Homework 1 for Advanced Machine Learning

Shanghao Shi

September 14, 2019

This is the writing part homework of the machine learning class CS5824. This homwork contains two sections, the first one is the answer of question 1 and the second section is the answer for question 2. In order to make my homework more readable, this homework will restate the question before giving out its answer.

1 Decision tree

1.1 Question

Show what the recursive decision tree learning algorithm would choose for the first split of the following dataset:

ID	X_1	X_2	X_3	X_4	Y
1	0	0	0	0	0
2	0	0	0	1	0
3	0	0	1	0	0
4	0	0	1	1	0
5	0	1	0	0	0
6	0	1	0	1	1
7	0	1	1	0	1
8	0	1	1	1	1
9	1	0	0	0	1
10	1	0	0	1	1

Assume that the criterion for deciding the best split is entropy reduction (i.e., information gain). If there are any ties, choose the first feature to split on tied for the best score. Show your calculations in your response.

1.2 Answer

First we will calculate the entrophy of the whole dataset $\mathcal{D}$. There are 5 samples having the Y values 1 and another 5 samples having Y values 0. Thus the entrophy of the whole dataset $\mathcal{D}$ is:

$$H(D) = \sum_{i=1}^{2} -p_i log(p_i) = 0.5log(2) + 0.5log(2) = 1bit \tag{1}$$

Then we choose feature X_1 to be the first decision rule. We will see the information gain of this feature. Using this rule, we divide the dataset $\mathcal{D}$ into two dataset $\mathcal{D}_1$ and $\mathcal{D}_2$. The information gain is:

$$\begin{aligned} infoGain(X_1) &= H(\mathcal{D}) - \{\frac{|\mathcal{D}_1|}{|\mathcal{D}|}H(\mathcal{D}_2) + \frac{|\mathcal{D}_2|}{|\mathcal{D}|}H(\mathcal{D}_2)\} \\ &= 1 - \frac{8}{10}\{\frac{5}{8}log(\frac{8}{5}) + \frac{3}{8}log(\frac{8}{3})\} = 0.2365bit \end{aligned} \tag{2}$$

Using the same method, we then calculate the information gain of feature X_2, X_3, X_4:

$$\begin{aligned} infoGain(X_2) &= H(\mathcal{D}) - \{\frac{|\mathcal{D}_1|}{|\mathcal{D}|}H(\mathcal{D}_2) + \frac{|\mathcal{D}_2|}{|\mathcal{D}|}H(\mathcal{D}_2)\} \\ &= 1 - \frac{6}{10}\{\frac{4}{6}log(\frac{6}{4}) + \frac{2}{6}log(\frac{6}{2})\} - \frac{4}{10}\{\frac{1}{4}log(\frac{4}{1}) + \frac{3}{4}log(\frac{4}{3})\} = 0.1245bit \end{aligned} \tag{3}$$

$$
\begin{aligned}
infoGain(X_3) &= H(\mathcal{D}) - \{\frac{|\mathcal{D}_1|}{|\mathcal{D}|}H(\mathcal{D}_2) + \frac{|\mathcal{D}_2|}{|\mathcal{D}|}H(\mathcal{D}_2)\} \\
&= 1 - \frac{6}{10}\{\frac{2}{4}log(\frac{4}{2}) + \frac{2}{4}log(\frac{4}{2})\} - \frac{4}{10}\{\frac{3}{6}log(\frac{6}{3}) + \frac{3}{6}log(\frac{6}{3})\} = 0bit
\end{aligned} \tag{4}
$$

$$
\begin{aligned}
infoGain(X_4) &= H(\mathcal{D}) - \{\frac{|\mathcal{D}_1|}{|\mathcal{D}|}H(\mathcal{D}_2) + \frac{|\mathcal{D}_2|}{|\mathcal{D}|}H(\mathcal{D}_2)\} \\
&= 1 - \frac{5}{10}\{\frac{2}{5}log(\frac{5}{2}) + \frac{3}{5}log(\frac{5}{3})\} - \frac{5}{10}\{\frac{2}{5}log(\frac{5}{2}) + \frac{3}{5}log(\frac{5}{3})\} = 0.0290bit
\end{aligned} \tag{5}
$$

Compare the four information gains we get above, we find that X_1 is the best decision rule.

Then we come to see dataset $\mathcal{D}_1$ and dataset $\mathcal{D}_2$. All data in $\mathcal{D}_2$ are in the same class, so we just need to consider $\mathcal{D}_1$. There are 8 examples in $\mathcal{D}_1$. 5 of them belong to class 0 and 3 of them belong to class 1. So the entrophy of $\mathcal{D}_1$ is:

$$
H(D_1) = \sum_{i=1}^{2} -p_i log(p_i) = \frac{5}{8}log(\frac{8}{5}) + \frac{3}{8}log(\frac{8}{3}) = 0.9544bit \tag{6}
$$

We will also choose another best decision rule for $\mathcal{D}_1$. This rule will divide $\mathcal{D}_1$ to two dataset $\mathcal{D}_3$ and $\mathcal{D}_4$. Calculate the information gain of X_2, X_3, X_4 to get this decision.

$$
\begin{aligned}
infoGain(X_2) &= H(\mathcal{D}_1) - \{\frac{|\mathcal{D}_3|}{|\mathcal{D}_1|}H(\mathcal{D}_3) + \frac{|\mathcal{D}_4|}{|\mathcal{D}_1|}H(\mathcal{D}_4)\} \\
&= 0.9544 - \frac{4}{8}\{\frac{1}{4}log(\frac{4}{1}) + \frac{3}{4}log(\frac{4}{3})\} = 0.5488bit
\end{aligned} \tag{7}
$$

$$
\begin{aligned}
infoGain(X_3) &= H(\mathcal{D}_1) - \{\frac{|\mathcal{D}_3|}{|\mathcal{D}_1|}H(\mathcal{D}_3) + \frac{|\mathcal{D}_4|}{|\mathcal{D}_1|}H(\mathcal{D}_4)\} \\
&= 0.9544 - \frac{4}{8}\{\frac{1}{4}log(\frac{4}{1}) + \frac{3}{4}log(\frac{4}{3})\} - \frac{4}{8}\{\frac{2}{4}log(\frac{4}{2}) + \frac{2}{4}log(\frac{4}{2})\} = 0.0488bit
\end{aligned} \tag{8}
$$

$$
\begin{aligned}
infoGain(X_4) &= H(\mathcal{D}_1) - \{\frac{|\mathcal{D}_3|}{|\mathcal{D}_1|}H(\mathcal{D}_3) + \frac{|\mathcal{D}_4|}{|\mathcal{D}_1|}H(\mathcal{D}_4)\} \\
&= 0.9544 - \frac{4}{8}\{\frac{1}{4}log(\frac{4}{1}) + \frac{3}{4}log(\frac{4}{3})\} - \frac{4}{8}\{\frac{2}{4}log(\frac{4}{2}) + \frac{2}{4}log(\frac{4}{2})\} = 0.0488bit
\end{aligned} \tag{9}
$$

From the result we can see that feature X_2 is the second best decision rule. This rule will divide the dataset into Two different parts $\mathcal{D}_3$ and $\mathcal{D}_4$. Both $\mathcal{D}_3$ and $\mathcal{D}_4$ contain 4 examples. All exmples in $\mathcal{D}_4$ are in the same class and we do not have to consider it. But in $\mathcal{D}_3$, 3 examples belong to class 1 and 1 example belongs to class 0. Therefore we can get that the entrophy of this dataset is:

$$
H(D_3) = \sum_{i=1}^{2} -p_i log(p_i) = \frac{1}{4}log(\frac{4}{1}) + \frac{3}{4}log(\frac{4}{3}) = 0.8113bit \tag{10}
$$

Using the same method we calculte the information gain of each feature. Using these features, dataset $\mathcal{D}_3$ is divided to two dataset $\mathcal{D}_5$ and $\mathcal{D}_6$.

$$
\begin{aligned}
infoGain(X_3) &= H(\mathcal{D}_3) - \{\frac{|\mathcal{D}_5|}{|\mathcal{D}_3|}H(\mathcal{D}_5) + \frac{|\mathcal{D}_6|}{|\mathcal{D}_3|}H(\mathcal{D}_6)\} \\
&= 0.8113 - \frac{2}{4}\{\frac{1}{2}log(\frac{2}{1}) + \frac{1}{2}log(\frac{2}{1})\} = 0.3113bit
\end{aligned} \tag{11}
$$

$$
\begin{aligned}
infoGain(X_4) &= H(\mathcal{D}_3) - \{\frac{|\mathcal{D}_5|}{|\mathcal{D}_3|}H(\mathcal{D}_5) + \frac{|\mathcal{D}_6|}{|\mathcal{D}_3|}H(\mathcal{D}_6)\} \\
&= 0.8113 - \frac{2}{4}\{\frac{1}{2}log(\frac{2}{1}) + \frac{1}{2}log(\frac{2}{1})\} = 0.3113bit
\end{aligned} \tag{12}
$$

Two information gain are the same and we choose this first feature X_3 as our best decision rule.

Finally, we will chose X_4 as the last decision rule. Having all of our decstion rules, we can get our decision tree and it is shown in the following picture.

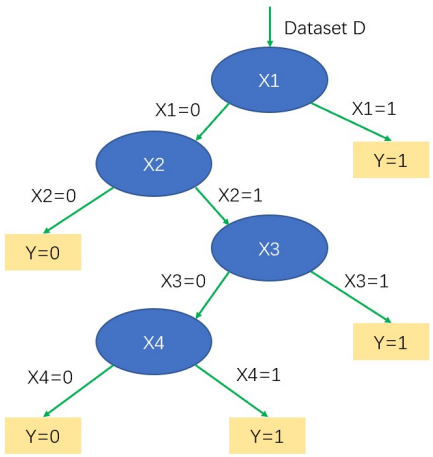

Figure 1: Decision Tree

2 Bayes analysis

2.1 Question

A Bernoulli distribution has the following likelihood function for a data set $\mathcal{D}$:

$$p(\mathcal{D}|\theta) = \theta^{N_1}(1-\theta)^{N_0}, \tag{13}$$

where N_1 is the number of instances in data set $\mathcal{D}$ that have value 1 and N_0 is the number in $\mathcal{D}$ that have value 0. The maximum likelihood estimate is

$$\hat{\theta} = \frac{N_1}{N_1 + N_0}. \tag{14}$$

1. Derive the maximum likelihood estimate above by solving for the maximum of the likelihood.

2. Suppose we now want to maximize a posterior likelihood

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}, \tag{15}$$

where we use the Bernoulli likelihood and a (slight variant[1] of a) symmetric Beta prior over the Bernoulli parameter

$$p(\theta) \propto \theta^{\alpha}(1-\theta)^{\alpha}. \tag{16}$$

Derive the maximum posterior mean estimate.

2.2 Answer

2.2.1 1

The maximum likelihood estimate of θ is the value that maximizing likelihood function. We have know that the likelihood function is $p(\mathcal{D}|\theta) = \theta^{N_1}(1-\theta)^{N_0}$. Thus we will derive on likelihood function to find its extremum.

$$\begin{aligned} \frac{\mathrm{d}p(\mathcal{D}|\theta)}{\mathrm{d}\theta} &= N_1\theta^{N_1-1}(1-\theta)^{N_0} - N_0\theta^{N_1}(1-\theta)^{N_0-1} \\ &= N_1(1-\theta\frac{N_1+N_0}{N_1})\theta^{N_1-1}(1-\theta)^{N_0-1} \end{aligned} \tag{17}$$

[1]For convenience, we are using the exponent of α instead of the standard $\alpha - 1$.

From the equation above, we can see that when $\theta = 0, 1, \frac{N_1}{N_1+N_0}$, $p'(\mathcal{D}|\theta)$ has the value of zero.

Further, we can also find that when $0 < \theta < \frac{N_1}{N_1+N_0}$, the derivative funtion of likelihood function $p'(\mathcal{D}|\theta)$ is strictly positive and when $\frac{N_1}{N_1+N_0} < \theta < 1$, the derivative funtion of likelihood function $p'(\mathcal{D}|\theta)$ is strictly negative. Thus we get that likelihood function reaches its maximum at point $\theta = \frac{N_1}{N_1+N_0}$. Therefore, the maximum likelihood estimate of θ is $\hat{\theta} = \frac{N_1}{N_1+N_0}$.

2.2.2 2

As $p(\mathcal{D})$ will not change when dataset $\mathcal{D}$ is fixed, the value of θ that maximum posterior likelihood $p(\theta|\mathcal{D})$ is the same with the value that maximum $p(\mathcal{D}|\theta)p(\theta)$. Thus we will only consider $p(\mathcal{D}|\theta)p(\theta)$ instead of $p(\theta|\mathcal{D})$. We will derive $p(\mathcal{D}|\theta)p(\theta)$ to see the result.

$$\begin{aligned}\frac{\mathrm{d}p(\mathcal{D}|\theta)p(\theta)}{\mathrm{d}\theta} &= \frac{\mathrm{d}\theta^{N_1+\alpha}(1-\theta)^{N_0+\alpha}}{\mathrm{d}\theta} \\ &= (\alpha+N_1)\theta^{N_1+\alpha-1}(1-\theta)^{N_0} - (\alpha+N_0)\theta^{N_1+\alpha}(1-\theta)^{N_0+\alpha-1} \\ &= (N_1+\alpha)(1-\theta\frac{N_1+N_0+2\alpha}{N_1+\alpha})\theta^{N_1-1}(1-\theta)^{N_0-1}\end{aligned} \tag{18}$$

From the equation above we can get that when $\theta = 0, 1$ and $\frac{N_1+\alpha}{N_1+N_0+2\alpha}$, $\frac{\mathrm{d}p(\mathcal{D}|\theta)p(\theta)}{\mathrm{d}\theta} = 0$.So only at these points can the function get its extremum.

When $0 < \theta < \frac{N_1+\alpha}{N_1+N_0+2\alpha}$, $\frac{\mathrm{d}p(\mathcal{D}|\theta)p(\theta)}{\mathrm{d}\theta}$ is strictly positive and function is monotone increasing. When $\frac{N_1+\alpha}{N_1+N_0+2\alpha} < \theta < 1$, $\frac{\mathrm{d}p(\mathcal{D}|\theta)p(\theta)}{\mathrm{d}\theta}$ is strictly negative and the function is monotone decreasing. Thus the function get its maximum at the point $\theta = \frac{N_1+\alpha}{N_1+N_0+2\alpha}$. So the maximum posterior estimate of parameter θ is $\hat{\theta} = \frac{N_1+\alpha}{N_1+N_0+2\alpha}$.