Homework 4

Graded

Student

Jinzhi Shen

Total Points

70.5 / 85 pts

Question 1

1 **17** / 20 pts

1.1 **a.1** **1** / 1 pt

✔ **– 0 pts** Correct

– 1 pt Wrong answer

– 1 pt If answer is false, missing or incorrect explanation

1.2 **a.2** **1** / 1 pt

✔ **– 0 pts** Correct

– 1 pt Wrong answer

– 1 pt If answer is false, missing or incorrect explanation

1.3 **a.3** **1** / 1 pt

✔ **– 0 pts** Correct

– 1 pt Wrong answer

– 1 pt If answer is false, missing or incorrect explanation

1.4 **a.4** **1** / 1 pt

✔ **– 0 pts** Correct

– 1 pt Wrong answer

– 1 pt If answer is false, missing or incorrect explanation

1.5 **a.5** **1** / 1 pt

✔ **– 0 pts** Correct

– 1 pt Wrong answer

– 1 pt If answer is false, missing or incorrect explanation

1.6 **b** **2** / 2 pts

✔ **– 0 pts** Correct

– 1 pt Not mentioning orthogonal

– 1 pt Not mentioning inverse of orthogonal matrix equal to its transpose

1.7 **c** **4** / 4 pts

✔ **– 0 pts** Correct

– 1 pt Missing or incorrect plot

– 3 pts Completely incorrect w

– 1 pt Forgetting to normalize w or normalize wrongly

– 1 pt Wrong signs. Note that PCA signs can be reversed, i.e. w and -w both work. But sign is wrong if only one dimension is reversed but the other isnt.

1.8 **d** **4** / 5 pts

– 0 pts Correct

– 1 pt wrong or missing plot

– 1 pt Not centering data

✔ **– 1 pt** Not diving by 4

– 1 pt Wrong final answer

– 4 pts Incorrect formula or approach

1.9 **e** **2** / 4 pts

– 0 pts Correct

– 2 pts wrong eigenvector or computing the eigenvector instead of getting it by inspection(The rubric asks us to take 2 points off if the student is computing the eigenvector. But I think that might be too strict. So I took 1/2 points off.)

✔ **– 2 pts** wrong optimal value

Question 2

2 **17** / 20 pts

2.1 **a** **1** / 2 pts

– **0 pts** Correct

– **1 pt** Did not take log

– **1 pt** Slight mistakes like wrong sign or forgetting subscripts like u_k

– **2 pts** Incorrect

✔ – **1 pt** Did not expand out $\mathcal{N}$

2.2 **b** **3** / 3 pts

✔ – **0 pts** Correct

– **1 pt** Wrong μ

– **1 pt** Wrong σ^2

– **1 pt** Wrong π

2.3 **c** **1** / 2 pts

– **0 pts** Correct

✔ – **1 pt** wrong distribution name

– **1 pt** Wrong solution

2.4 **d** **1** / 2 pts

– **0 pts** Correct

– **2 pts** Incorrect

✔ – **1 pt** Minor errors like missing subscripts

2.5 **e** **3** / 3 pts

✔ – **0 pts** Correct

– **1 pt** Missing or wrong condition

– **1 pt** Missing or wrong condition

– **1 pt** Missing or wrong condition

2.6 **f** **4** / 4 pts

✔ – **0 pts** Correct

– **4 pts** Incorrect or missing proof

– **2 pts** Minor errors

– **2 pts** Wrong or missing explanation on how as $\epsilon \rightarrow 0,$ we get min_k

2.7 g **4** / 4 pts

✔ **– 0 pts** Correct

– 4 pts Incorrect

– 2 pts Missing steps. eg not writing down GMM objective, taking limit etc

– 2 pts Not relating to hard assignment of kmeans

– 1 pt Minor errors

Question 3

3 **19** / 25 pts

3.1 **a.i** **1** / 1 pt

✔ **– 0 pts** Correct

– 1 pt incorrect

3.2 **a.ii** **2** / 2 pts

✔ **– 0 pts** Correct

– 2 pts Incorrect

– 1 pt Minor errors

3.3 **a.iii** **1** / 1 pt

✔ **– 0 pts** Correct

– 1 pt Incorrect

3.4 **b.i** **1** / 1 pt

✔ **– 0 pts** Correct

– 1 pt Incorrect

3.5 **b.ii** **0** / 2 pts

– 0 pts Correct

✔ **– 2 pts** Incorrect

3.6 **b.iii** **3** / 3 pts

✔ **+ 3 pts** Correct

+ 0 pts Incorrect proof

+ 1 pt Correct eq 2 in solution

+ 1 pt Correct eq 3 in solution

3.7 **b.iv** **4** / 4 pts

✔ **– 0 pts** Correct

– 4 pts Incorrect

– 1 pt Minor mistakes

3.8 **b.v** **4** / 4 pts

✔ **– 0 pts** Correct

– 4 pts Incorrect

– 1 pt Minor errors

3.9 **c.i** **3** / 3 pts

✔ **– 0 pts** Correct

– 3 pts Incorrect or not taking derivative

– 1 pt minor errors

– 2 pts Correct approach but wrong solution

3.10 **c.ii** **0** / 4 pts

– 0 pts Correct

✔ **– 4 pts** Incorrect

– 2 pts Wrong lagrangian

– 1 pt Wrong π_k

– 1 pt Wrong λ

– 3 pts Wrong solution but correct approach

Question 4

4 **5** / 5 pts

4.1 **a** **1** / 1 pt

✔ **– 0 pts** Correct

– 1 pt Incorrect

4.2 **b** **1** / 1 pt

✔ **– 0 pts** Correct

– 1 pt Incorrect

4.3 **c** **1** / 1 pt

✔ **– 0 pts** Correct

– 1 pt Incorrect

4.4 **d** **1** / 1 pt

✔ **– 0 pts** Correct

– 1 pt Incorrect

4.5 **e** **1** / 1 pt

✔ **– 0 pts** Correct

– 1 pt Incorrect

– 0.5 pts Incorrect explanation

Question 5

5 **12.5** / 15 pts

5.1 **a** **1** / 1 pt

✔ **– 0 pts** Correct

– 1 pt Incorrect

5.2 **b** **2** / 2 pts

✔ **– 0 pts** Correct

– 2 pts Incorrect

– 1 pt Minor errors like subscripts

5.3 **c** **2** / 2 pts

✔ **– 0 pts** Correct

– 2 pts Incorrect

– 1 pt Minor errors like subscripts

5.4 **d** **4** / 5 pts

– 0 pts Correct

– 1 pt Wrong pseudocode

– 1 pt Says not guaranteed to converge

✔ **– 1 pt** Wrong reasons for convergence guarantee or insufficient reason. It is important to state that domain is finite.

– 1 pt Says guaranteed global convergence

– 1 pt Wrong or insufficient reasons for global convergence gurantee

5.5 **e** **1** / 1 pt

✔ **– 0 pts** Correct

– 1 pt Incorrect

5.6 **f** **0** / 1 pt

– 0 pts Correct

✔ **– 1 pt** Incorrect

5.7 g **2.5** / 3 pts

– **0 pts** Correct

– **1 pt** Wrong number of steps

– **1 pt** Wrong cost value

– **1 pt** Wrong cluster centers

✔ – **0.5 pts** Forgetting to round the cluster centers

Questions assigned to the following page: 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, and 1.7

Homework 4

Question 1

(a)

1. False. PCA minimizes the sum of squared distances (not vertical distance) to the subspace.
2. True. Reconstruction error is the distance between points and the subspace. Since PCA minimizes the sum of squared distances (not vertical distance) to the subspace, it minimizes the least squared reconstruction error.
3. False. Principal components are orthogonal to each other, so they are linearly independent with each other.
4. False. Solving PCA using SVD results in the same solution as original PCA.
5. False. If all eigenvectors are used, points have zero distance from the subspace, and therefore there is no reconstruction error. In this case, PCA guarantees perfect reconstruction error.

(b)

Since W is the eigenvectors found by PCA, corresponding to its definition, we have

$$WW^T = I$$

Therefore, we have

$$ZW^T = XWW^T = XI = X$$

which means we can recover X by doing $X = ZW^T$

(c)

The plot is shown in Figure 1. The first principal component w of this dataset is $w = (\frac{1}{\sqrt{50}}, \frac{-7}{\sqrt{50}})$

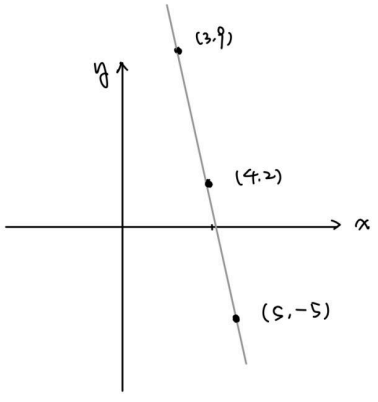

Figure 1: Plot of the points for 1(c).

Questions assigned to the following page: 2.1, 2.2, 2.3, 1.7, 1.8, and 1.9

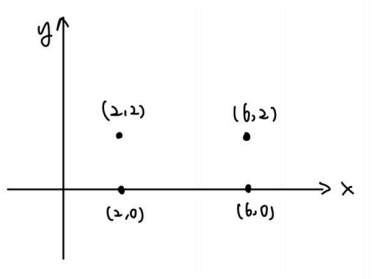

Figure 2: Plot of the points for 1(d).

(d)

The plot is shown in Figure 2. The dimension of the covariance matrix Σ is 2. $\Sigma = \begin{bmatrix} 16 & 0 \\ 0 & 4 \end{bmatrix}$

(e)

$$det\Sigma = (1-\lambda)(6-\lambda)(5-\lambda)(10-\lambda)$$

So there are four eigenvalues: $1, 5, 6, 10$, where the largest eigenvalue is 10. With $\Sigma w = 10w$, the eigenvector corresponding to the largest eigenvalue 10 is $[0, 0, 0, 1]^T$ and the optimal value of the program in Eq.(1) is 16.

Question 2

(a)

According to the Gaussian Mixture Model,

$$p(\mathcal{D}) = \prod_{i=1}^{N} p(x_i) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i|\mu_k, \sigma_k)$$

therefore, the log-likelihood of the data is

$$log\ p(\mathcal{D}) = \sum_{i=1}^{N} log \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_i|\mu_k, \sigma_k) \right)$$

(b)

If $K = 1$, then $\pi_1 = 1$, and we have

$$log\ p(\mathcal{D}) = \sum_{i=1}^{N} log\ \mathcal{N}(x_i|\mu_1, \sigma_1) = \sum_{i=1}^{N} \left(-\frac{1}{2} log\ (2\pi\sigma_1^2) - \frac{(x_i - \mu_1)^2}{2\sigma_1^2} \right)$$

Optimality condition:

$$\frac{\partial log\ p(\mathcal{D})}{\partial \mu_1} = \sum_{i=1}^{N} \frac{2(x_i - \mu_1)}{2\sigma_i^2} = 0 \quad \Longrightarrow \quad \mu_1 = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\frac{\partial log\ p(\mathcal{D})}{\partial \sigma_1} = \sum_{i=1}^{N} \left(-\frac{1}{\sigma_1} + \frac{(x_i - \mu_1)^2}{\sigma_1^3} \right) = 0 \quad \Longrightarrow \quad \sigma_1^2 = \frac{\sum_{i=1}^{N} (x_i - \mu_1)^2}{N}$$

Therefore, the maximum likelihood estimate for the parameters are $(\mu_1, \sigma_1^2, \pi_1) = (\frac{\sum_{i=1}^{N} x_i}{N}, \frac{\sum_{i=1}^{N} (x_i - \mu_1)^2}{N}, 1)$

(c)

$$p(z_{i,1}, z_{i,2}, ..., z_{i,K}) = \prod_{k=1}^{K} \pi_k^{z_{i,k}}$$

Its name is marginal distribution

Questions assigned to the following page: 2.4, 2.5, 2.6, and 2.7

(d)

$$p(z_{ik}=1|x_i)=\frac{p(x_i|z_{ik}=1)p(z_{ik}=1)}{\sum_{\hat{k}=1}^{K}p(x_i|z_{i\hat{k}}=1)p(z_{i\hat{k}}=1)}$$

(e)

1. These two methods are both unsupervised clustering methods.
2. k-Means uses hard assignments of a sample to its cluster, while Gaussian Mixture Model uses soft assignments of a sample to the clusters.
3. They both have cluster centers μ_k, and when the $\sigma_k=\epsilon$ are the same for all clusters, these two methods have the same cost functions in the limit of $\epsilon\to 0$

(f)

$$\lim_{\epsilon\to 0}-\epsilon log\sum_{k=1}^{K}exp(-F_k/\epsilon)=\lim_{\epsilon\to 0}\frac{log\sum_{k=1}^{K}exp(-\frac{F_k}{\epsilon})}{-\frac{1}{\epsilon}}$$

which has the form of $\frac{\infty}{\infty}$, so we can use l'Hopital rule as follow:

$$\lim_{\epsilon\to 0}\frac{log\sum_{k=1}^{K}exp(-\frac{F_k}{\epsilon})}{-\frac{1}{\epsilon}}=\lim_{\epsilon\to 0}\frac{\sum_{k=1}^{K}exp(-\frac{F_k}{\epsilon})\cdot\frac{F_k}{\epsilon^2}/\sum_{k=1}^{K}exp(-\frac{F_k}{\epsilon})}{\frac{1}{\epsilon^2}}=\lim_{\epsilon\to 0}\frac{\sum_{k=1}^{K}exp(-\frac{F_k}{\epsilon})\cdot F_k}{\sum_{k=1}^{K}exp(-\frac{F_k}{\epsilon})}$$

Let F_{min} denote the $\min_k F_k$, we have

$$\begin{aligned}
\lim_{\epsilon\to 0}\frac{\sum_{k=1}^{K}exp(-\frac{F_k}{\epsilon})\cdot F_k}{\sum_{k=1}^{K}exp(-\frac{F_k}{\epsilon})}&=\lim_{\epsilon\to 0}\frac{\sum_{k=1}^{K}exp(-\frac{F_{min}}{\epsilon}-\frac{(F_k-F_{min})}{\epsilon})\cdot F_k}{\sum_{k=1}^{K}exp(-\frac{F_{min}}{\epsilon}-\frac{(F_k-F_{min})}{\epsilon})}\\
&=\lim_{\epsilon\to 0}\frac{\sum_{k=1}^{K}exp(-\frac{(F_k-F_{min})}{\epsilon})\cdot F_k}{\sum_{k=1}^{K}exp(-\frac{(F_k-F_{min})}{\epsilon})}\\
&=\lim_{\epsilon\to 0}\sum_{k=1}^{K}\frac{exp(-\frac{(F_k-F_{min})}{\epsilon})}{\sum_{k=1}^{K}exp(-\frac{(F_k-F_{min})}{\epsilon})}\cdot F_k\\
&=\sum_{k=1}^{K}\lim_{\epsilon\to 0}\frac{exp(-\frac{(F_k-F_{min})}{\epsilon})}{\sum_{k=1}^{K}exp(-\frac{(F_k-F_{min})}{\epsilon})}\cdot F_k
\end{aligned}\tag{1}$$

The term

$$\lim_{\epsilon\to 0}exp\left(-\frac{(F_k-F_{min})}{\epsilon}\right)$$

is zero except the k which minimizes F_k, therefore

$$\sum_{k=1}^{K}\lim_{\epsilon\to 0}\frac{exp(-\frac{(F_k-F_{min})}{\epsilon})}{\sum_{k=1}^{K}exp(-\frac{(F_k-F_{min})}{\epsilon})}\cdot F_k=\lim_{\epsilon\to 0}\frac{exp(-\frac{(F_{min}-F_{min})}{\epsilon})}{exp(-\frac{(F_{min}-F_{min})}{\epsilon})}\cdot F_{min}=F_{min}=\min_k F_k$$

(g)

From (f), the 0-temperature limit of Gaussian Mixture Model objective is:

$$\min_{\mu}\sum_{x_i\in\mathcal{D}}\min_k(x-\mu_k)^2$$

We can replace the $\min_k$ with the assignment of r_{ik} which takes value from $\{0,1\}$ and satisfies the requirement of $\sum_{k=1}^{K}r_{ik}=1,\forall i$ as follow:

$$\min_{\mu}\min_{\mathbf{r}}\sum_{x_i\in\mathcal{D}}\sum_{k=1}^{K}r_{ik}(x-\mu_k)^2$$

which is the same as k-Means objective of

$$\min_{\mu}\min_{\mathbf{r}}\sum_{x_i\in\mathcal{D}}\sum_{k=1}^{K}r_{ik}||x-\mu_k||_2^2$$

Questions assigned to the following page: 3.2, 3.3, 3.4, 3.1, 3.5, 3.6, and 3.7

Question 3

(a)

(i)

$$P(x|q) = \prod_{d=1}^{D} q_d^{x_d}(1-q_d)^{1-x_d}$$

(ii)

$$P(x^{(i)}|p,\pi) = \sum_{k=1}^{K} \pi_k P(x^{(i)}|p^{(k)})$$

(iii)

$$\begin{aligned} log\ P(\mathcal{D}|\pi,p) &= log \prod_{x^{(i)}\in\mathcal{D}} P(x^{(i)}|p,\pi) \\ &= log \prod_{x^{(i)}\in\mathcal{D}} \sum_{k=1}^{K} \pi_k P(x^{(i)}|p^{(k)}) \\ &= log \prod_{x^{(i)}\in\mathcal{D}} \sum_{k=1}^{K} \pi_k \prod_{d=1}^{D} (p_d^{(k)})^{x_d^{(i)}}(1-p_d^{(k)})^{1-x_d^{(i)}} \\ &= \sum_{x^{(i)}\in\mathcal{D}} log \sum_{k=1}^{K} \pi_k \prod_{d=1}^{D} (p_d^{(k)})^{x_d^{(i)}}(1-p_d^{(k)})^{1-x_d^{(i)}} \end{aligned} \tag{2}$$

(b)

(i)

$$P(z^{(i)}|\pi) = \prod_{k=1}^{K} \pi_k^{z_k^{(i)}}$$

(ii)

$$P(x^{(i)}|z^{(i)},p,\pi) = \sum_{k=1}^{K} z_k^{(i)} P(x^{(i)}|p^{(k)})$$

(iii)

$$P(Z,\mathcal{D}|\pi,p) = \prod_{x^{(i)}\in\mathcal{D}} P(z^{(i)}|\pi)P(x^{(i)}|z^{(i)},p,\pi) = \prod_{x^{(i)}\in\mathcal{D}} \left(\prod_{k=1}^{K} \pi_k^{z_k^{(i)}}\right)\left(\sum_{k=1}^{K} z_k^{(i)} P(x^{(i)}|p^{(k)})\right)$$

(iv)

$$\begin{aligned} \eta(z_k^{(i)}) &= \mathbb{E}[z_k^{(i)}|x^{(i)},\pi,p] \\ &= P(z_k^{(i)}=1|x^{(i)},\pi,p) \\ &= \frac{P(x^{(i)}|z_k^{(i)}=1,\pi,p)P(z_k^{(i)}=1|\pi,p)}{P(x^{(i)}|\pi,p)} \\ &= \frac{P(x^{(i)}|z_k^{(i)}=1,\pi,p)P(z_k^{(i)}=1|\pi,p)}{\sum_j P(x^{(i)}|z_j^{(i)}=1,\pi,p)P(z_j^{(i)}=1|\pi,p)} \\ &= \frac{P(x^{(i)}|p^{(k)})\pi_k}{\sum_j P(x^{(i)}|p^{(j)})\pi_j} \\ &= \frac{\pi_k \prod_{d=1}^{D}(p_d^{(k)})^{x_d^{(i)}}(1-p_d^{(k)})^{1-x_d^{(i)}}}{\sum_j \pi_j \prod_{d=1}^{D}(p_d^{(j)})^{x_d^{(i)}}(1-p_d^{(j)})^{1-x_d^{(i)}}} \end{aligned} \tag{3}$$

Questions assigned to the following page: 5.1, 4.2, 4.3, 4.4, 4.1, 4.5, 3.8, 3.9, and 3.10

(v)

$$\begin{aligned}
\mathbb{E}[log\ P(Z, \mathcal{D}|\tilde{p}, \tilde{\pi})|\mathcal{D}, p, \pi] &= \sum_{i=1}^{n}\sum_{k=1}^{K} log\ P(z_k^{(i)} = 1, x^{(i)}|\tilde{p}, \tilde{\pi}) \cdot P(z_k^{(i)} = 1|x^{(i)}, p, \pi) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} log\ (P(z_k^{(i)} = 1|\tilde{\pi})P(x^{(i)}|z_k^{(i)} = 1, \tilde{p}, \tilde{\pi})) \cdot \eta(z_k^{(i)}) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} log\ (\tilde{\pi}_k \prod_{d=1}^{D} (\tilde{p}_d^{(k)})^{x_d^{(i)}} (1 - \tilde{p}_d^{(k)})^{1-x_d^{(i)}}) \cdot \eta(z_k^{(i)}) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} \eta(z_k^{(i)}) \left[log\ \tilde{\pi}_k + \sum_{d=1}^{D} (x_d^{(i)} log\ \tilde{p}_d^{(k)} + (1 - x_d^{(i)}) log\ (1 - \tilde{p}_d^{(k)})) \right]
\end{aligned} \tag{4}$$

(c)

(i)

Optimality condition:

$$\frac{\partial \mathbb{E}[log\ P(Z, \mathcal{D}|\tilde{p}, \tilde{\pi})|\mathcal{D}, p, \pi]}{\partial \tilde{p}^{(k)}} = \sum_{i=1}^{n} \eta(z_k^{(i)}) \left(\frac{x^{(i)}}{\tilde{p}^{(k)}} - \frac{1 - x^{(i)}}{1 - \tilde{p}^{(k)}} \right) = 0 \quad \Longrightarrow \quad \tilde{p}^{(k)} = \frac{\sum_{i=1}^{n} \eta(z_k^{(i)}) x^{(i)}}{\sum_{i=1}^{n} \eta(z_k^{(i)})} = \frac{\sum_{i=1}^{n} \eta(z_k^{(i)}) x^{(i)}}{N_k}$$

(ii)

Question 4

(a)

K-Means is an unsupervised learning method that is used for clustering, while KNN algorithm is a supervised learning methods used for classification. Although these two method both have k in their name, these two k means very differently: the k in k-Means means k centers for k clusters, while the k in KNN algorithm indicates how many nearest neighbors to consider when classification.

(b)

This is ensured by $r_{ik} \in \{0, 1\}, \forall i \in \mathcal{D}, k \in \{1, 2, ..., K\}$ and $\sum_{k=1}^{K} r_{ik} = 1, \forall i \in \mathcal{D}$

(c)

We can change the requirements to $r_{ik} \in [0, 1], \forall i \in \mathcal{D}, k \in \{1, 2, ..., K\}$ and $\sum_{k=1}^{K} r_{ik} = 1, \forall i \in \mathcal{D}$

(d)

5 is the best choice for the number of clusters since it is the smallest number that gives the almost lowest cost function value.

(e)

K-Means is not an efficient algorithm to cluster the data. In this data, points are considered similar when they are closer alone the curve, rather than the euclidean distance. However, k-Means algorithm is based on euclidean distance which is not suitable to capture the pattern in this data.

Question 5

(a)

The domain for μ_k is $\mu_k \in \mathbb{R}^2, k \in \{1, 2, ..., K\}$

Questions assigned to the following page: 5.2, 5.3, 5.4, 5.5, and 5.6

(b)

The optimal $r_{x,k}$ is

$$r_{x,k} = \begin{cases} 1 & \mu_k \text{ is the closest cluster center to } x. \\ 0 & \text{Otherwise} \end{cases} \tag{5}$$

In this way, for each $x \in \mathcal{D}$, $\sum_{k\in\{1,\dots,K\}} \frac{1}{2} r_{x,k} ||x - \mu_k||_2^2$ is minimized, so does the entire cost function.

(c)

Optimize for μ given r:

$$\nabla_{\mu_k} : \sum_{x\in\mathcal{D}} r_{x,k}(x - \mu_k) = 0 \quad \Longrightarrow \quad \mu_k = \frac{\sum_{x\in\mathcal{D}} r_{x,k} x}{\sum_{x\in\mathcal{D}} r_{x,k}}$$

(d)

Algorithm 1 K-Means

Randomly initialize k cluster centers $\{\mu_k\}$
while Points' assignment change **do**
    Assign each point $x \in \mathcal{D}$ to its closest cluster center
    Update the cluster centers to be the mean of the points that are assigned to this cluster center.
end while

This algorithm is guaranteed to converge because the cost function monotonically decreases during the iterations until convergence.

This algorithm is not guaranteed to find the global optimum, which is illustrated in Figure 3 where colors indicate cluster center assignment. In this example, the global optimum solution should assign two clusters to the two clusters on the right.

Figure 3: K-Means local optimum example.

(e)

We can't use gradient descent because the cost function is not a differential function with respect to $r_{x,k}$ since the domain of $r_{x,k}$ is discrete.

(f)

We can initialize the cluster centers and run the algorithm multiple times.

Question assigned to the following page: 5.7

(g)

After two updates the algorithm converges. It converges to a cost function value of 8.5123 and the obtained cluster centers are $(-2.5087, 2.2610)$ and $(1.8960, -2.1884)$. The visualizations at step 0, step 1 and step 2 are shown in Figure 4, 5 and 6 respectively.

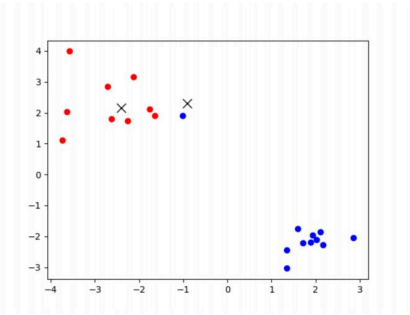

Figure 4: K-Means step 0.

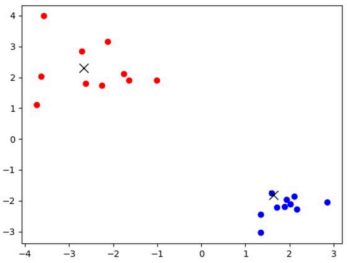

Figure 5: K-Means step 1.

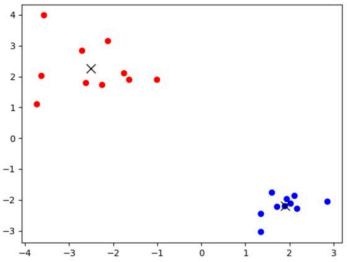

Figure 6: K-Means step 2.