CM146, Winter 2023
Problem Set 4: Boosting, Unsupervised learning

1 AdaBoost

1.1 (a)

Solution: Denoting the cost function as J, the partial derivative of J with respect to β_t can be expressed as:

$$\frac{\partial J}{\partial \beta_t} = \frac{\partial}{\partial \beta_t}\left(\left(e^{\beta_t} - e^{-\beta_t}\right)\varepsilon_t + e^{-\beta_t}\right)$$

It is important to note that ε_t is not dependent on β_t and can be treated as a constant. Setting the derivative to 0, we get:

$$\frac{\partial}{\partial \beta_t}\left(\left(e^{\beta_t} - e^{-\beta_t}\right)\varepsilon_t + e^{-\beta_t}\right) = 0$$
$$\left(e^{\beta_t} + e^{-\beta_t}\right)\varepsilon_t - e^{-\beta_t} = 0$$
$$\varepsilon_t e^{\beta_t} = e^{-\beta_t}\left(1 - \varepsilon_t\right)$$
$$\frac{e^{\beta_t}}{e^{-\beta_t}} = \frac{1-\varepsilon_t}{\varepsilon_t}$$
$$e^{2\beta_t} = \frac{1-\varepsilon_t}{\varepsilon_t}$$
$$\beta_t = \frac{1}{2}\log\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$$

Essentially, the above expressions demonstrate how to calculate the partial derivative of the cost function with respect to β_t. Note that ε_t is a constant and by setting the derivative to 0, we solve for β_t in terms of ε_t.

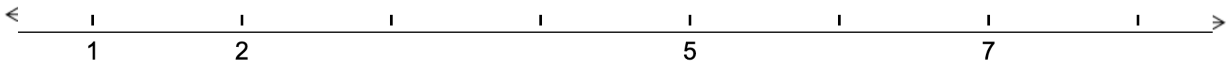
1.2 (b)

Solution: In the scenario where the training set can be separated linearly and no slack is permitted, there will be no instances of misclassification error. As a result, ε_t will tend towards 0. Referring back to the outcome of part (a), when ε_t approaches 0, β_1 will tend towards infinity.

2 K-means for single dimensional data

2.1 (a)

Solution: When we place these points on a number line, we obtain:



When $K = 3$, the optimal clustering solution involves placing the centers at $\mu_1 = 1.5, \mu_2 = 5$, and $\mu_3 = 7$. This arrangement assigns x_1 and x_2 to μ_1, x_3 to μ_2, and x_4 to μ_3. The value of the objective, then, can be calculated as:

$$(1 - 1.5)^2 + (2 - 1.5)^2 + (5 - 5)^2 + (7 - 7)^2 = 0.5$$

2.2 (b)

Solution: A possible suboptimal assignment would be $\mu_1 = 1, \mu_2 = 2$, and $\mu_3 = 6$, where x_1 is assigned to μ_1, x_2 is assigned to μ_2, and x_3 and x_4 are assigned to μ_3. The value of the objective is:

$$(1 - 1)^2 + (2 - 2)^2 + (5 - 6)^2 + (7 - 6)^2 = 2$$

The resulting objective value is 2, which is larger than the value of 0.5 obtained in part (a). However, if we apply Lloyd's algorithm to this assignment, we find that x_1 is closest to μ_1, x_2 is closest to μ_2, and both x_3 and x_4 are closest to μ_3. As a result, there is no change in assignments or centroids. Therefore, we have arrived at a suboptimal solution that represents only a local minimum, rather than the global minimum.

3 Gaussian Mixture Models

3.1 (a)

Solution: The multivariate normal distribution represents a variable with d dimensions and is characterized by its mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. It is given by the formula:

$$N(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

After substituting this in the formula for $l(\boldsymbol{\theta})$, the gradient with respect to $\boldsymbol{\mu}_j$ is computed as follows:

$$l(\boldsymbol{\theta}) = \sum_k \sum_n \gamma_{nk} \log \omega_k + \sum_k \sum_n \gamma_{nk} \log \left(\frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_k|}} \exp\left(-\frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)\right) \right)$$

$$l(\boldsymbol{\theta}) = \sum_k \sum_n \gamma_{nk} \log \omega_k + \sum_k \sum_n \gamma_{nk} \left(\left(\log \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_k|}} \right) - \frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) \right)$$

The first summation is not a function of $\boldsymbol{\mu}_{\mathrm{j}}$ and the first term in the second summation is constant, which means that their gradients are both equal to zero. Therefore, taking the gradient results in:

$$\nabla_{\boldsymbol{\mu}_j} l(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\mu}_j} \sum_n \left(-\frac{1}{2} \gamma_{nj} \left(\mathbf{x}_n - \boldsymbol{\mu}_j \right)^T \boldsymbol{\Sigma}_j^{-1} \left(\mathbf{x}_n - \boldsymbol{\mu}_j \right) \right)$$
$$\nabla_{\boldsymbol{\mu}_j} l(\boldsymbol{\theta}) = \sum_n \left(-\frac{1}{2} \gamma_{nj} (2)(-1) \left(\mathbf{x}_n - \boldsymbol{\mu}_j \right) \boldsymbol{\Sigma}_j^{-1} \right)$$
$$\nabla_{\boldsymbol{\mu}_j} l(\boldsymbol{\theta}) = \boldsymbol{\Sigma}_j^{-1} \sum_n \left(\gamma_{nj} \left(\mathbf{x}_n - \boldsymbol{\mu}_j \right) \right)$$

3.2 (b)

Solution: To obtain the desired answer, we set the result obtained in part (a) equal to zero and solve for $\boldsymbol{\mu}_j$.

$$\nabla_{\boldsymbol{\mu}_j} l(\boldsymbol{\theta}) = \boldsymbol{\Sigma}_j^{-1} \sum_n \left(\gamma_{nj} \left(\mathbf{x}_n - \boldsymbol{\mu}_j \right) \right) = \mathbf{0}$$
$$\sum_n \gamma_{nj} \mathbf{x}_n = \boldsymbol{\mu}_j \sum_n \gamma_{nj}$$
$$\boldsymbol{\mu}_j = \frac{\sum_n \gamma_{nj} \mathbf{x}_n}{\sum_n \gamma_{nj}}$$

3.3 (c)

Solution: We learned in lecture that we can express ω_k and $\boldsymbol{\mu}_k$ as follows:

$$\omega_k = \frac{\sum_n \gamma_{nk}}{\sum_k \sum_n \gamma_{nk}} \quad \boldsymbol{\mu}_k = \frac{\sum_n \gamma_{nk} \mathbf{x}_n}{\sum_n \gamma_{nk}}$$

After substituting the values from the table, we obtain the following values:

$$\omega_1 = \frac{0.2 + 0.2 + 0.8 + 0.9 + 0.9}{0.2 + 0.2 + 0.8 + 0.9 + 0.9 + 0.8 + 0.8 + 0.2 + 0.1 + 0.1} = \frac{3}{5} = 0.6$$
$$\omega_2 = \frac{0.8 + 0.8 + 0.2 + 0.1 + 0.1}{5} = \frac{2}{5} = 0.4$$
$$\mu_1 = \frac{1}{3}(0.2(5) + 0.2(15) + 0.8(25) + 0.9(30) + 0.9(40)) = 29$$
$$\mu_2 = \frac{1}{2}(0.8(5) + 0.8(15) + 0.2(25) + 0.1(30) + 0.1(40)) = 14$$