Introduction to Big Data Analysis

Homework 5 Reference Answer

2.

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 $w_{+,mis}^{(3)} = \frac{3}{5}$, while the weights of the other three samples are unchanged.

After the third iteration, G3(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 $G(x) = sgn(\log 4G_1(x) + \log 3G_2(x) + \log 5G_3(x))$. 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.

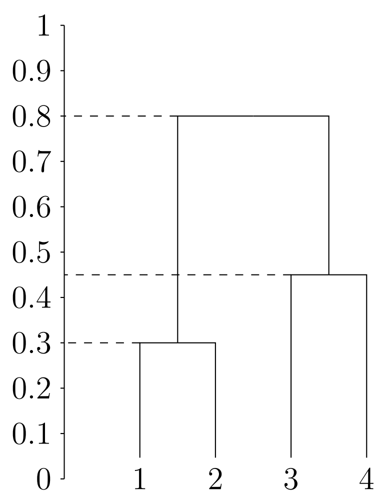
A general explanation is that after 2 iterations there are always a pair of a "+" sample and a "-" sample that come in the same class in each of the two iteration. This leads to the same weights of them and thus the same classification results, which obviously gives rise to nonzero training error.

3.

1. We would have

$$\begin{array}{cc} & \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} & \begin{pmatrix} 0 & 0.3 & 0.4 & 0.7 \\ 0.3 & 0 & 0.5 & 0.8 \\ 0.4 & 0.5 & 0 & 0.45 \\ 0.7 & 0.8 & 0.45 & 0 \end{pmatrix} \end{array} \Rightarrow \begin{array}{cc} & \begin{array}{ccc} (1,2) & 3 & 4 \end{array} \\ \begin{array}{c} (1,2) \\ 3 \\ 4 \end{array} & \begin{pmatrix} 0 & 0.5 & 0.8 \\ 0.5 & 0 & 0.45 \\ 0.8 & 0.45 & 0 \end{pmatrix} \end{array} \Rightarrow \begin{array}{cc} & \begin{array}{cc} (1,2) & (3,4) \end{array} \\ \begin{array}{c} (1,2) \\ (3,4) \end{array} & \begin{pmatrix} 0 & 0.8 \\ 0.8 & 0 \end{pmatrix} \end{array}$$

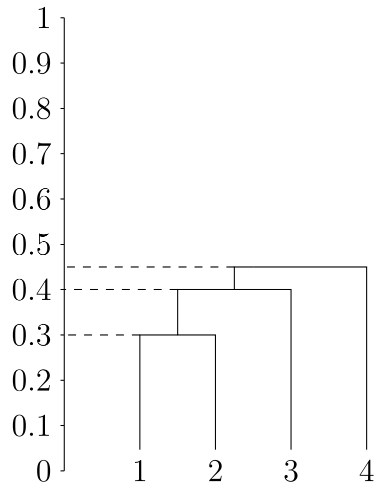
Hence, the dendrogram is



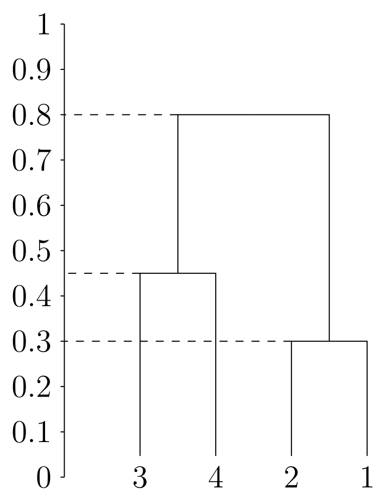
2. We would have

$$\begin{array}{cc} & \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} & \begin{pmatrix} 0 & 0.3 & 0.4 & 0.7 \\ 0.3 & 0 & 0.5 & 0.8 \\ 0.4 & 0.5 & 0 & 0.45 \\ 0.7 & 0.8 & 0.45 & 0 \end{pmatrix} \end{array} \Rightarrow \begin{array}{cc} & \begin{array}{ccc} (1,2) & 3 & 4 \end{array} \\ \begin{array}{c} (1,2) \\ 3 \\ 4 \end{array} & \begin{pmatrix} 0 & 0.4 & 0.7 \\ 0.4 & 0 & 0.45 \\ 0.7 & 0.45 & 0 \end{pmatrix} \end{array} \Rightarrow \begin{array}{cc} & \begin{array}{cc} ((1,2),3) & 4 \end{array} \\ \begin{array}{c} ((1,2),3) \\ 4 \end{array} & \begin{pmatrix} 0 & 0.45 \\ 0.45 & 0 \end{pmatrix} \end{array}$$

Hence, the dendrogram is



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



4.

1. The OLS coefficients are given by

$$\begin{aligned}
\hat{\omega}_1^{*b} &= \frac{\sum_{i=1}^n (x_i^{*b} - \bar{x}^{*b})(y_i^{*b} - \bar{y}^{*b})}{\sum_{i=1}^n (x_i^{*b} - \bar{x}^{*b})^2} \\
&= \frac{\sum_{i=1}^n x_i^{*b} y_i^{*b} - n\bar{x}^{*b}\bar{y}^{*b}}{\sum_{i=1}^n (x_i^{*b})^2 - n(\bar{x}^{*b})^2} \\
&= \frac{n\sum_{i=1}^n x_i^{*b} y_i^{*b} - \sum_{i=1}^n x_i^{*b} \sum_{i=1}^n y_i^{*b}}{n\sum_{i=1}^n (x_i^{*b})^2 - (\sum_{i=1}^n x_i^{*b})^2}, \\
\hat{\omega}_0^{*b} &= \bar{y}^{*b} - \hat{\omega}_1^{*b}\bar{x}^{*b} \\
&= \frac{n\bar{y}^{*b}\sum_{i=1}^n (x_i^{*b})^2 - \bar{y}^{*b}(\sum_{i=1}^n x_i^{*b})^2 - n\bar{x}^{*b}\sum_{i=1}^n x_i^{*b} y_i^{*b} - \bar{x}^{*b}\sum_{i=1}^n x_i^{*b}\sum_{i=1}^n y_i^{*b}}{n\sum_{i=1}^n (x_i^{*b})^2 - (\sum_{i=1}^n x_i^{*b})^2} \\
&= \frac{n\sum_{i=1}^n x_i^{*b} y_i^{*b} - \sum_{i=1}^n x_i^{*b} \sum_{i=1}^n y_i^{*b}}{n\sum_{i=1}^n (x_i^{*b})^2 - (\sum_{i=1}^n x_i^{*b})^2}.
\end{aligned}$$

2. From that $\hat{\omega}_i^{*b}$ depends only on Z^* with $(x_i^*, y_i^*) \sim \hat{P}$, by the law of large number, $\frac{1}{B}\sum_{b=1}^B \hat{\omega}_i^{*b} \xrightarrow{P} E_{\hat{P}}\hat{\omega}_i^*$ as $B \to \infty$.

3. Similarly, denote by $Z^{*b(-i)} \subset Z\backslash\{x_i, y_i\}$ each group of bootstrap data, b=1, 2, $\cdots$, B, then rewrite the OLS fitted coefficients $\hat{\omega}_0^{*b(-i)}$, $\hat{\omega}_1^{*b(-i)}$, and the univariate linear model $\hat{f}^{*b(-i)}(x) = \hat{\omega}_0^{*b(-i)} + \hat{\omega}_1^{*b(-i)}x$.

The bagging estimate of $\hat{f}^{(-i)}(x)$ is $\hat{f}_{bag}^{(-i)} = \frac{1}{B}\sum_{b=1}^B \hat{f}^{*b(-i)}(x)$.

From part 2, $\hat{f}_{bag}^{(-i)} \xrightarrow{P} E_{\hat{P}}\hat{f}^{*(-i)}(x)$, $B \to \infty$.