CM146, Winter 2020
Problem Set 5: Boosting, Unsupervised learning
Due March 15, 2020
Submitted by Guy Ohayon

1 AdaBoost

Solution:

(a) Let $L(h_t(x), \beta_t) = (e^{\beta_t} - e^{-\beta_t}) \sum_n w_t(n) \mathbb{I}[y_n \neq h_t(x_n)] + e^{-\beta_t} \sum_n w_t(n) = (e^{\beta_t} - e^{-\beta_t})\epsilon_t + e^{-\beta_t}$.

$$\Rightarrow \frac{\partial L}{\partial \beta_t} = (e^{\beta_t} + e^{-\beta_t})\epsilon_t - e^{-\beta_t} = 0$$

$$\Rightarrow (e^{2\beta_t} + 1)\epsilon_t = 1 \Rightarrow e^{2\beta_t}\epsilon_t = 1 - \epsilon_t \Rightarrow 2\beta_t = \log(\frac{1-\epsilon_t}{\epsilon_t})$$

$$\Rightarrow \boxed{\beta_t^* = \frac{1}{2}\log(\frac{1-\epsilon_t}{\epsilon_t})}$$

(b) The data is linearly separable, and we are using a hard margin SVM. This means that after training the SVM, all training points will be correctly classified, which means that $\epsilon_1 = 0$. Thus:

$$\boxed{\beta_t^* = \frac{1}{2}\log(\frac{1}{0}) = \infty}$$

.

2 K-means for a single dimensional data

Solution:

(a) All of the clustering possibilities (for K=3) and their objective losses are:

i. $\{1, 7\}, \{2\}, \{5\}$. $L = 3^2 + 3^2 = 18$.
ii. $\{1\}, \{2, 7\}, \{5\}$. $L = 2.5^2 + 2.5^2 = 12.5$.

iii. $\{1\},\{2\},\{5,7\}$. $L = 1^2 + 1^2 = 2$.
iv. $\{1,5\},\{2\},\{7\}$. $L = 2^2 + 2^2 = 8$.
v. $\{1\},\{2,5\},\{7\}$. $L = 1.5^2 + 1.5^2 = 4.5$.
vi. $\{1,2\},\{5\},\{7\}$. $L = 0.5^2 + 0.5^2 = 0.5$.

Thus, the optimal clustering is $\{1,2\},\{5\}\{7\} = \{x_1,x_2\},\{x_3\},\{x_4\}$ and the corresponding objective is 0.5.

(b) Suppose we initialize 3 clusters of x_1, x_2 and x_3 with centroids $c_1 = x_1 = 1, c_2 = x_2 = 2, c_3 = x_3 = 5$. On the first iteration of the algorithm, x_2 is closest to the centroid $c_2 = 2$, x_1 is closest to the centroid $c_1 = 1$, x_3 is closest to the centroid $c_3 = 5$, and x_4 is closest to the centroid $c_3 = 5$. Thus, x_4 will join the cluster of c_3, the cluster's new centroid will be $c_3 = 6$, and no further changes will happen because the other points sit exactly on their cluster's centroid. On the second iteration of the algorithm, x_1 and x_2 will remain in seperate clusters because both of these points still sit exactly on their cluster's centroid. Also, x_3 and x_4 will remain in their cluster because they are closer to their cluster's centroid c_3 than to the other centroids c_1, c_2. Thus, on the 2nd iteration there will be no changes in the clusters and the algorithm will terminate and result in the clustering $\{x_1\},\{x_2\},\{x_3,x_4\} = \{1\},\{2\},\{5,7\}$, which has an objective loss of 2 and therefore is not optimal (the clustering objective will not improve anymore because the algorithm terminates when the clusters stop changing).

3 Gaussian Mixture Models

Solution:

(a)

$$\frac{\partial l(\theta)}{\partial \mu_j} = \frac{\partial \sum_k \sum_n \gamma_{nk} \log N(x_n|\mu_k\Sigma_k)}{\partial \mu_j} = \frac{\sum_k \sum_n \gamma_{nk} \partial \log N(x_n|\mu_k\Sigma_k)}{\partial \mu_j}$$

$$= \sum_k \sum_n \frac{\gamma_{nk}}{N(x_n|\mu_k\Sigma_k)} \frac{\partial N(x_n|\mu_k\Sigma_k)}{\partial \mu_j}$$

Lets calculate $\frac{\partial N(x_n|\mu_k\Sigma_k)}{\partial \mu_j}$:

$$N(x_n|\mu_k\Sigma_k) = \frac{1}{\sqrt{(2\pi)^d|\Sigma_k|}} e^{-\frac{1}{2}(x_n-\mu_k)^T\Sigma_k^{-1}(x_n-\mu_k)}$$

$$= \frac{1}{\sqrt{(2\pi)^d|\Sigma_k|}} e^{-\frac{1}{2}(x_n^T\Sigma^{-1}x_n - 2\mu_k^T\Sigma_k^{-1}x_n + \mu_k^T\Sigma_k^{-1}\mu_k)}$$

$$\Rightarrow \frac{\partial N(x_n|\mu_k\Sigma_k)}{\partial \mu_j} = -\mathbb{1}_{(k=j)}\frac{1}{2}N(x_n|\mu_k\Sigma_k)[-2\Sigma_k^{-1}x_n + ((\Sigma_k^{-1})^T + \Sigma_k^{-1})\mu_k]$$

$$= -\mathbb{1}_{(k=j)}\frac{1}{2}N(x_n|\mu_k\Sigma_k)[-2\Sigma_k^{-1}x_n + ((\Sigma_k^T)^{-1} + \Sigma_k^{-1})\mu_k] = \mathbb{1}_{(k=j)}N(x_n|\mu_k\Sigma_k)[\Sigma_k^{-1}x_n - \Sigma_k^{-1}\mu_k]$$

$$= \mathbb{1}_{(k=j)}N(x_n|\mu_k\Sigma_k)\Sigma_k^{-1}(x_n - \mu_k)$$

Therefore:

$$\frac{\partial l(\theta)}{\partial \mu_j} = \sum_k \sum_n \frac{\gamma_{nk}}{N(x_n|\mu_k\Sigma_k)} \frac{\partial N(x_n|\mu_k\Sigma_k)}{\partial \mu_j}$$

$$= \sum_k \sum_n \frac{\gamma_{nk}}{N(x_n|\mu_k\Sigma_k)} \mathbb{1}_{(k=j)}N(x_n|\mu_k\Sigma_k)\Sigma_k^{-1}(x_n - \mu_k)$$

$$= \sum_k \sum_n \gamma_{nk}\mathbb{1}_{(k=j)}\Sigma_k^{-1}(x_n - \mu_k) = \boxed{\sum_n \gamma_{nj}\Sigma_j^{-1}(x_n - \mu_j)}$$

(b) By setting the gradient to zero:

$$\sum_n \gamma_{nj}\Sigma_j^{-1}(x_n - \mu_j) = 0 \Rightarrow \sum_n \gamma_{nj}\Sigma_j^{-1}x_n = \sum_n \gamma_{nj}\Sigma_j^{-1}\mu_j$$

Multiply both sides by Σ_j from the left, and then we get:

$$\Rightarrow \boxed{\mu_j = \frac{1}{\sum_n \gamma_{nj}} \sum_n \gamma_{nj}x_n}$$

(c) We are given γ_{nk} for each n, k at the end of step 1 at iteration 5 of the EM algorithm. At step 2 (of the same iteration), the EM algorithm treats γ_{nk} as fixed to estimate θ based on the current γ_{nk}. Thus, to calculate w_1, w_2, μ_1, μ_2, we simply need to use the given γ_{nk} to calculate the optimal μ_1, μ_2 based on the formula we found in (b), and also to use the formula from the lecture for the optimal w_k (it's also very simple to derive). We get:

$$\sum_n \gamma_{n1} = 0.2 \cdot 2 + 0.8 + 0.9 \cdot 2 = 3$$

$$\sum_n \gamma_{n2} = 0.2 + 0.8 \cdot 2 + 0.1 \cdot 2 = 2$$

$$\mu_1 = \frac{1}{\sum_n \gamma_{n1}} \sum_n \gamma_{n1} x_n = \frac{1}{3}(0.2{\cdot}(5{+}15){+}0.8{\cdot}25{+}0.9{\cdot}(30{+}40)) = \frac{87}{3} = \boxed{29}$$

$$\mu_2 = \frac{1}{\sum_n \gamma_{n2}} \sum_n \gamma_{n2} x_n = \frac{1}{2}(0.8{\cdot}(5{+}15){+}0.2{\cdot}25{+}0.1{\cdot}(30{+}40)) = \frac{28}{2} = \boxed{14}$$

$$w_1 = \frac{\sum_n \gamma_{n1}}{\sum_n \sum_k \gamma_{nk}} = \frac{3}{3+2} = \boxed{\frac{3}{5}}$$

$$w_2 = \frac{\sum_n \gamma_{n2}}{\sum_n \sum_k \gamma_{nk}} = \frac{2}{3+2} = \boxed{\frac{2}{5}}$$