MA234 · Introduction to Big Data Science

Assignment 2

刘乐奇 / 12011327

April 27, 2022

1 Exercise 1

It was finished on the website.

2 Exercise 2

1. The initial entropy of "Appealing" is

$$Entropy(t_0) = -\frac{5}{10}log_2\frac{5}{10} = \frac{1}{2} \tag{1}$$

2. "Taste" is the root of the decision tree. Then the information gain is

$$\begin{aligned} InfoGain &= Entropy(t_0) - \sum_{k=1}^{K}\frac{n_k}{n}Entropy(t_k) \\ &= \frac{1}{2} - (\frac{0}{3}log_2\frac{0}{3} + \frac{3}{3}log_2\frac{3}{3}) - (\frac{2}{4}log_2\frac{2}{4} + \frac{2}{4}log_2\frac{2}{4}) - (\frac{3}{3}log_2\frac{3}{3} + \frac{0}{3}log_2\frac{0}{3}) \\ &= \frac{1}{2} - 0 + 1 - 0 \\ &= 1.5 \end{aligned} \tag{2}$$

3. The full decision tree without pruning is as below:

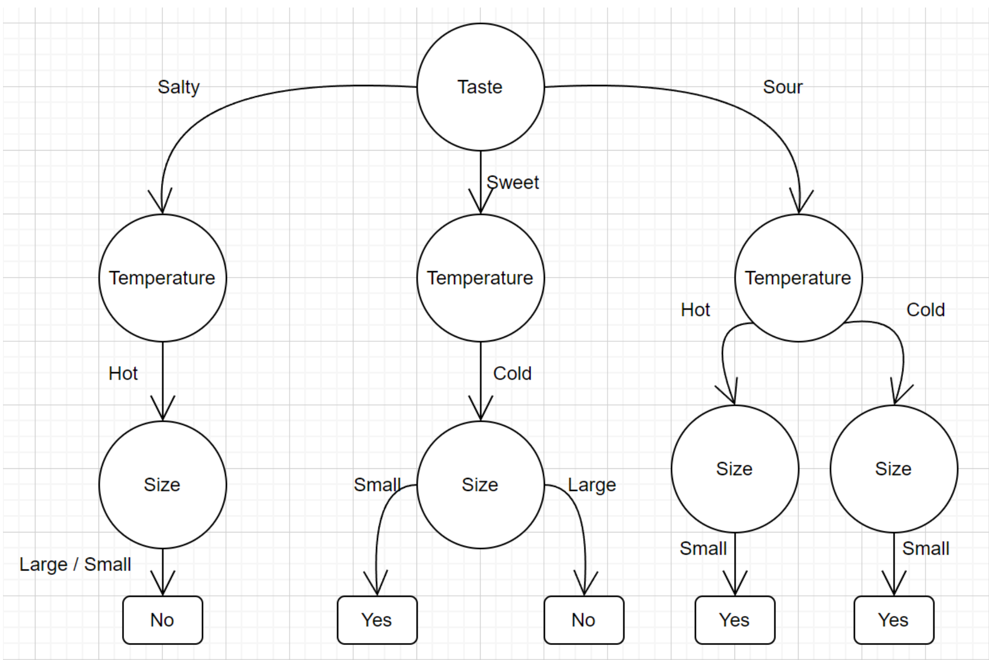

Figure 1: Full decision tree without pruning

3 Exercise 3

1. Since

$$f_\theta = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{1}{2\sigma^2}(x-\mu)^2) \tag{3}$$

then we have

$$\begin{aligned} lnL(\theta) &= ln\prod_{i=1}^{n} f_\theta(x_i) = \sum_{i=1}^{n} lnf_\theta(x_i) \\ &= \sum_{i=1}^{n}(-\frac{1}{2}ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_i-\mu)^2) \\ &= -\frac{n}{2}ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \end{aligned} \tag{4}$$

So

$$\frac{\partial lnL(\theta)}{\partial\mu} = \frac{\mu}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = \frac{\mu}{\sigma^2}(\sum_{i=1}^{n}x_i - n\mu) \tag{5}$$

$$\frac{\partial lnL(\theta)}{\partial\sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 \tag{6}$$

Let equation (3) and (4) equal to 0 and combine them, then

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}x_i,\ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

2.

$$E\hat{\mu} = E(\frac{1}{n}\sum_{i=1}^{n}x_i) = \frac{1}{n}\sum_{i=1}^{n}E(x_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}n\mu = \mu \tag{7}$$

$$\begin{aligned} E(\frac{n}{n-1}\hat{\sigma}^2) &= E(\frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2) = \frac{1}{n-1}E(\sum_{i=1}^{n}(x_i-\overline{x})^2) = \frac{1}{n-1}E(\sum_{i=1}^{n}(x_i-\mu+\mu-\overline{x})^2) \\ &= \frac{1}{n-1}E(\sum_{i=1}^{n}((x_i-\mu)^2 - 2(x_i-\mu)(\overline{x}-\mu) + (\overline{x}-\mu)^2)) \\ &= \frac{1}{n-1}E(\sum_{i=1}^{n}(x_i-\mu)^2 - 2\sum_{i=1}^{n}(x_i-\mu)(\overline{x}-\mu) + n(\overline{x}-\mu)^2) \\ &= \frac{1}{n-1}E(\sum_{i=1}^{n}(x_i-\mu)^2 - 2n(\overline{x}-\mu)^2 + n(\overline{x}-\mu)^2) \\ &= \frac{1}{n-1}E(\sum_{i=1}^{n}(x_i-\mu)^2 - n(\overline{x}-\mu)^2) \\ &= \frac{1}{n-1}(\sum_{i=1}^{n}E(x_i-\mu)^2 - nE(\overline{x}-\mu)^2) \\ &= \frac{1}{n-1}(nVar(x) - nVar(\overline{x})) \\ &= \frac{n}{n-1}(\sigma^2 - \frac{\sigma^2}{n}) = \sigma^2 \end{aligned} \tag{8}$$

4 Exercise 4

For prior probability, we have

$$P(Y=k) = p_k^{I(y=k)}(1-p_k)^{I(y\neq k)} \tag{9}$$

$$
\begin{aligned}
L(P) &= arg\max_{p_k} \sum_{y\in Y} ln(p_k^{I(y=k)}(1-p_k)^{I(y\neq k)}) \\
&= arg\max_{p_k} (\sum_{y=k} lnp_k + \sum_{y\neq k} ln(1-p_k)) \\
&= arg\max_{p_k} (n_1 lnp_k + n_2 ln(1-p_k)) \quad (n_1 = \sum_{i=1}^{n} I(y=k), n_2 = \sum_{i=1}^{n} I(y\neq k))
\end{aligned} \tag{10}
$$

$$
\begin{aligned}
\frac{\partial L}{\partial p_k} &= \frac{\partial(n_1 lnp_k + n_2 ln(1-p_k))}{\partial p_k} \\
&= \frac{n_1}{p_k} - \frac{n_2}{1-p_k}
\end{aligned} \tag{11}
$$

Let equation (11) be zero, then we can get

$$
\hat{p}_k = \frac{n_1}{n_1+n_2} = \frac{\sum_{i=1}^{n} I(y=k)}{n} \tag{12}
$$

For conditional probability, we have

$$
P(X=s|Y=k) = \frac{P(X=s, Y=k)}{P(Y=k)} \tag{13}
$$

Since we have found the value of $P(Y=k)$, we need to find the MLE for $P(X=s, Y=k)$.

$$
\begin{aligned}
L(P) &= arg\max_{p} \sum_{x\in X, y\in Y} ln(p^{I(x=s,y=k)}(1-p)^{1-I(x=s,y=k)}) \\
&= arg\max_{p} (n_1 lnp + n_2 ln(1-p)) \quad (n_1 = \sum_{i=1}^{n} I(x=s, y=k), n_2 = \sum_{i=1}^{n}(1 - I(x=s, y=k)))
\end{aligned} \tag{14}
$$

$$
\begin{aligned}
\frac{\partial L}{\partial p} &= \frac{\partial(n_1 lnp + n_2 ln(1-p))}{\partial p} \\
&= \frac{n_1}{p} - \frac{n_2}{1-p}
\end{aligned} \tag{15}
$$

Let equation (15) be zero, then we can get $\hat{p} = \frac{n_1}{n_1+n_2} = \frac{\sum_{i=1}^{n} I(x=s,y=k)}{n}$. To sum up, the MLE of p_{sk} is

$$
\hat{p}_{sk} = \frac{\hat{p}}{\hat{p}_k} = \frac{\sum_{i=1}^{n} I(x=s, y=k)}{\sum_{i=1}^{n} I(y=k)} \tag{16}
$$

5 Exercise 5

$$
\begin{aligned}
E_{S\sim P^n}\mathcal{E}(f^{1NN}) &= E_{S\sim P^n} E_{(\mathbf{X},Y)\sim P} 1_{Y\neq f^{1NN}} \\
&= E_{S\sim P^n} E_{(\mathbf{X},Y)\sim P} 1_{Y\neq y_{\pi_S(\mathbf{X})}} \\
&= E_{S\sim P^n} (1-\eta(\mathbf{X})) p_{\mathbf{X}}(\mathbf{X}) \\
&= E_{S\sim P^n} [\int_P \eta(\mathbf{X})(1-\eta(\mathbf{X}') + \eta(\mathbf{X}')(1-\eta(\mathbf{X}) dx]
\end{aligned} \tag{17}
$$

$$
2\mathcal{E}(f^*) + cE_{S\sim P^n} E_{\mathbf{X}\sim p_{\mathbf{X}}} ||\mathbf{X} - \mathbf{X}_{\pi_S(\mathbf{X})}|| \geq E_{S\sim P^n}[2\int_P |\eta(\mathbf{X}) - \eta(\mathbf{X}_{\pi_S(\mathbf{X})})| dx + \int_P |\eta(\mathbf{X}) - \eta(\mathbf{X}')| dx] \tag{18}
$$

So we need to prove that

$$
a(1-b) + b(1-a) \leq 2|a - a_\pi| + |a-b| \tag{19}
$$

where $a, b \in [0,1]$.
Let $f(b) = 2|a - a_\pi| + |a-b| + 2ab - a - b$, we have $\min_b f(b) = min f(0), f(1), f(a) = min 2|a-a_\pi|, 2|a-a_\pi|, 2(a^2 - a + |a - a_\pi|$
0. Therefore, it was proved.