Intro to Big Data Science: Assignment 4 Reference Answer

TAN 12212523@mail.sustech.edu.cn

June 6, 2025

Exercise 1 (AdaBoost)

1. The "$-$" sample is misclassified and thus has its weight increased, since the best classifier $G_1(x)$ is to classify all samples to be "$+$" in order to minimize ϵ_1.

2. It takes exactly 3 iterations to achieve zero training error.

   Initially, $\omega_i^{(1)} = \frac{1}{5}$. After the 1st iteration, $\epsilon_1 = \frac{1}{5}$ and $\alpha_1 = \log 4$, all the weights are kept the same except the weight of the "$-$" sample is updated to be $w^{(2)} = \frac{4}{5}$.

   After the second iteration, $G_2(x)$ classifies the two "$+$" samples on the same side (left two, or right two, or upper two, or lower two) to be "$+$" while classifying the others to be "$-$". Then $\epsilon_2 = \frac{1}{4}$ and $\alpha_2 = \log 3$, the weights of the misclassified two "$+$" are updated to be $w_{+,\text{mis}}^{(3)} = \frac{3}{5}$, while the weights of the other three samples are unchanged.

   After the third iteration, $G_3(x)$ classifies the two "$+$" samples that are misclassified in 2nd iteration to be "$+$", while classifying the others to be "$-$". Then $\epsilon_3 = \frac{1}{6}$ and $\alpha_3 = \log 5$.

   The boosting classifier is given by $G(x) = sgn(\log 4 \cdot G_1(x) + \log 3 \cdot G_2(x) + \log 5 \cdot G_3(x))$. It is straightforward to check that this classifier classifies every sample correctly.

3. The AdaBoost algorithm can never achieve zero training error in two steps no matter whether we add a "$+$" or "$-$" sample and where we put the new sample.

   This can be considered in separate cases of adding a "$+$" sample and adding a "$-$" sample. The discussion proceeds in the same way as we have done in part 2 with much more involved calculations.

   A general explanation is that after 2 iterations there are always a pair of a "$+$" sample and a "$-$" sample that come in the same class in each of the two iteration. This leads to the same weights of them and thus the same classification results, which obviously gives rise to nonzero training error.

Exercise 2 (Alternative objective functions)

(a) Using coordinate descent:

1. Select base classifier minimizing weighted error:

$$b_u = \arg\min_{b\in\mathcal{H}} \sum_{i=1}^{m} w_i^{(t)} \mathbb{I}(y_i \neq b(x_i)), \quad w_i^{(t)} = \Phi'(-y_i f_{t-1}(x_i))$$

2. Solve for optimal weight α_t:

$$\alpha_t = \arg\min_{\alpha} \sum_{i=1}^{m} \Phi\left(-y_i(f_{t-1}(x_i) + \alpha b_u(x_i))\right)$$

(b) Loss Function Validation Required properties: strictly increasing, convex, differentiable, and

$$\forall x \geq 0, \Phi(x) \geq 1; \quad \forall x < 0, \Phi(x) > 0.$$

1. 0-1 loss: $\Phi_1(-u) = \mathbb{I}_{u \le 0}$
   Not differentiable, discontinuous. **Not satisfy**

2. Least squared loss: $\Phi_2(-u) = (1-u)^2$
   Not strictly increasing ($\downarrow$ for $x < -1$). **Not satisfy**

3. SVM loss: $\Phi_3(-u) = \max\{0, 1-u\}$
   Non-differentiable at $u = 1$. **Not satisfy**

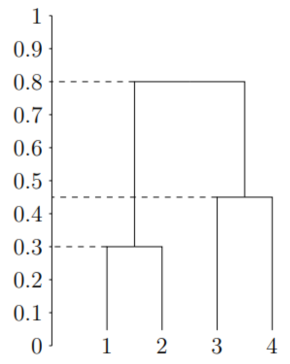
4. Logistic loss: $\Phi_4(-u) = \log(1 + e^{-u})$
   $\Phi(0) = \log_2 2 = 1$ satisfies $\forall x \ge 0, \Phi(x) \ge 1$. **Satisfy**

Exercise 3 (Hierarchical Clustering)

1. We would have

$$\begin{pmatrix} 0 & 0.3 & 0.4 & 0.7 \\ 0.3 & 0 & 0.5 & 0.8 \\ 0.4 & 0.5 & 0 & 0.45 \\ 0.7 & 0.8 & 0.45 & 0 \end{pmatrix} \Rightarrow \begin{pmatrix} 0 & 0.5 & 0.8 \\ 0.5 & 0 & 0.45 \\ 0.8 & 0.45 & 0 \end{pmatrix} \Rightarrow \begin{pmatrix} 0 & 0.5 & 0.8 \\ 0.5 & 0 & 0.45 \\ 0.8 & 0.45 & 0 \end{pmatrix}$$

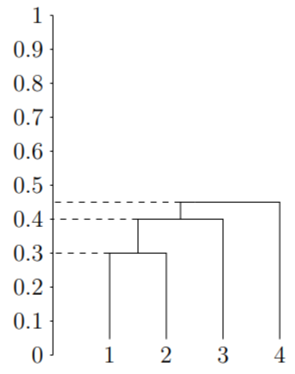
Hence, the dendrogram is



2. We would have

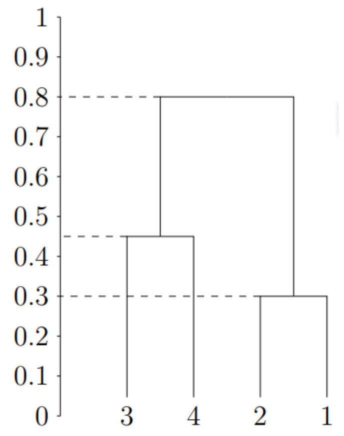
$$\begin{pmatrix} 0 & 0.3 & 0.4 & 0.7 \\ 0.3 & 0 & 0.5 & 0.8 \\ 0.4 & 0.5 & 0 & 0.45 \\ 0.7 & 0.8 & 0.45 & 0 \end{pmatrix} \Rightarrow \begin{pmatrix} 0 & 0.4 & 0.7 \\ 0.4 & 0 & 0.45 \\ 0.7 & 0.45 & 0 \end{pmatrix} \Rightarrow \begin{pmatrix} 0 & 0.45 \\ 0.45 & 0 \end{pmatrix}$$

Hence, the dendrogram is



3. (1, 2), (3, 4)

4. ((1, 2), 3), (4)
5. Example:



Exercise 4

5. (1) M is average of dataset. $M_j = \frac{1}{N}\sum_{i=1}^{d} x_i^j$

$$||x_i - M||^2 = \sum_{j=1}^{d}(x_{ij} - M_j)^2$$

$$SSE_T = \sum_{i=1}^{d} ||x_i - M||^2, \text{ and we know } SSE_T = \sum_{j=1}^{d} SSE_j$$

$$SSE_j = \sum_{i=1}^{d}(x_{ij} - M_j)^2 = \sum_{i=1}^{d} x_{ij}^2 - 2M_j \sum_{i=1}^{d} x_{ij}^2 + NM_j^2$$

$$= \sum_{i=1}^{d} x_{ij}^2 - 2M_j \cdot NM_j + NM_j^2 = \sum_{i=1}^{d} x_{ij}^2 - NM_j^2$$

(2)

$$SSE^{(j)} = \sum_{i\in G_j}\sum_{k=1}^{d}(x_{ik} - m_j M_{jk})^2 = \sum_{i\in G_j}\sum_{k=1}^{d}(x_{ik}^2 - 2x_{ik}M_{jk} + M_{jk}^2)$$

$$= \sum_{i\in G_j}\sum_{k=1}^{d} x_{ik}^2 - n_j \sum_{k=1}^{d} M_{jk}^2$$

(3)

$$SSE_T = \sum_{i=1}^{d}\sum_{k=1}^{d} x_{ik}^2 - N\sum_{k=1}^{d} M_k^2$$

$$SSE^{(j)} + SSE^{(2)} = \sum_{i=1}^{d}\sum_{k=1}^{d} x_{ik}^2 - \left(n_1 \sum_{k=1}^{d} M_{ik} + n_2 \sum_{k=1}^{d} M_{kk}^2\right)$$

$$N\sum_{k=1}^{d} M_k^2 \leq n_1 \sum_{k=1}^{d} M_{ik}^2 + n_2 \sum_{k=1}^{d} M_{kk}^2$$

$$\Rightarrow SSE_T \geq SSE^{(1)} + SSE^{(2)}$$

Exercise 5 (EM Algorithm)

1. The likelihood function is

$$L(\mu, a, b) = \left(\frac{1}{2}\right)^a (\mu)^b (2\mu)^c \left(\frac{1}{2} - 3\mu\right)^d$$

It follows that the log-likelihood function is

$$l(\mu, a, b) = \log L(\mu, a, b) = a \log(1/2) + b \log(\mu) + c \log(2\mu) + d \log(1/2 - 3\mu).$$

2. Since $\mathbb{P}(A) = 1/2$ and $\hat{\mathbb{P}}(B) = \hat{\mu}^{(m)}$, then we get

$$\hat{a}^{(m)} = \frac{\mathbb{P}(A)}{\mathbb{P}(A) + \hat{\mathbb{P}}(B)} h = \frac{1/2}{1/2 + \hat{\mu}^{(m)}} h \quad \text{and} \quad \hat{b}^{(m)} = \frac{\hat{\mathbb{P}}(B)}{\mathbb{P}(A) + \hat{\mathbb{P}}(B)} h = \frac{\hat{\mu}^{(m)}}{1/2 + \hat{\mu}^{(m)}} h.$$

3. By differentiate l with respect to μ and let it be zero, we have

$$\frac{\partial l}{\partial \mu} = \frac{b}{\mu} + \frac{c}{\mu} + \frac{-3d}{1/2 - 3\mu} = 0.$$

This gives that

$$\hat{\mu}^{(m+1)} = \frac{\hat{b}^{(m)} + c}{6(\hat{b}^{(m)} + c + d)} = \frac{c + 2\hat{\mu}^{(m)}(h + c)}{6\left((c + d) + 2\hat{\mu}^{(m)}(h + c + d)\right)}.$$

4. **True.** Because the value of the log-likelihood function (lower bound) increases at each iteration.