MA234 · Introduction to Big Data Science

Assignment 4

刘乐奇 / 12011327

April 27, 2022

1 Exercise 1

It was done on the website.

2 Exercise 2

1. At the end of the first iteration, the weight of samples in the red circle will be increased.

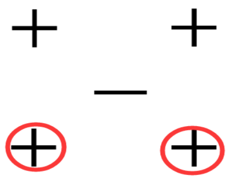

Figure 1: First iteration

2. Initialize the weights of the five samples as $w_i^{(1)} = \frac{1}{5}, i = 1,2,3,4,5.$
First iteration:

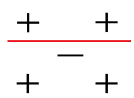

Figure 2: First iteration

$$\varepsilon_1 = \frac{\sum_{i=1}^{5} w_i^{(1)} I(y_i \neq G(x_i))}{\sum_{i=1}^{5} w_i^{(1)}} = 0.4000$$

$$\alpha_1 = \ln(\frac{1-\varepsilon_1}{\varepsilon_1}) = 0.4054$$

$$w_1^{(2)} = 0.20000, \quad w_2^{(2)} = 0.20000, \quad w_3^{(2)} = 0.20000, \quad w_4^{(2)} = 0.29998, \quad w_5^{(2)} = 0.29998$$

$f_1(x) = \frac{1}{2}\alpha_1 G_1(x)$, $sign(f_1(x))$ has 2 misclassified point.
Second iteration:

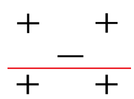

Figure 3: Second iteration

$$\varepsilon_2 = \frac{\sum_{i=1}^{5} w_i^{(2)} I(y_i \neq G(x_i))}{\sum_{i=1}^{5} w_i^{(2)}} = 0.3333$$

$$\alpha_2 = \ln(\frac{1-\varepsilon_1}{\varepsilon_1}) = 0.6933$$

$$w_1^{(2)} = 0.40006, \quad w_2^{(2)} = 0.40006, \quad w_3^{(2)} = 0.20000, \quad w_4^{(2)} = 0.29998, \quad w_5^{(2)} = 0.29998$$

$f_2(x) = \frac{1}{2}\alpha_1 G_1(x) + \frac{1}{2}\alpha_2 G_2(x)$, $sign(f_2(x))$ has no misclassified point.
Therefore, it needs 2 iterations.

3. Yes. The data set will be

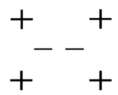

Figure 4: Adding one more sample

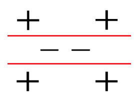

Figure 5: Classification boundary

3 Exercise 3

1. For complete linkage

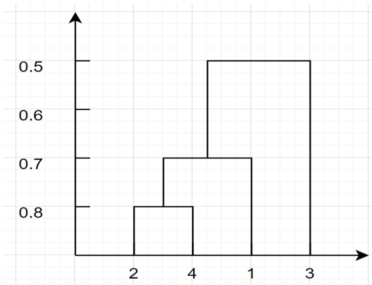

Figure 6: Complete linkage

2. For complete linkage

3. The two clusters result is $\{\{2,4\},\{1\}\}$ and $\{3\}$.

4. The two clusters result is $\{\{2,1\},\{3\}\}$ and $\{4\}$.

5. After swapped

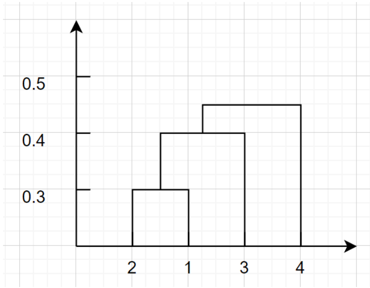

Figure 7: Single linkage

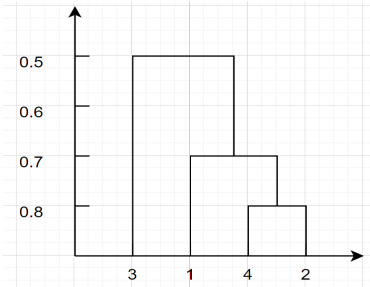

Figure 8: Swapped

4 Exercise 4

1. Let $\varepsilon = \sum_{i=1}^{n}(y_i^{*b} - \hat{w}_0^{*b} - \hat{w}_1^{*b}x_i^{*b})^2$, we want to minimize it.
   Let

$$\frac{\partial \varepsilon}{\partial \hat{w}_0^{*b}} = -2\sum_{i=1}^{n}(y_i^{*b} - \hat{w}_0^{*b} - \hat{w}_1^{*b}x_i^{*b}) = 0$$
$$\frac{\partial \varepsilon}{\partial \hat{w}_0^{*b}} = -2\sum_{i=1}^{n}x_i^{*b}(y_i^{*b} - \hat{w}_0^{*b} - \hat{w}_1^{*b}x_i^{*b}) = 0$$

   Then we have

$$\hat{w}_0^{*b} = \frac{\sum_{i=1}^{n}(x_i^{*b})^2\sum_{i=1}^{n}y_i^{*b} - \sum_{i=1}^{n}x_i^{*b}y_i^{*b}\sum_{i=1}^{n}x_i^{*b}}{n\sum_{i=1}^{n}(x_i^{*b})^2 - (\sum_{i=1}^{n}x_i^{*b})^2}$$
$$\hat{w}_1^{*b} = \frac{n\sum_{i=1}^{n}x_i^{*b}y_i^{*b} - \sum_{i=1}^{n}x_i^{*b}\sum_{i=1}^{n}y_i^{*b}}{n\sum_{i=1}^{n}(x_i^{*b})^2 - (\sum_{i=1}^{n}x_i^{*b})^2}$$

2. Since $(x_i^{*b}, y_i^{*b}) \sim \hat{P}$ and (x_i^{*b}, y_i^{*b}) is independent with each other for $i = 1, 2, \ldots, N$, then by Sinchin's law of large number, we have

$$\lim_{B\to\infty}\mathbf{P}\{|\frac{1}{B}\sum_{b=1}^{B}\hat{w}_i^{*b} - E_{\hat{P}}\hat{w}_i| < \varepsilon\} = 1$$

   Thus, $B \to \infty$, $\frac{1}{B}\sum_{b=1}^{B}\hat{w}_i^{*b} \xrightarrow{P} E_{\hat{P}}\hat{w}_i$.

3. For data $\mathbf{Z}\backslash\{x_i, y_i\}$, each group of bootstrap data is $\mathbf{Z}^{*b} = \{(x_1^{*b}, y_1^{*b}), (x_2^{*b}, x_1^{*b}), \ldots, (x_{N-1}^{*b}, y_{N-1}^{*b})\}$ for $b = 1, 2, \ldots, B$, which follows that $(x_j^{*b}, y_j^{*b}) \sim \hat{P}$, where the cumulative distribution of $\hat{P}$ is $F_{\hat{P}(x,y)=\frac{1}{N-1}\sum_{j=1,j\neq i}^{N}I(x_j\leq x, y_j\leq y)}$. We can fit $\hat{f}^{(-i)*b}(x) = \hat{w}_0^{*b} + \hat{w}_1^{*b}x$
   Then the definition of the bagging estimate of $\hat{f}^{(-i)}(x)$ is

$$\hat{f}^{(-i)}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{(-i)*b}(x)$$

Similarly, by law of large number, as $B \to \infty$, $\hat{f}^{(-i)}(x) \to E_{\hat{P}}\hat{f}^{(-i)*}(x)$, where $\hat{f}^{(-i)*}(x) = \hat{w}_0^* + \hat{w}_1^* x$.

4.