SOUTH UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA

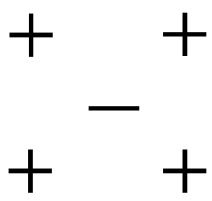
Intro to Big Data Science: Assignment 5

Due Date: April 28, 2022

✏ **Exercise 1**

Log into "cookdata.cn", and enroll the course "数据科学导引". Finish the online exercise there.

✏ **Exercise 2** Consider training an AdaBoost classifier using decision stumps on the five-point data set (4 "+" samples and 1 "-" sample):



1. Which examples will have their weights increased at the end of the first iteration? Circle them.
2. How many iterations will it take to achieve zero training error? Explain by doing some computation using the above algorithm.
3. Can you add one more sample to the training set so that AdaBoost will achieve zero training error in two steps? If not, explain why.

✏ **Exercise 3** (Hierarchical Clustering)

Suppose that we have four observations, for which we compute a dissimilarity matrix,

given by

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

For instance, the dissimilarity between the first and second observations is 0.3, and the dissimilarity between the second and fourth observations is 0.8.

1. On the basis of this dissimilarity matrix, sketch the dendrogram that results from hierarchically clustering these four observations using complete linkage. Be sure to indicate on the plot the height at which each fusion (merge) occurs, as well as the observations corresponding to each leaf in the dendrogram.
2. Repeat 1, this time using single linkage clustering.
3. Suppose that we cut the dendrogram obtained in 1 such that two clusters result. Which observations are in each cluster?
4. Suppose that we cut the dendrogram obtained in 2 such that two clusters result. Which observations are in each cluster?
5. It is mentioned in the chapter that at each fusion in the dendrogram, the position of the two clusters being fused can be swapped without changing the meaning of the dendrogram. Draw a dendrogram that is equivalent to the dendrogram in 1, for which two or more of the leaves are repositioned, but for which the meaning of the dendrogram is the same.

✏ **Exercise 4**(Out-of-bag error of random forest approaches its leave-one-out CV error)

Consider the regression problem with data $\mathbf{Z} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$. The bootstrap aggregation draws N samples from $\mathbf{Z}$ *independently with replacement.* Let $\hat{P}$ be the empirical distribution putting equal probability $1/N$ on each of the data points (x_i, y_i), i.e., the cumulative distribution function of $\hat{P}$ is $F_{\hat{P}}(x, y) = \frac{1}{N}\sum_{i=1}^{N} I(x_i \leqslant x, y_i \leqslant y)$. For each group of bootstrap data $\mathbf{Z}^{*b} = \{(x_1^{*b}, y_1^{*b}), (x_2^{*b}, y_2^{*b}), \ldots, (x_N^{*b}, y_N^{*b})\}$, $b = 1, 2, \ldots, B$, it follows that $(x_i^{*b}, y_i^{*b}) \sim \hat{P}$. We can fit a univariate linear model $\hat{f}^{*b}(x) = \hat{w}_0^{*b} + \hat{w}_1^{*b} x$.

1. Show that the OLS fitted coefficients are

$$\hat{w}_0^{*b} = \frac{\sum_{i=1}^{n}(x_i^{*b})^2 \sum_{i=1}^{n} y_i^{*b} - \sum_{i=1}^{n} x_i^{*b} y_i^{*b} \sum_{i=1}^{n} x_i^{*b}}{n\sum_{i=1}^{n}(x_i^{*b})^2 - (\sum_{i=1}^{n} x_i^{*b})^2}$$

$$\hat{w}_1^{*b} = \frac{n\sum_{i=1}^{n} x_i^{*b} y_i^{*b} - \sum_{i=1}^{n} x_i^{*b} \sum_{i=1}^{n} y_i^{*b}}{n\sum_{i=1}^{n}(x_i^{*b})^2 - (\sum_{i=1}^{n} x_i^{*b})^2}$$

2. The bagging estimate is defined by

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x).$$

Show that as $B \to \infty$, $\frac{1}{B}\sum_{b=1}^{B} \hat{w}_i^{*b} \xrightarrow{P} \mathrm{E}_{\hat{P}}\, \hat{w}_i^*$, where $f^*(x) = \hat{w}_0^* + \hat{w}_1^* x$ is a univariate linear model fitted from $Z^* = \{(x_1^*, y_1^*), (x_2^*, y_2^*), \ldots, (x_N^*, y_N^*)\}$ with each $(x_i^*, y_i^*) \sim \hat{P}$. (Hint: Use law of large number.)

(**Remark**: Therefore, $\hat{f}_{bag}(x) \xrightarrow{P} \mathrm{E}_{\hat{P}} \hat{f}^*(x)$ as $B \to \infty$.)

3. The leave-one-out fitted model $\hat{f}^{(-i)}(x)$ is defined as the fitted function from the data set $\mathbf{Z}\backslash\{(x_i, y_i)\}$. Repeat the definition in 1, what is the definition of the bagging estimate of $\hat{f}^{(-i)}(x)$? And what is its limit as $B \to \infty$?

4. (Optional) Let $L(y, f)$ be the loss function for the model f and the target y. The out-of-bag estimate on the sample x_i is defined as

$$\hat{f}_{oob}(x_i) = \frac{1}{|C^{(-i)}|} \sum_{b \in C^{(-i)}} \hat{f}^{*b}(x_i)$$

The out-of-bag error is defined by using the definition of out-of-bag samples:

$$\widehat{Err}_{oob} = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}_{oob}(x_i)),$$

where $C^{(-i)}$ is the set of indices of the bootstrap samples b that do not contain observation i, and $|C^{(-i)}|$ is the number of such samples.

The leave-one-out cross-validation error for the model $\hat{f}$ is defined as

$$\widehat{Err}_{CV} = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}^{(-i)}(x_i))$$

Please show that leave-one-out CV error of bagging estimate $\hat{f}_{bag}$ approaches its OOB error $\widehat{Err}_{oob}$ in probability as $B \to \infty$, and they share the same limit $\frac{1}{N}\sum_{i=1}^{N} L(y_i, \mathrm{E}_{\hat{P}^{(-i)}} \hat{f}^{(-i)*}(x_i))$, where $\hat{P}^{(-i)}$ is the empirical distribution putting equal probability $1/(N-1)$ on each of the data points except for (x_i, y_i).
(**Remark**: As a result, for bagging and random forest, we can use OOB error instead of CV error to validate the model and determine the tuning parameters.)