AI 534: Machine Learning

Assignment IV

Name: Nam Nguyen

ID: 934-422-327

Problem 1:

$$\sum_{i=1}^{N} D_{\ell+1}(i)\, I\left(h_\ell(x_i) \neq y_i\right) = 0.5$$

- Let ϵ_i be the weighted error of h_i

$$\Rightarrow \quad \epsilon_i = \sum_{j=1}^{N} D_i(j)\, I\left(h_i(x_j) \neq y_j\right)$$

- The weights of the correct examples, $\epsilon e^{-\alpha}$
- The weights of the incorrect examples, $(1-\epsilon)\, e^{\alpha}$
- In additions, we have

$$\epsilon e^{-\alpha} = \epsilon\, e^{-\left(\frac{1}{2}\log\frac{\epsilon}{1-\epsilon}\right)}, \quad \text{where } \alpha = \frac{1}{2}\log\frac{\epsilon}{1-\epsilon}$$

$$= \epsilon\left(e^{\log\frac{\epsilon}{1-\epsilon}}\right)^{(-112)}$$

$$= \epsilon\left(\frac{\epsilon}{1-\epsilon}\right)^{(-112)} = \frac{\epsilon^{112}}{(1-\epsilon)^{(-112)}} \quad ①$$

$$(1-\epsilon)\, e^{\alpha} = (1-\epsilon)\, e^{\left(\frac{1}{2}\log\frac{\epsilon}{1-\epsilon}\right)}$$

$$= (1-\epsilon)\left(\frac{\epsilon}{1-\epsilon}\right)^{112}$$

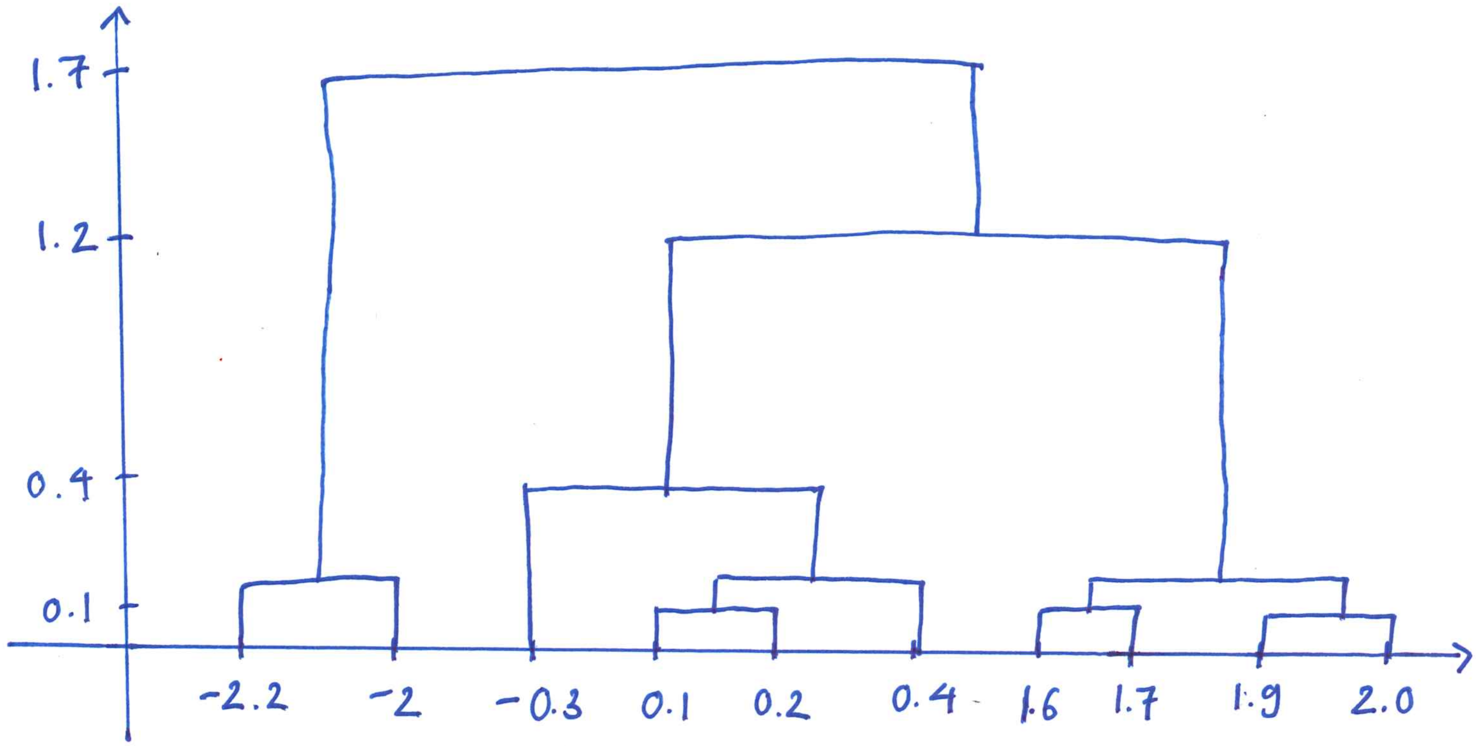
$$= \frac{\epsilon^{112}}{(1-\epsilon)^{(-112)}} \quad ②$$

From ① and ②, we have

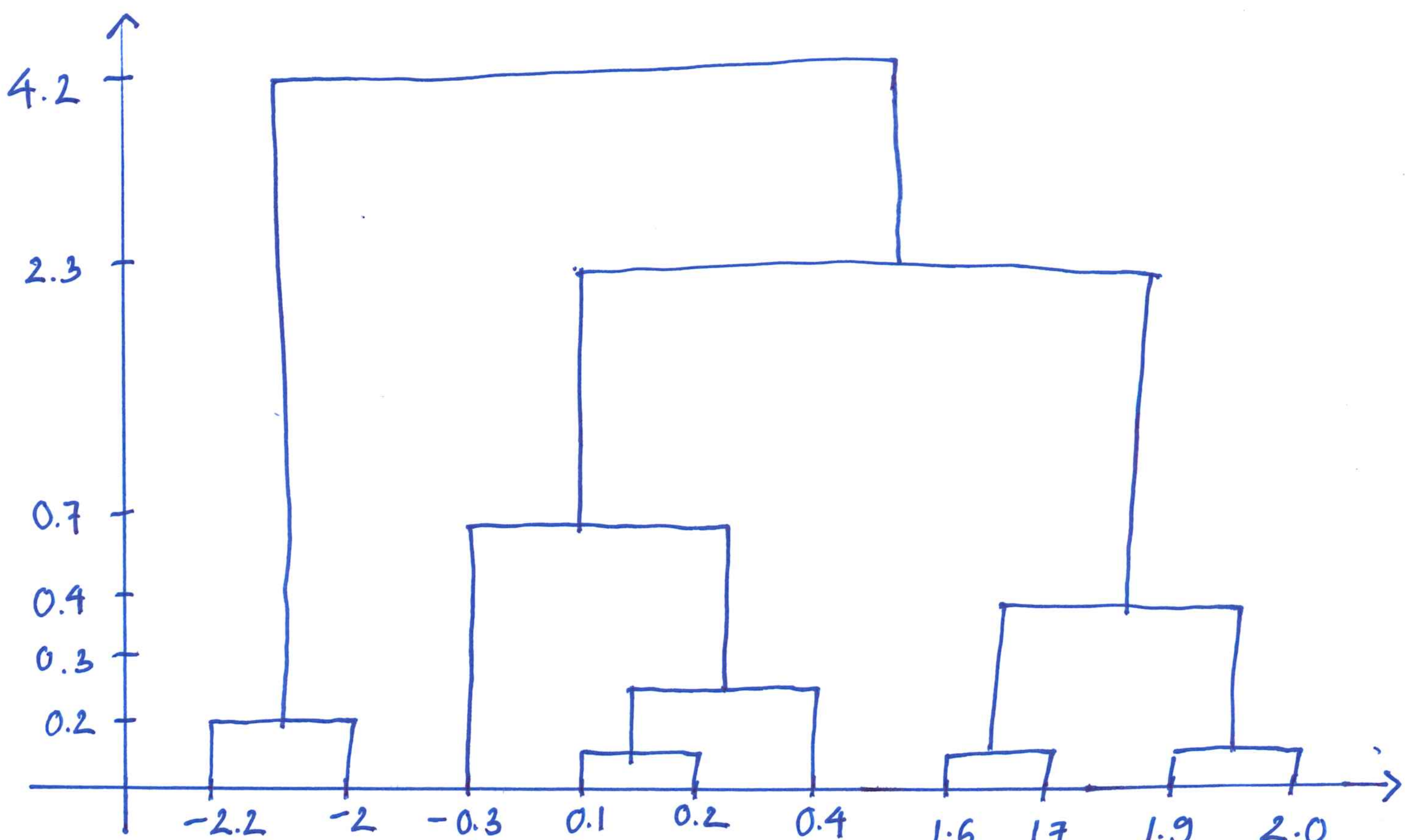
$$\epsilon e^{-\alpha} = (1-\epsilon) e^{\alpha}$$

It implys that the weighted error of h_i on the updated weights D_{i+1} is exactly 50%.

Problem 2 :



a) Using single link



b) Using complete link

Problem 3:

$$\min_{\mu_1,\dots\mu_K, C_1,\dots,C_K} \sum_{i=1}^{K} \sum_{\underline{X}\in C_i} |\underline{X} - \mu_i|$$

a) • Considering the j-th element of μ_i, we have

$$\min_{\mu_1,\dots\mu_K, C_1,\dots,C_K} \sum_{i=1}^{K} \sum_{X\in C_i} |X(j) - \mu_i(j)|$$

• We rewrite the objective function as follows

$$\min_{\mu_1,\dots,\mu_K, C_1,\dots,C_K} \sum_{i=1}^{K} \left(\sum_{X\in C_i} |X(j) - \mu_i(j)| + \text{constant} \right)$$

• In order to find the optimal $\mu_i(j)^*$, we need to take the derivative of $\sum_{X\in C_i} |X(j) - \mu_i(j)|$ and set to zero.

• We also have,

$$\frac{d \sum_{X\in C_i} |X(j) - \mu_i(j)|}{d\mu_i(j)} = \begin{cases} 1 & \text{if } \mu_i(j) > X(j) \\ -1 & \text{if } \mu_i(j) < X(j) \end{cases}$$

• Hence, $\frac{d}{d\mu_i(j)} \sum_{X\in C_i} |X(j) - \mu_i(j)| = 0 \Leftrightarrow$ the number of x with $x[j] < \mu_i(j)$ need to be equal the number of x with $x(j) > \mu_i(j)$.

This implys that μ_i that optimizes the above objective can be obtain by taking the median of each dimension for cluster i.

b) L_1 based objective k-means algorithm:

- Input: N data points, desired # of clusters k.
- Initialize: $\mu_1, \ldots, \mu_k$, the k cluster centers (by randomly selecting k points)
- Iterate:
  1. Assigning each of the N data points to the closest μ_i by using L_1 based objective
  2. Re-estimate the cluster center by assuming the current assignment is correct.

     Estimating $\mu_i(j)$ is the j-th dimension's median for all examples that are assigned to cluster i.
- Termination:

  If none of the data points changed membership in the last iteration, exist.

  Otherwise, go to 1.

c) . The L_1 based objective algorithm is more robust to outliers since it doesn't have the quadratic term as the L_2 based objective algorithm. And, using mean is not robust to outliers when comparing to use median.

Problem 4 :

a)

- Show that the minimum of J is a decreasing function of k.

- Using the induction argument:

- Assume that till k, J hase been non decreasing in k.

Let add another cluster center to some arbitrary location.

- After running the K-means, since it has not yet converged, the minimum possible J for k+1 clusters is not attainted.

- We know that, at least for the point that is now a cluster center, the term in J will be 0. This implys that J has dereased when we have added a new cluster.

b). This strategy is a bad idea. since the optimal (the minimum of) J will always decrease when we increase k.

- It is noted that when k = n, (the number of points in data set) that J = 0. Hence this approach will always give k = n.

Problem 5:

- $f(x|\theta_1) \sim N(\mu_1, \sigma^2)$, where $\begin{cases} \mu_1 = 0 \\ \sigma_1^2 = 1 \end{cases}$

- $f(x|\theta_2) \sim N(\mu_2, \sigma^2)$, where $\begin{cases} \mu_2 = 0 \\ \sigma_2^2 = 0.5 \end{cases}$

- α is mixing parameter (the prior probability of θ_1)

- We have,

$$L(\alpha) = p(x_1|\alpha) = \frac{\alpha}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} + \frac{(1-\alpha)}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}$$

$$= \frac{\alpha}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} + \frac{(1-\alpha)}{\sqrt{\pi}} e^{-x^2} \quad (\text{where } 0 \leq \alpha \leq 1)$$

- Maximum likelihood estimation of α:

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

$$+ \frac{1}{\sqrt{\pi}} e^{-x_1^2}$$

$$\Rightarrow \boxed{\max_{\alpha} \; p(x_1|\alpha).} = \max_{\alpha} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{x_1^2}{2}} - \frac{1}{\sqrt{\pi}} e^{-x_1^2} \right) \alpha$$

$$+ \frac{1}{\sqrt{\pi}} e^{-x_1^2}$$

- We have,

$$\frac{dp(x_1|\alpha)}{d\alpha} = \frac{1}{\sqrt{2\pi}} e^{-\frac{x_1^2}{2}} - \frac{1}{\sqrt{\pi}} e^{-x_1^2}$$

- $\frac{dp(x_1|\alpha)}{d\alpha} > 0 \Rightarrow \frac{1}{\sqrt{2\pi}} e^{-\frac{x_1^2}{2}} > \frac{1}{\sqrt{\pi}} e^{-x_1^2}$

$\Rightarrow -\frac{1}{2}x_1^2 > \ln(2) - x_1^2 \Rightarrow x_1^2 \geq \ln 2.$

- Hence, if $\frac{dp(x_1|\alpha)}{d\alpha} > 0 \Leftrightarrow x_1^2 \geq \ln 2$, then

$$\max_{\alpha} p(x_1|\alpha) = 1 \quad (\text{sinie } 0 \leq \alpha \leq 1)$$

else

$$\max_{\alpha} p(x_1|\alpha) = 0$$

Problem 6:

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

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

$$\mu_k(j) = p(\underline{x}(j) = 1 \mid z = k)$$

$$\sum_j \mu_k(j) = 1$$

- E-step:

- We have, $Q_i(z_i) = p(z_i \mid x_i ; \theta)$ is the probability that observation i belong to each of the K cluster.

- Hence,

$$\boxed{Q_i(z_i) = p(z_i \mid x_i ; \theta)}$$

$$\Rightarrow \quad Q_i(z_i = k) = p(z_i = k \mid x_i ; \theta)$$

$$= \frac{p(x_i \mid z_i = k ; \theta)\, p(z_i = k \mid \theta)}{p(x_i \mid \theta)} \quad \text{(Bayes rule)}$$

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

$$= \frac{\pi_k \, p(x_i \mid \mu_k)}{\sum_{j=1}^{K} \pi_j \prod_{m=1}^{M} \mu_j(m)^{x_i(m)}}$$

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

- M-step:

$$\boxed{\theta = \arg\max_{\theta} \sum_{i=1}^{N} \sum_{z_i} Q_i(z_i) \log \frac{p(x_i, z_i; \theta)}{Q_i(z_i)}}$$

$$\Rightarrow \theta = \arg\max_{\theta} \sum_{i=1}^{N} \sum_{j=1}^{K} Q_i(z_i = j) \log \frac{p(x_i, z_i = j; \theta)}{Q_i(z_i = j)}$$

$$= \arg\max_{\theta} \sum_{i=1}^{N} \sum_{j=1}^{K} Q_i(z_i = j) \log p(x_i, z_i = j; \theta)$$

$$= \arg\max_{\theta} \sum_{i=1}^{N} \sum_{j=1}^{K} Q_i(z_i = j) \log p(x_i \mid z_i = j; \theta)\, p(z_i = j; \theta)$$

$$= \arg\max_{\theta} \sum_{i=1}^{N} \sum_{j=1}^{K} Q_i(z_i = j) \log \pi_j \prod_{k=1}^{M} \mu_j(k)^{x_i(k)}$$

$$= \arg\max_{\theta} \sum_{i=1}^{N} \sum_{j=1}^{K} \left(Q_i(z_i=j) \log \pi_j + Q_i(z_i=j) \sum_{k=1}^{M} \log \mu_j(k)^{x_i(k)} \right)$$

$$= \arg\max_{\theta} \sum_{i=1}^{N} \sum_{j=1}^{K} \left(Q_i(z_i=j) \log \pi_j + Q_i(z_i=j) \sum_{k=1}^{M} x_i(k) \log \mu_j(k) \right)$$

- <u>For μ_ℓ:</u>

$$\Rightarrow \cancel{\arg\max_{\theta}} \quad \text{constant} + \cancel{\sum_{i=1}^{N} \sum_{j=1}^{K}} \; Q$$

$$\boxed{\sum_{i=1}^{N} Q_i(z_i=\ell) \sum_{k=1}^{M} x_i(k) \log \mu_\ell(k)}$$

- We use the Lagrangian multiplier method to solve:

$$\begin{cases} \arg\max_{\mu_\ell} \sum_{i=1}^{N} Q_i(z_i=\ell) \sum_{k=1}^{M} x_i(k) \log \mu_\ell(k) \\ \text{s.t.} \quad \sum_{j=1}^{M} \mu_\ell(j) = 1. \end{cases}$$

$$\Rightarrow L(\mu_\ell) = \sum_{i=1}^{N} Q_i(z_i=\ell) \sum_{k=1}^{M} x_i(k) \log \mu_\ell(k) + \beta \left(\sum_{j=1}^{M} \mu_\ell(j) - 1 \right)$$

$$(\beta \geqslant 0)$$

$$\Rightarrow \nabla_{\mu_\ell(k)} L(\mu_\ell) = \frac{\partial L(\mu_\ell)}{\partial \mu_\ell(k)} = \sum_{i=1}^{N} Q_i(z_i=\ell) \frac{x_i(k)}{\mu_\ell(k)} + \beta = 0$$

$$\Rightarrow \quad \mu_\ell(k) = \frac{\sum_{i=1}^{N} Q_i(z_i = \ell)\, x_i(k)}{(-\beta)}$$

- And, $\sum_{j=1}^{M} \mu_\ell(j) = 1$ (the constraint)

$$\Rightarrow \quad \sum_{j=1}^{M} \left(\frac{\sum_{i=1}^{N} Q_i(z_i = \ell)\, x_i(j)}{(-\beta)} \right) - 1 = 0$$

$$\Rightarrow \quad (-\beta) = \sum_{j=1}^{M} \sum_{i=1}^{N} Q_i(z_i = \ell)\, x_i(j)$$

- Hence,

$$\mu_\ell(k) = \frac{\sum_{i=1}^{N} Q_i(z_i = \ell)\, x_i(k)}{\sum_{j=1}^{M} \sum_{i=1}^{N} Q_i(z_i = \ell)\, x_i(j)}$$

$$= \frac{\sum_{i=1}^{N} Q_i(z_i = \ell)\, x_i(k)}{\sum_{i=1}^{N} Q_i(z_i = \ell)} \quad \left(\text{since } \sum_{j=1}^{M} x_i(j) = 1 \right)$$

- For π_ℓ :

$$\boxed{\sum_{i=1}^{N} Q_i(z_i = \ell) \log \pi_\ell}$$

- We also use the Lagrangian multiplier method to solve:

$$\begin{cases} \underset{\pi_\ell}{\arg\max} \sum_{i=1}^{N} Q_i(z_i=\ell) \log \pi_\ell \\ \text{s.t.} \sum_{j=1}^{K} \pi_j = 1 \end{cases}$$

$$\Rightarrow L(\pi_\ell) = \sum_{i=1}^{N} Q_i(z_i=\ell) \log \pi_\ell + \beta\left(\sum_{j=1}^{K} \pi_j - 1\right) \quad (\beta \geq 0)$$

$$\nabla_{\pi_\ell} L(\pi_\ell) = \frac{\partial L(\pi_\ell)}{\partial \pi_\ell} = \sum_{i=1}^{N} \frac{Q_i(z_i=\ell)}{\pi_\ell} + \beta = 0$$

$$\Rightarrow \pi_\ell = \frac{\sum_{i=1}^{N} Q_i(z_i=\ell)}{(-\beta)}$$

- And, $\sum_{j=1}^{K} \pi_j = 1$ (the constraint)

$$\Rightarrow \sum_{j=1}^{K} \left(\frac{\sum_{i=1}^{N} Q_i(z_i=j)}{(-\beta)} \right) - 1 = 0$$

$$\Rightarrow (-\beta) = \sum_{j=1}^{K} \sum_{i=1}^{N} Q_i(z_i=j)$$

- Hence,

$$\pi_\ell = \frac{\sum_{i=1}^{N} Q_i(z_i=\ell)}{\sum_{j=1}^{K} \sum_{i=1}^{N} Q_i(z_i=j)} = \frac{\sum_{i=1}^{N} Q_i(z_i=\ell)}{N}$$