Homework 3

2021年10月9日 星期六 下午3:06

Problem 1

1. First, we plug in $y=0$ in $P[y|x;w]$:

$$P[y|x;w] = P[y=1|x;w]^0 \cdot P[y=0|x;w]^1$$
$$= P[y=0|x;w]$$

Then, we plug in $y=1$ in $P[y|x;w]$:

$$P[y|x;w] = P[y=1|x;w]^1 \cdot P[y=0|x;w]^0$$
$$= P[y=1|x;w]$$

So $P[y|x;w] = P[y=1|x;w]^y \cdot P[y=0|x;w]^{(1-y)}$

2.

$$\log\left(\prod_{i=1}^n P[y_i|x_i;w]\right) = \sum_{i=1}^n \log P[y_i|x_i;w]$$
$$= \sum_{i=1}^n \log\left[\left(\frac{1}{1+e^{-w^Tx}}\right)^{y_i} \cdot \left(\frac{1}{1+e^{w^Tx}}\right)^{(1-y_i)}\right]$$
$$= \sum_{i=1}^n y_i \log\left[\frac{1}{1+e^{-w^Tx}}\right] + (1-y_i)\log\left[\frac{1}{1+e^{w^Tx}}\right]$$
$$= -\sum_{i=1}^n y_i \log(1+e^{-w^Tx}) + (1-y_i)\log(1+e^{w^Tx})$$

$$\arg\min_{w \in \mathbb{R}^d} \sum_{i=1}^n y_i \log(1+e^{-w^Tx}) + (1-y_i)\log(1+e^{w^Tx})$$

3.

$$w_{t+1} = w_t - \eta_t\left(\sum_{i=1}^n (w_t^T x_i - y_i)x_i + 2\lambda w_t\right)$$

$$\frac{\partial}{\partial w} = \sum_{i=1}^n y_i \frac{-x_i e^{-w^Tx_i}}{1+e^{-w^Tx_i}} + \sum_{i=1}^n (1-y_i)\frac{x_i e^{w^Tx_i}}{1+e^{w^Tx_i}}$$
$$= \sum_{i=1}^n y_i \frac{-x_i(e^{-w^Tx_i}+1-1)}{1+e^{-w^Tx_i}} + \sum_{i=1}^n (1-y_i)\frac{x_i(e^{w^Tx}+1-1)}{1+e^{w^Tx_i}}$$
$$= \sum_{i=1}^n -x_i y_i (1-\sigma(w^Tx_i)) + \sum_{i=1}^n (1-y_i)x_i(1-\sigma(-w^Tx_i))$$
$$= \sum_{i=1}^n -x_i y_i \sigma(-w^Tx_i) + \sum_{i=1}^n (1-y_i)x_i\sigma(w^Tx_i)$$
$$= \sum_{i=1}^n -x_i y_i \sigma(-w^Tx_i) + \sum_{i=1}^n -x_i y_i \sigma(w^Tx_i) + x_i\sigma(w^Tx_i)$$
$$= \sum_{i=1}^n -x_i y_i \sigma(-w^Tx_i) - x_i y_i \sigma(w^Tx_i) + x_i\sigma(w^Tx_i)$$
$$= \sum_{i=1}^n -x_i y_i[\sigma(-w^Tx_i) + \sigma(w^Tx_i)] + x_i\sigma(w^Tx_i)$$
$$= \sum_{i=1}^n -x_i y_i \cdot 1 + x_i\sigma(w^Tx_i)$$
$$= \sum_{i=1}^n [x_i(\sigma(w^Tx_i) - y_i)]$$

correctly classified:

We know that when $y_i = 1$, the confidence should be $\sigma(w^Tx_i)$, So the difference between $\sigma(w^Tx_i)$ and y_i will be small. That means the weight of the x_i is small, also the summation will be small. Therefore, the contribution of the correctly classified example to updated solution in every iteration of GD

misclassified example:

To the contrary, if that example is misclassified. the confidence should be far from $\sigma(w^Tx_i)$, So the difference will be significant. And the summation will be large in the end. It will lead bigger contribution than correctly classified example.

Problem 2

label ↓ feature ↓

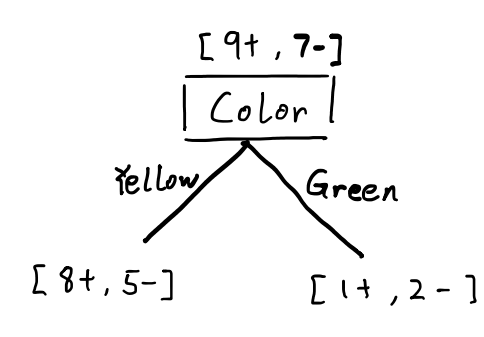
$$H(Y|X) = \sum_{x \in values(X)} p(X=x)H(Y|X=x)$$

$$H(X) = -\sum_{x \in values(X)} p(x)\log_2(p(x))$$
$$= -[p\log_2 p + (1-p)\log_2(1-p)]$$

1. Attributes and their values:
   - Color: Yellow, Green
   - Size: Small, Large
   - Shape: Round, Irregular

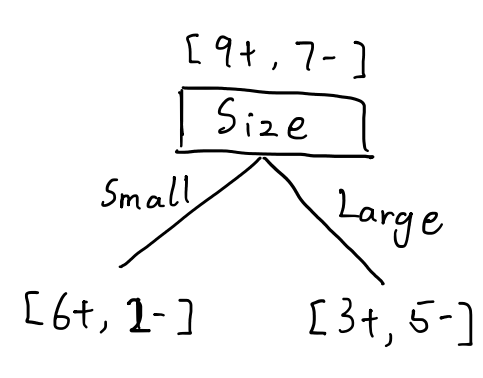
Target label: {Yes, No}

Splitting for mushroom data



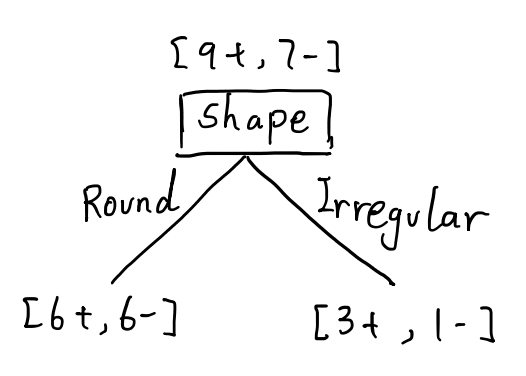
$$H(Y) = -\left[\frac{9}{16}\log_2\left(\frac{9}{16}\right) + \frac{7}{16}\log_2\left(\frac{7}{16}\right)\right] = 0.989$$
$$H(Y|Yellow) = -\left[\frac{8}{13}\log_2\left(\frac{8}{13}\right) + \frac{5}{13}\log_2\left(\frac{5}{13}\right)\right] = 0.961$$
$$H(Y|Green) = -\left[\frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right] = 0.918$$

$$InfoGain(Color) = H(Y) - H(Y|Color)$$
$$= 0.989 - \left[\frac{13}{16}H(Y|Yellow) + \frac{3}{16}H(Y|Green)\right]$$
$$= 0.989 - \left[\frac{13}{16}(0.961) + \frac{3}{16}(0.918)\right]$$
$$= 0.036$$



$$H(Y|Small) = -\left[\frac{6}{8}\log_2\left(\frac{6}{8}\right) + \frac{2}{8}\log_2\left(\frac{2}{8}\right)\right] = 0.811$$
$$H(Y|Large) = -\left[\frac{3}{8}\log_2\left(\frac{3}{8}\right) + \frac{5}{8}\log_2\left(\frac{5}{8}\right)\right] = 0.954$$

$$InfoGain(Size) = H(Y) - H(Y|Size)$$
$$= 0.989 - \left[\frac{8}{16}(0.811) + \frac{8}{16}(0.954)\right]$$
$$= 0.107 \checkmark$$

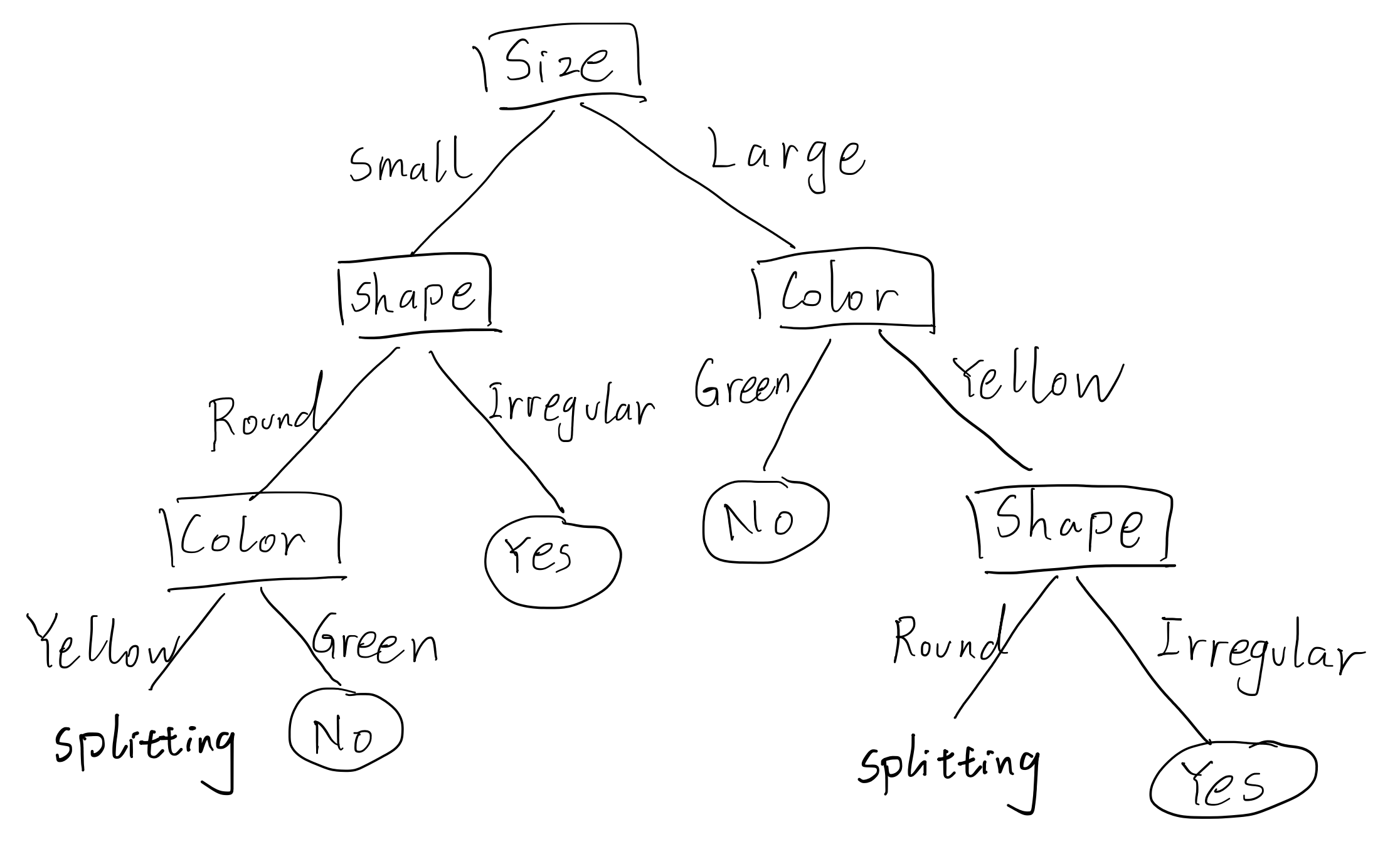


$$H(Y|Round) = -\left[\frac{6}{12}\log_2\left(\frac{6}{12}\right) + \frac{6}{12}\log_2\left(\frac{6}{12}\right)\right] = 1$$
$$H(Y|Irregular) = -\left[\frac{3}{4}\log_2\left(\frac{3}{4}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right] = 0.811$$

$$InfoGain(Shape) = H(Y) - H(Y|Shape)$$
$$= 0.989 - \left[\frac{12}{16}(1) + \frac{4}{16}(0.811)\right] = 0.036$$

"Size" would the algorithm choose to use for the root of tree

2.



Problem 3

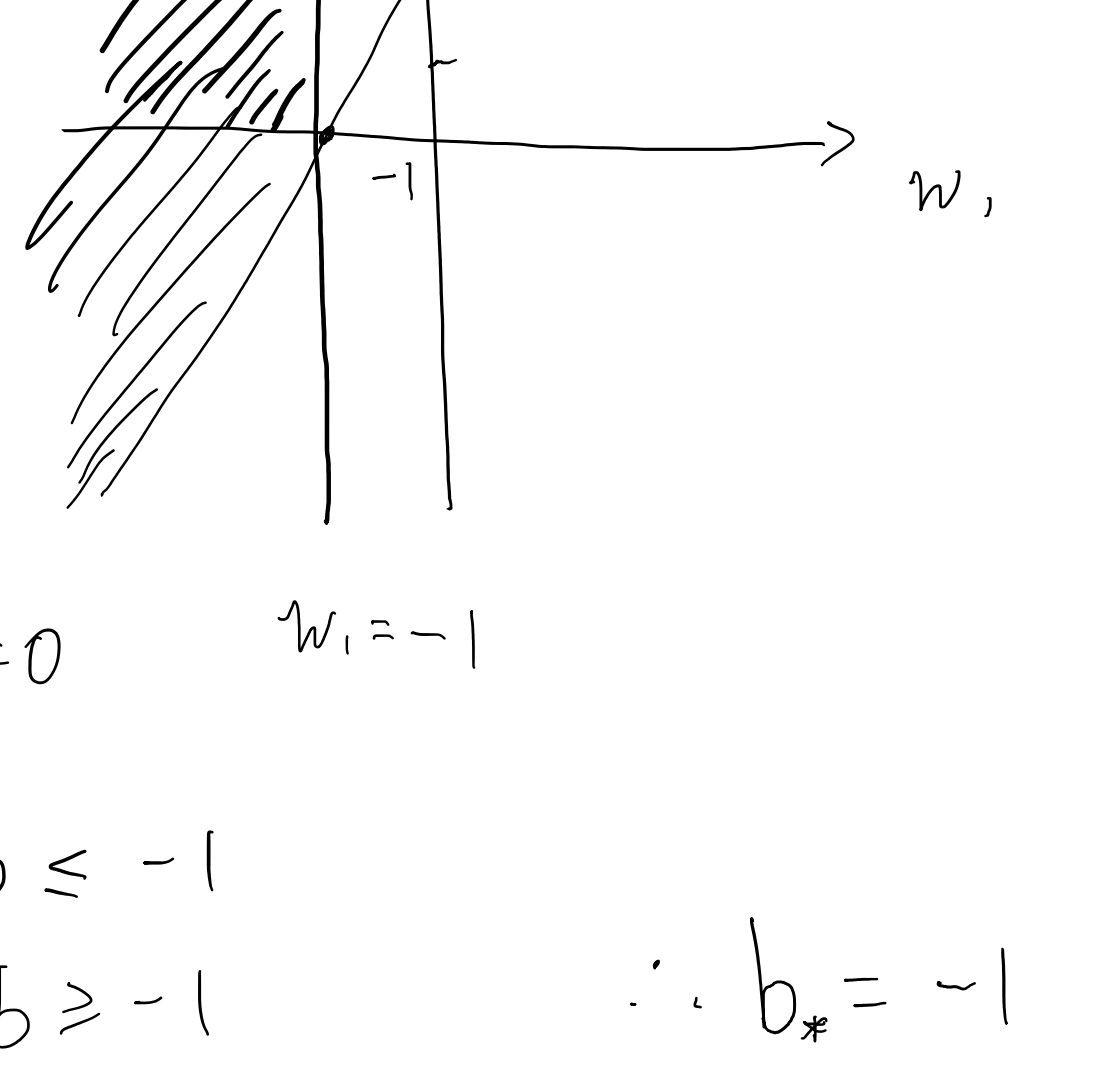
There is a possibility of overfitting. During the design of decision trees, overfitting occurs when all samples from the training set are perfectly fitted into the tree. Therefore, it should branches with sparse data.

Problem 4

$$\vec{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

$$\begin{cases} (-1)(w_1(0) + w_2(0) + b) \geq 1 \\ (-1)(w_1(0) + w_2(-1) + b) \geq 1 \\ (+1)(w_1(-2) + w_2(0) + b) \geq 1 \end{cases}$$

$$\begin{cases} -b \geq 1 & ① \\ w_2 - b \geq 1 & ② \\ -2w_1 + b \geq 1 & ③ \end{cases}$$



① + ③ ⟶ $-2w_1 \geq 2$
$w_1 \leq -1$

② + ③ ⟶ $w_2 - b - 2w_1 + b \geq 2$
$-2w_1 + w_2 \geq 2$

To $\min_{w,b} \frac{1}{2}\|w\|_2^2$, $w_1 = -1$, $w_2 = 0$

$$\begin{cases} -b \geq 1 \\ 0 - b \geq 1 \\ 2 + b \geq 1 \end{cases} \quad \begin{cases} -b \geq 1 \\ 2 + b \geq 1 \\ b \leq -1 \end{cases} \quad \begin{cases} b \leq -1 \\ b \geq -1 \end{cases} \quad \therefore b_* = -1$$

optimal hyperplane $\left(\begin{bmatrix} -1 \\ 0 \end{bmatrix}, -1\right)$

margin:

$$margin(w, b, S) = \min_{x_i \in S} d(x_i, w, b)$$
$$= 1$$