\Rightarrow L(\mu_t) = \sum_{i=1}^{N} (P(y_i = t \mid x_i) \sum_{k=1}^{m} x_i(k) \log [\mu_t(k)) + b(\sum_{j=1}^{m} \mu_t(j) - 1)$$

differentiating wrt $\mu_t(j)$ & keeping to 0.

$$\mu_t(k) = [\sum_{i=1}^{N} P(y_i = t \mid x_i) \cdot x_i(k) \times \frac{1}{-\alpha} \quad \text{--- (III)}$$

We know that $\sum_{j=1}^{n} \mu_t(j) = 1$

$$\Rightarrow \sum_{j=1}^{M} \sum_{i=1}^{N} P(y_i = t \mid x_i)\, x_i(j) \times \frac{1}{-\alpha}$$

$$\Rightarrow -\alpha = \sum_{j=1}^{M} \sum_{i=1}^{N} P(y_i = t \mid x_i)\, x_i(j)$$

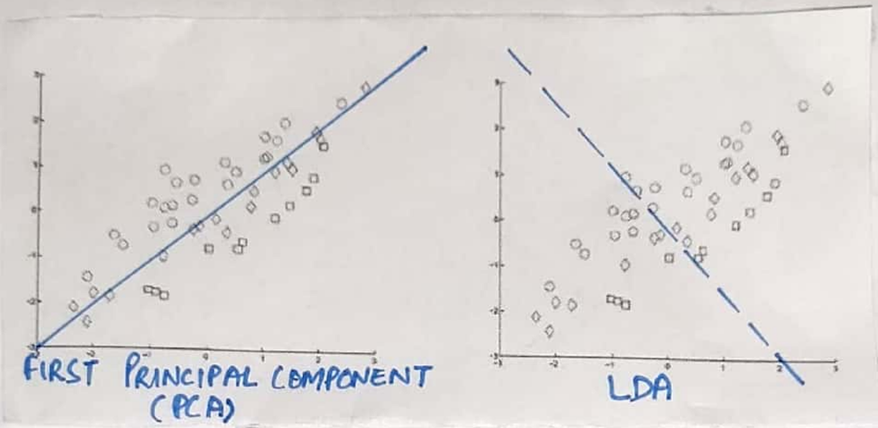
$\therefore$ In eq^n $$\mu_t(k) = \frac{\sum_{i=1}^{N} P(y_i = t \mid x_i)\, x_i(k)}{\sum_{j=1}^{M} \sum_{i=1}^{N} P(y_i = t \mid x_i)\, x_i(j)}$$

$$\Rightarrow \sum_{j=1}^{n} x_j(j) = 1$$

$$\mu_t(k) = \sum_{i=1}^{N} \frac{P(y_i = t \mid x_i)\, x_i(k)}{\sum_{i=1}^{N} P(y_i = t \mid x_i)}$$

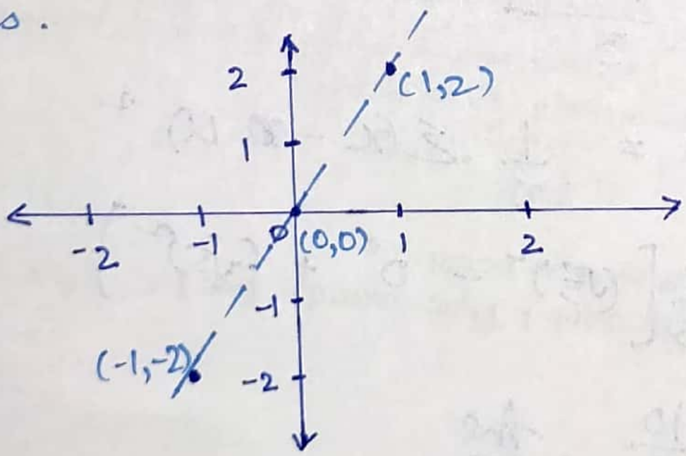
Ques 5 (A) Consider the following dataset, please draw on the picture the 1^{st} Principal component direction, & the direction for LDA respectively. Note for PCA, please ignore the markers & for LDA, we treat the circles as one class & the rest as the other class.

Solution:



Ques 5 (B) Given three data points, (0,0); (1,2); (-1,-2) in a 2-D space. What is the first principal component direction? 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?

Solution: let us draw the given points on the coordinate axes.



Points: (0,0); (1,2); (-1,-2)

$\therefore$ eq^n of line: $y - 2x = 0$

$\Rightarrow \boxed{y = 2x}$

The first Principal component should be along the direction of line $y = 2x$.

Let the Vector be w

$$\therefore \; w \equiv \left(\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}} \right)^T$$

We'll use the above w to project the data points in the 1-D space.

i.e. $w^T x$

for every datapoint :

$w^T x$ for $(-1,-2)$ $\Rightarrow$ $\begin{bmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{bmatrix} \begin{bmatrix} -1 \\ -2 \end{bmatrix} = -\sqrt{5}$

$w^T x$ for $(1,2)$ $\Rightarrow$ $\begin{bmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \sqrt{5}$

$w^T x$ for $(0,0)$ $\Rightarrow$ $\begin{bmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} = 0$

$\therefore$ 1-D Space $\Rightarrow$

$$\text{Mean } (\mu) = \frac{\sqrt{5} + 0 - \sqrt{5}}{3} = 0$$

$$\text{Variance } (\sigma^2) = \frac{1}{n} \Sigma (x - \mu)^2$$

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

$$= \frac{10}{3} \quad \underline{\text{Ans}}.$$