# Hw5

2022年5月3日 下午 7:57

# Homework 5

Machine Learning

Due on May 17 at 11:59 AM on Canvas

## 1 Implement Decision tree and AbaBoost algorithms manually

### 1.1 Decision Tree (15 pts)

We plan to train a decision tree model on the following dataset to predict whether an email is a spam or not. We have reprocessed the data and the binary-valued features indicate whether I know the author, whether the email is long, and whether it contain certain key words. Finally, the last column indicates whether the email is determined as a spam. ($y = +1$ for "spam" and $y = -1$ for "ham")

| $x_1$ | $x_2$ ✓ | $x_3$ | $x_4$ | $x_5$ | $y$ |
|---|---|---|---|---|---|
| know author? | is long? | has "research" | has "grade" | has "lottery" | spam? |
| 0 | 0 | 1 | 1 | 0 | +1 |
| 1 | 1 | 0 | 1 | 0 | +1 |
| 0 | 1 | 1 | 1 | 1 | +1 |
| 0 | 1 | 0 | 0 | 0 | +1 |
| 0 | 1 | 0 | 0 | 0 | +1 |
| 1 | 0 | 1 | 1 | 1 | -1 |
| 0 | 0 | 1 | 0 | 0 | -1 |
| 1 | 0 | 0 | 0 | 0 | -1 |
| 1 | 0 | 1 | 1 | 0 | -1 |
| 1 | 1 | 1 | 1 | 1 | +1 |

(i) Which feature should we choose first to spilt the data using the information gain as the splitting criterion?

(ii) Grow the decision tree to the fullest (i.e., no more feature/samples left or no more information gain for spliting) manually using the information gain as spliting criterion. Draw the decision tree and the predicted labels for each leaf node.

(i). "O" Denotes samples with label "+1";
while "□" Denotes samples with label "-1".

$$I(S,x_1) = -\left(\tfrac{3}{5}\log_2\tfrac{3}{5} + \tfrac{2}{5}\log_2\tfrac{2}{5}\right) + \underset{\downarrow 1}{\tfrac{1}{2}}\left(\tfrac{2}{5}\log_2\tfrac{2}{5} + \tfrac{3}{5}\log_2\tfrac{3}{5}\right) + \underset{\downarrow 0}{\tfrac{1}{2}}\left(\tfrac{1}{5}\log_2\tfrac{1}{5} + \tfrac{4}{5}\log_2\tfrac{4}{5}\right) = 0.971 - 0.485 - 0.361 = \boxed{0.125}$$

$$I(S,x_2) = -\left(\tfrac{3}{5}\log_2\tfrac{3}{5} + \tfrac{2}{5}\log_2\tfrac{2}{5}\right) + \tfrac{1}{2}\left(\tfrac{1}{5}\log_2\tfrac{1}{5} + \tfrac{4}{5}\log_2\tfrac{4}{5}\right) = 0.971 - 0.361 = \boxed{0.610}$$ ☆. So we choose $x_2$.

$$I(S,x_3) = -\left(\tfrac{3}{5}\log_2\tfrac{3}{5} + \tfrac{2}{5}\log_2\tfrac{2}{5}\right) - \tfrac{3}{5}\times 1 + \tfrac{2}{5}\left(\tfrac{1}{4}\log_2\tfrac{1}{4} + \tfrac{3}{4}\log_2\tfrac{3}{4}\right) = 0.971 - 0.6 - 0.325 = \boxed{0.046.}$$

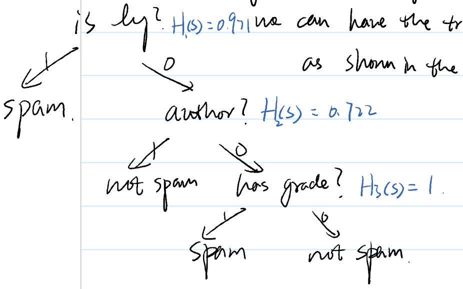
$$I(S,x_4) = -\left(\tfrac{3}{5}\log_2\tfrac{3}{5} + \tfrac{2}{5}\log_2\tfrac{2}{5}\right) + \tfrac{3}{5}\times\left(\tfrac{1}{3}\log_2\tfrac{1}{3} + \tfrac{2}{3}\log_2\tfrac{2}{3}\right) = 0.971 - 0.549 = \boxed{0.422}$$

$$I(S,x_5) = -\left(\tfrac{3}{5}\log_2\tfrac{3}{5} + \tfrac{2}{5}\log_2\tfrac{2}{5}\right) + \tfrac{3}{10}\times\left(\tfrac{1}{3}\log_2\tfrac{1}{3} + \tfrac{2}{3}\log_2\tfrac{2}{3}\right) + \tfrac{7}{10}\left(\tfrac{3}{7}\log_2\tfrac{3}{7} + \tfrac{4}{7}\log_2\tfrac{4}{7}\right)$$

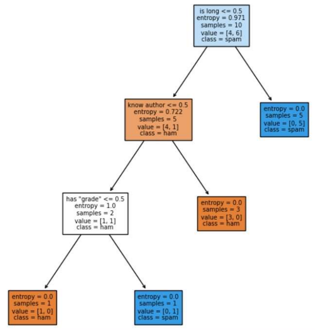
$$= 0.971 - 0.274 - 0.690 = \boxed{0.007}$$

(ii). We then divide the dataset into two groups ("is long?" or not "is long") and follow the same process

we can have the tree as shown in the left.



With validation of the python program:



### 1.2 AdaBoost (15 pts)

We will apply the AdaBoost algorithm on the following dataset with the weak learners of the form (i) "$x \geq \theta_x$" **or** (ii) "$y \geq \theta_y$" for some integers $\theta_x$ and $\theta_y$ (either one of the two forms), i.e.,

$$\text{label} = \begin{cases} + & \text{if } x \geq \theta_x \\ - & \text{otherwise} \end{cases} \quad \textbf{or} \text{ label} = \begin{cases} + & \text{if } y \geq \theta_y \\ - & \text{otherwise} \end{cases}$$

| $i$ | $x$ | $y$ | Label | |
|---|---|---|---|---|
| 1 | ~~1~~ | ~~10~~ | − | |
| 2 | ~~4~~ | ~~4~~ | − | |
| 3 | ~~8~~ | ~~7~~ | + | |
| 4 | ~~5~~ | ~~6~~ | − | |
| 5 | ~~3~~ | ~~16~~ | − | |
| 6 | ~~7~~ | ~~7~~ | + | |
| 7 | ~~10~~ | ~~14~~ | + | |
| 8 | ~~4~~ | ~~2~~ | − | |
| 9 | ~~4~~ | ~~10~~ | + | 0.25 |
| 10 | ~~8~~ | ~~8~~ | − | 0.25 |

(i) First round, $w_i^1 = \frac{1}{10}$.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| label | − | − | − | − | + | − | + | − | + | + |
| x | 1 | 3 | 4 | 4 | 4 | 5 | 7 | 8 | 8 | 10 |
| $w_i^2$ | 0.0625 | 0.0625 | 0.0625 | 0.0625 | 0.2500 | 0.0625 | 0.0625 | 0.2500 | 0.0625 | 0.0625 |
| label | − | − | − | + | + | − | − | + | + | − |
| y | 2 | 4 | 6 | 7 | 7 | 8 | 10 | 10 | 14 | 16 |
| $w_j^2$ | 0.0625 | 0.0625 | 0.0625 | 0.0625 | 0.0625 | 0.25 | 0.0625 | 0.25 | 0.0625 | 0.0625 |

$$h_1(x) = \begin{cases} 1, & x \geq 6 \\ -1, & x < 6. \end{cases}$$

$\text{rate}_1 = \frac{1}{5}$.

$\text{err}_1 = 0.1 \times 1 + 0.1 \times 1 = 0.2$

$$h_1(y) = \begin{cases} 1, & x \geq 6.5 \\ -1, & x < 6.5 \end{cases}$$

$\text{rate}_1' = \frac{3}{10} > \text{rate}_1 = \frac{1}{5}$.

So, the weak learner $h_1(x)$ is better

the weight of $h_1(x)$, $\alpha_1 = \frac{1}{2}\ln\frac{1-\text{err}_1}{\text{err}_1} = 0.6931$.

(i) Start the first round with a uniform distribution $D_1$ over the data. Find the weak learner $h_1$ that can minimize the weighted misclassification rate and predict the data samples using $h_1$.

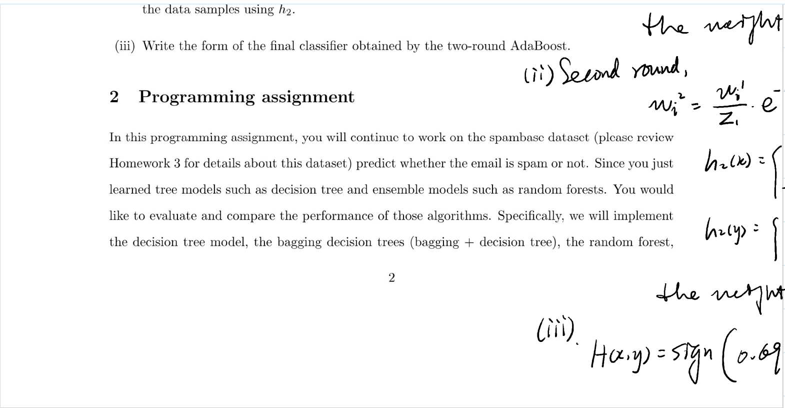
(ii) Update the weight of each data sample, denoted by $D_2$, based on the results in (1). Find the weak learner $h_2$ that can minimize the weighted misclassification rate with $D_2$, and predict the data samples using $h_2$.

(iii) Write the form of the final classifier obtained by the two-round AdaBoost.

(ii) Second round,

$w_i^2$ $w_i^1$ $e^{-\alpha_1 y_i h_1(x_i)}$ $w_i^1$ $e^{-0.6931 y_i h_1(x_i)}$.

## 2 Programming assignment



the data samples using $h_2$.

(iii) Write the form of the final classifier obtained by the two-round AdaBoost.

## 2 Programming assignment

In this programming assignment, you will continue to work on the spambase dataset (please review Homework 3 for details about this dataset) predict whether the email is spam or not. Since you just learned tree models such as decision tree and ensemble models such as random forests. You would like to evaluate and compare the performance of those algorithms. Specifically, we will implement the decision tree model, the bagging decision trees (bagging + decision tree), the random forest,

the weight of $h_1(x)$, $\alpha_1 = \frac{1}{2}\ln\frac{1-err_1}{err_1} = 0.6931$.

(ii) Second round,

$$w_i^2 = \frac{w_i^1}{Z_1} \cdot e^{-\alpha_1 y_i h_1(x_i)} = \frac{w_i^1}{2\sqrt{\frac{1}{5}\cdot\frac{4}{5}}} \cdot e^{-0.6931 y_i h_1(x_i)}.$$

shown as $w_i^2$ above.

$$h_2(x) = \begin{cases} 1, & x \geq 4 \\ -1, & x < 4 \end{cases}$$

$err_2 = 0.0625 \times 3 + 0.25 = 0.4375$

$$h_2(y) = \begin{cases} 1, & x \geq 10 \\ -1, & x < 10. \end{cases}$$

$err_2' = 0.0625 \times 4 = 0.25. < err_2.$

So, the weak learner $h_2(y)$ is better

the weight of $h_2(y)$, $\alpha_2 = \frac{1}{2}\ln\frac{1-err_2'}{err_2'} = 0.5493$.

(iii).

$$H(x,y) = \text{sign}\left(0.6931\, h_1(x) + 0.5493\, h_2(y)\right).$$