CS 534 Machine Learning
Written Assignment 4
Aashish Adhikari

New Link
Date
Page

Q.1 Solution

We have our kmeans objective

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

Proving that it monotonically decreases means that we need to prove that it continuously decreases and eventually converges.

We know that if there are n data points, then we can have only k^n ways to make k clusters. During each new iteration, a new clustering is generated based on the original one. There are two possible cases for this: either the old and the new clustering are the same or the new clustering has a cost that is lower than the original. It is because each vector is assigned to the closest centroid, so the distance it contributes to the new cost decreases. Also, the cost decreases in the computation step because the new centroid is a vector for which the cost reaches its minimum.

Since we have a finite set of possible clusterings, it will eventually arrive at a minimum.

We know that taking the partial derivative of the cost wrt the m^{th} components of the vector μ

$$\frac{\partial(J)}{\partial \mu_{im}} = \sum_{x_i \in C_c} 2(\mu_m - x_m)$$

Setting it to zero,

New Link
Date
Page

$$u_m = \frac{1}{|C_i|} \sum_{x_i \in C_i} x_m$$

Thus, we minimize the cost when the old centroid is replaced with the new centroid. Hence the sum of different components of the cost, J, must also decrease while recomputing the values.

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

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

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

In this newe iteration, there is a point which belongs to C_{k+1}. for this particular point, the distance from itself to the centroid is zero. i.e., $|x - \mu_{k+1}| = 0$

In case more than one point are in this new cluster, the distance between the points and the new centroid will be less than the distance between the points and the centroid in the previous iteration k.

This implies

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

Hence, the increase in the number of clusters, ~~the~~ results in a decrease in the objective J.

⇒ Since the increase in the number of clusters will always reduce the distance to the datapoints, increasing k will always decrease this metric, that is, it reaches upto zero when its value is equal to the no. of data points. Hence, it is a BAD IDEA to choose the number of ~~param~~ clusters by minimizing J.

New Link
Date
Page

A solution to this is elbow method.
-Mean distance to the centroid as a function of k is plotted and the point, also called the elbow point, where the rate of decrease sharply shifts can be roughly used to determine which k to use. An example of the plot is shown below:

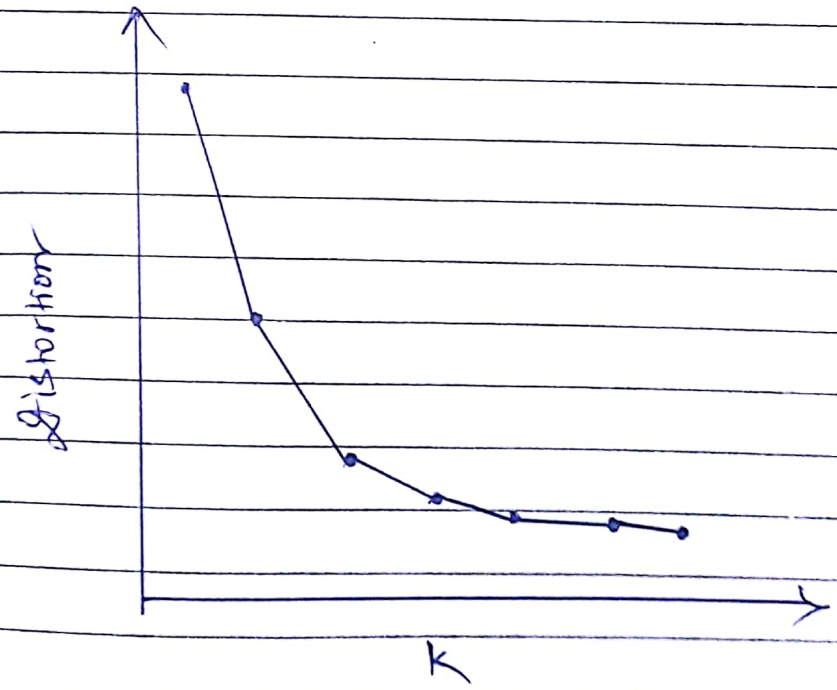


fig plot for distortion vs k

Q3. We have two univariate Gaussian distributions:

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

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

Then,

The mixture gives us

$$f(x) = \sum_{i=1}^{k} f(x, y=i)$$

$$= \sum_{i=1}^{k} f(x|y=i)\, f(y=i)$$

$$= \sum_{i=1}^{k} \alpha_i f(x|\theta_i)$$

In our case since we have only θ_1 & θ_2

$$= \alpha f(x|\theta_1) + (1-\alpha) f(x|\theta_2)$$

where α & $(1-\alpha)$ are class priors.

& $\theta_i = \{\mu_i, \Sigma_i\}$

The log likelihood is

$$l(\theta; x_n) = \sum_{i=1}^{k} \log \left[\alpha f(x|\theta_1) + (1-\alpha) f(x|\theta_2) \right]$$

Direct maximization of $l(\theta|x)$ is difficult because of the sums of logarithmic terms.

So, we consider unobserved latent variables Δ_i.

$$l(\theta; x, \Delta) = \sum_{i=1}^{k} \left[(1-\Delta) \log f(x|\theta_2) + \Delta \log f(x|\theta_1) \right] +$$

$$\sum_{i=1}^{k} \left[(1-\Delta) \log \alpha + \Delta \log (1-\alpha) \right]$$

Since the value of Δ is unknown, we iterate the value of Δ with its expected value.

$$\gamma_i(\theta) = E(\Delta_i | \theta, x) = P(\Delta_i = 1 | \theta, x)$$

- Take initial guesses for parameters

$$\hat{\mu_1}, \hat{\sigma_1^2}; \hat{\mu_2}, \hat{\sigma_2^2}, \hat{\alpha}$$

- Expectation step: Compute the responsibilities.

$$\hat{\gamma_i} = \frac{\hat{\alpha} f(y_i)}{(1-\hat{\alpha}) f_{\hat{\theta}_1}(y_i) + \hat{\alpha} f_{\hat{\theta}_2}(y_i)}, \quad i = 1, 2, \ldots k$$

where $f_{\hat{\theta}_i}(y_i) = f_{\hat{\theta}_i}(x|\theta_i)$

- Maximization: compute the weighted mean & variances & the mixing probability

New Link
Date
Page

$$\Rightarrow \hat{\alpha} = \sum_{i=1}^{K} \frac{\hat{\alpha}_i}{K}$$

which is the required answer. #

Q4. We are given $p(x) = \sum_{k=1}^{K} \pi_k p(x|\mu_k)$

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

$j = 1 \ldots M$

$x(j) = 1$ if $x = j$

E step:

$$p(y_i = k|x_i) = \frac{p(x_i|y_i = k)\, p(y_i = k)}{p(x_i)}$$

$$= \frac{\pi_k \cdot p(x_i|\mu_k)}{\sum_{j=1}^{K} \pi_j p(x_i|\mu_j)}$$

$$= \frac{\pi_k \prod_{i=1}^{M} \mu_k(j)^{x(j)}}{\sum_{j=1}^{K} \pi_j \prod_{l=1}^{M} \mu_j(l)^{x_i(l)}}$$

M step:

$$\arg\max_{\theta} \sum_{i=1}^{N} \sum_{j=1}^{K} p(y_i = j|x_i) \log \frac{p(x_i|y=j)}{p(y_i = j|x_i)}$$

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

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

$$= \sum_{i=1}^{N} \sum_{j=1}^{K} p(y_i = j | x_i) \left(\log \pi_j + \sum_{k=1}^{M} x_i(k) \log u_j(k) \right)$$

This is our first equation (A)

Taking only π_t terms:

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

Now, we can use Lagrangian with a constraint $\sum_{j=1}^{K} \pi_j = 1$

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

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

$$\pi_t = -\frac{1}{\beta} \sum_{i=1}^{N} p(y_i = t | x_i) \quad \text{(B)}$$

we know that $\sum_{j=1}^{K} \pi_j = 1$

$$\sum_{j=1}^{K} \left(-\frac{1}{\beta} \sum_{i=1}^{N} p(y_i = j | x) \right) = 1$$

$$-\beta = \sum_{j=1}^{K} \sum_{i=1}^{N} p\left(y_i = j \mid x_i\right) = N$$

plugging the values back in (B)

$$\pi_t = \frac{\sum_{i=1}^{N} p\left(y_i = t \mid x_i\right)}{N} \qquad \text{(C)}$$

Taking the terms of μ_t from (A), we get

$$\sum_{i=1}^{N} \left\{ p(y_i = t_k) \sum_{k=1}^{M} x_i(k) \log \mu_t(k) \right.$$

Using the Lagrangian

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

$$\mathcal{L}(\mu_t) = \sum_{i=1}^{N} \left(p\left(y_i = t_k\right) \sum_{k=1}^{M} x_i(k) \log \mu_t(k) \right) + p\left(\sum_{j=1}^{M} \mu_t(j) - 1 \right)$$

Taking derivative wrt $\mu_t(k)$

$$\frac{\partial}{\partial \mu_t(k)} = \sum_{i=1}^{N} \left(p(y_i = t \mid x_i) \cdot x_i(k) \frac{1}{\mu_t(k)} \right) + \beta = 0$$

New Link
Date
Page

$$u_t(k) = \left[\sum_{i=1}^{N} p(y_i = t | x_i) \cdot x_i(k) \right] \frac{1}{-\beta} \quad \text{---(A)}$$

We know $\sum_{j=1}^{N} u_t(j) = 1$

This gives,

$$\sum_{j=1}^{N} \sum_{i=1}^{N} p(y_i = t | x_i)\, x_i(j) \frac{1}{-\beta} = 1$$

or, $-\beta = \sum_{j=1}^{N} \sum_{i=1}^{N} p(y_i = t | x_i)\, x_i(j)$

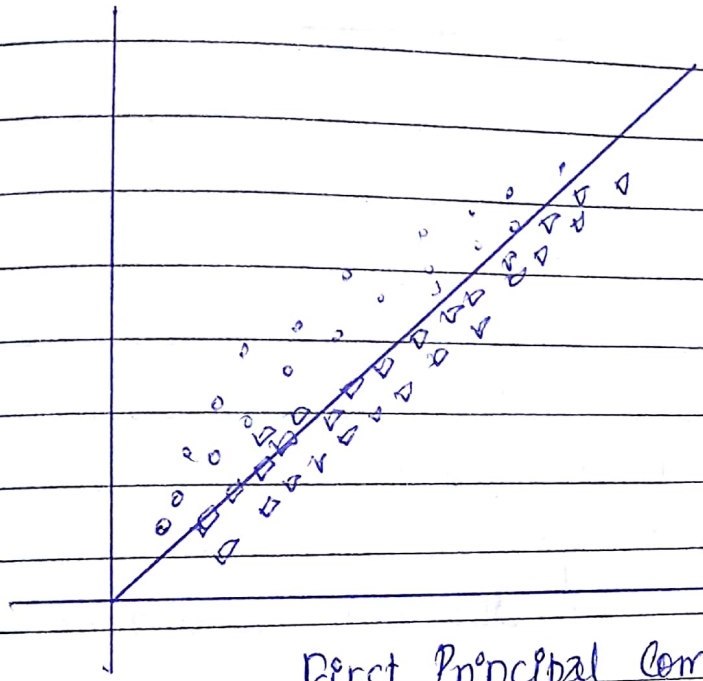
From (A)

$$u_t(k) = \frac{\sum_{i=1}^{N} p(y_i = t | x_i)\, x_i(k)}{\sum_{j=1}^{N} \sum_{i=1}^{N} p(y_i = t | x_i)\, x_i(j)}$$

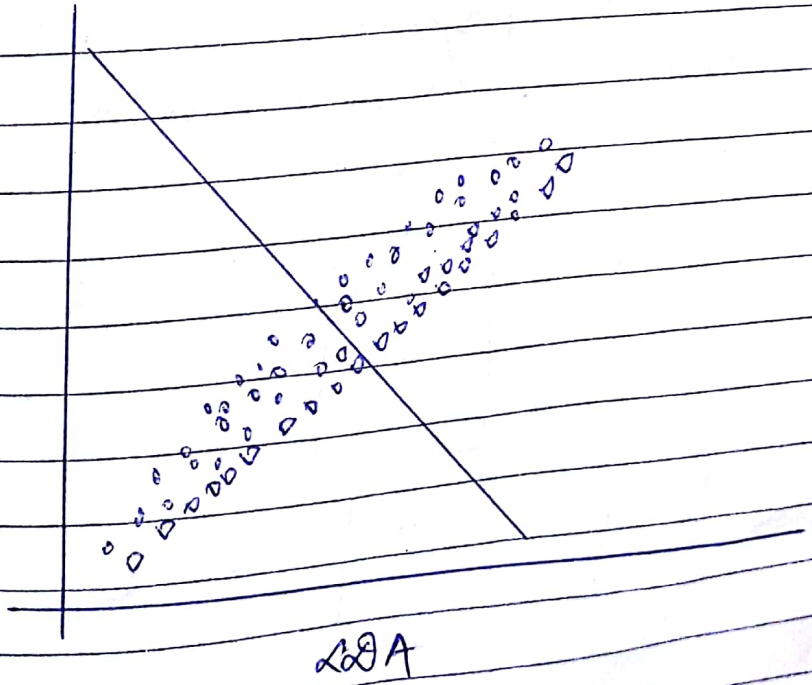
However, $\sum_{j=1}^{N} x_i(j) = 1$

$$\Rightarrow u_t(k) = \frac{\sum_{i=1}^{N} p(y_i = t | x_i)\, x_i(k)}{\sum_{i=1}^{N} p(y_i = t | x_i)} \quad \#$$

Q.5.a)

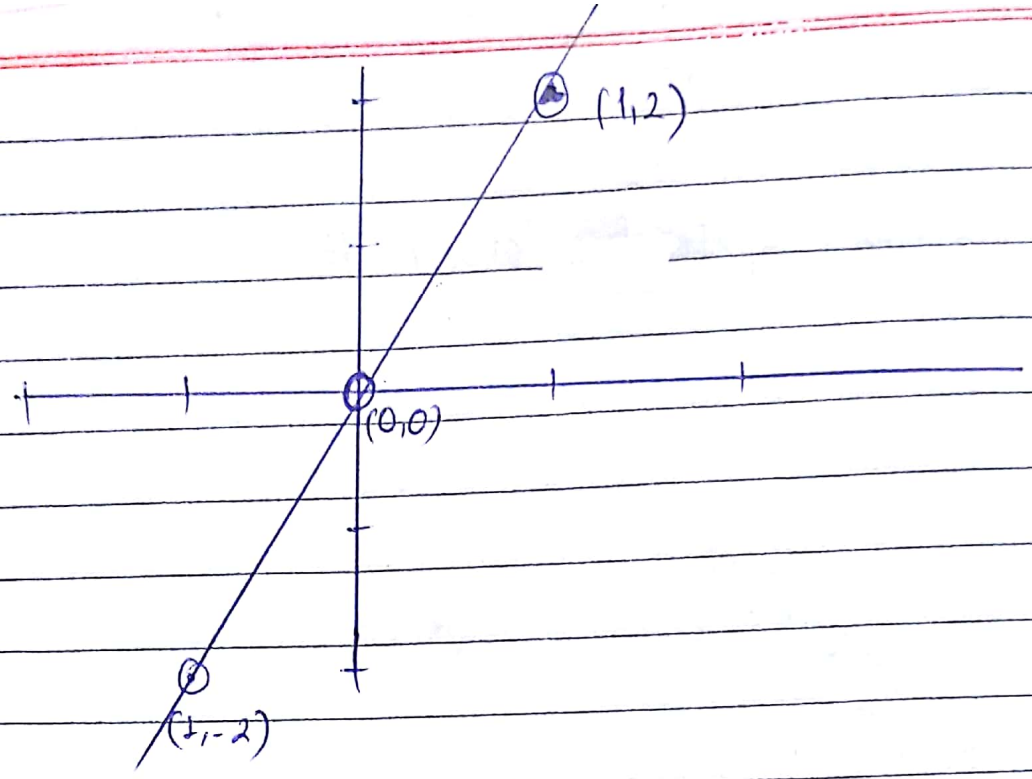


First Principal Component Direction



LDA

b)



Our data points are $(0,0), (1,2), (-1,-2)$.

Equation of the line is $y - 2x = 0$

First Principal Component direction would be along this line $y - 2x = 0$

$$w = \left(1/\sqrt{5} \cdot 2/\sqrt{5} \right)^T$$

This vector is used to project the data points into a new one-dimensional space as $w^T x$.

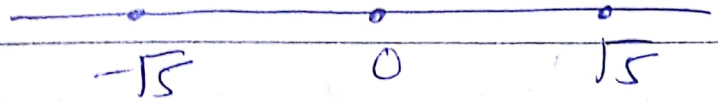
$$\Rightarrow \begin{bmatrix} 1/\sqrt{5} & 2/\sqrt{5} \end{bmatrix} \begin{bmatrix} -1 \\ -2 \end{bmatrix} = -\sqrt{5}$$

$$\Rightarrow \begin{bmatrix} 1/\sqrt{5} & 2/\sqrt{5} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} = 0$$

New Link
Date
Page

$$\Rightarrow \begin{bmatrix} 1/\sqrt{5} & 2/\sqrt{5} \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \sqrt{5}$$

↓ In 1-D space



Variance of the projected data $= \frac{1}{N} \Sigma (x-\mu)^2$

$\mu = (\sqrt{5} + 0 + (-\sqrt{5}))/3$

$= 0$

$\therefore \quad \sigma^2 = \frac{1}{3} \times \Sigma (x-\mu)^2$

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

$= 10/3$ #