Statistical Inference and Machine Learning | Homework 2

Team members : Xia Shengzhao, Chica Linares Andrés, Gloria Dal Santo, Loïc Piccot, Elio Ovide Sanchez

November 25, 2020

1 Feature Selection

1.1 Answer

The definition of mutual information can be reformulated as follows:

$$\begin{aligned}
I(X_i, Y) &= H(Y) - H(Y|X_i) \\
&= -\sum_{y\in\mathcal{Y}} p(y)\log_2 p(y) + \sum_{y\in\mathcal{Y}, x_i\in\mathcal{X}} p(x_i, y)\log_2 \frac{p(x_i, y)}{p(x_i)} \\
&= -\sum_{y\in\mathcal{Y}} \left(\sum_{x_i\in\mathcal{X}} p(x_i, y)\right)\log_2 p(y) + \sum_{y\in\mathcal{Y}}\sum_{x_i\in\mathcal{X}} p(x_i, y)\log_2 \frac{p(x_i, y)}{p(x_i)} \\
&= \sum_{y\in\mathcal{Y}, x_i\in\mathcal{X}} p(x_i, y)\log_2 \frac{p(x_i, y)}{p(y)p(x_i)}
\end{aligned}$$

1.2 Answer

To show that the mutual information is symmetric we can proceed as follows:

$$\begin{aligned}
I(X_i, Y) &= H(Y) - H(Y|X_i) \qquad (1) \\
&= \sum_{y\in\mathcal{Y}, x_i\in\mathcal{X}} p(x_i, y)\log_2 \frac{p(x_i, y)}{p(y)p(x_i)} \\
&= -\sum_{y\in\mathcal{Y}, x_i\in\mathcal{X}} p(x_i, y)\log_2 p(x_i) + \sum_{y\in\mathcal{Y}, x_i\in\mathcal{X}} p(x_i, y)\log_2 \frac{p(x_i, y)}{p(y)} \\
&= -\sum_{x_i\in\mathcal{X}} \left(\sum_{y\in\mathcal{Y}} p(x_i, y)\right)\log_2 p(x_i) - H(X_i|Y) \\
&= -\sum_{x_i\in\mathcal{X}} p(x_i)\log_2 p(x_i) - H(X_i|Y) \\
&= H(X_i) - H(X_i|Y) = I(Y, X_i)
\end{aligned}$$

1.3 Answer

To evaluate the Information Gains according to eq.(1) we have computed the entropy $H(Y)$, where Y is the label 'play', as follows:

$$H(Y) = -\sum p_i \log_2 p_i = -\left[\frac{9}{14}\log_2\left(\frac{9}{14}\right) + \frac{5}{14}\log_2\left(\frac{5}{14}\right)\right] = 0.9403$$

Then, each conditional entropy $H(Y|X_i)$ was found according to the formula already mentioned above:

$$H(Y|X_i) = \sum_{y \in \mathcal{Y}, x_i \in \mathcal{X}} p(x_i, y) \log_2 \frac{p(x_i, y)}{p(y)}$$

The obtained information gains are reported in Table 1:

Table 1: Informativeness of Different Features

Features	Outlook	Temp	Humidity	Wind
Informativeness	0.2467	0.0292	0.1518	0.0481

From the table, we can see that the reduction in uncertainty of feature **Outlook** is most significant, which means **Outlook** is the most informative feature.

2 Decision Trees

As showed in exercise 1, we know that the feature Outlook has the biggest mutual information. That is why we set it as the root node of the tree. We then divide our dataset into 3 subsets (Tables 2, 4 and 5) according to the feature's values. We will repeat the mutual information computation of all features from each subset. We will use the features with maximum value to use them as decision nodes.

2.1 First subset: Outlook = sunny

The following table shows the new subset when the outlook feature's value is sunny.

Table 2: First subset : Outlook = sunny

Day	Temperature	Humidity	Wind	Play
1	hot	high	strong	no
2	hot	high	weak	no
8	mild	high	weak	no
9	cool	normal	weak	yes
11	mild	normal	strong	yes

We first compute the mutual information of the first subset's features:

$$H(Y) = -\sum p_i \log_2 p_i = -\left[\frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right] = 0.971$$

Table 3: Mutual information results for first subset

Feature	H(Y)	H(Y|X)	I(X,Y)
Temperature		0.4	0.571
Humidity	0.971	0	0.971
Wind		0.951	0.02

In this case, humidity has the most important mutual information so we make it a decision node. Moreover we can see that it branches immediately tells us the output.

2.2 Second subset: Outlook = overcast

The following table shows the new subset when the outlook feature's value is overcast.

Table 4: Second subset : Outlook = overcast

Day	Temperature	Humidity	Wind	play
3	hot	high	weak	yes
7	cool	normal	strong	yes
12	mild	high	strong	yes
13	hot	normal	weak	yes

Concerning the second subset (Table 4), we can observe that when the value of Outlook is overcast, the output is yes, regardless of the other features values. That is why no computation is needed for this subset and the overcast branch directly finishes on a leaf node.

2.3 Third subset: Outlook = rain

The following table shows the new subset when the outlook feature's value is rain.

Table 5: Third subset : Outlook = rain

Day	Temperature	Humidity	Wind	play
4	mild	high	weak	yes
5	cool	normal	weak	yes
6	cool	normal	strong	no
10	mild	normal	weak	yes
14	mild	high	strong	no

Finally, we compute the mutual information of the third subset's features:

$$H(Y) = -\sum p_i \log_2 p_i = -\left[\frac{3}{5}\log_2\left(\frac{3}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right] = 0.971$$

Table 6: Mutual information result for third subset

Feature	H(Y)	H(Y|X)	I(X,Y)
Temperature		0.951	0.02
Humidity	0.971	0.951	0.02
Wind		0	0.971

The wind feature has the biggest mutual information so it will be a decision node. Taking a look at the third subset (Table 5), we even notice that knowing the value of this feature is sufficient to determinate the output.

We can now plot the decision tree showed in Fig.1. The ID3 algorithm makes it the simplest decision tree concerning our dataset.

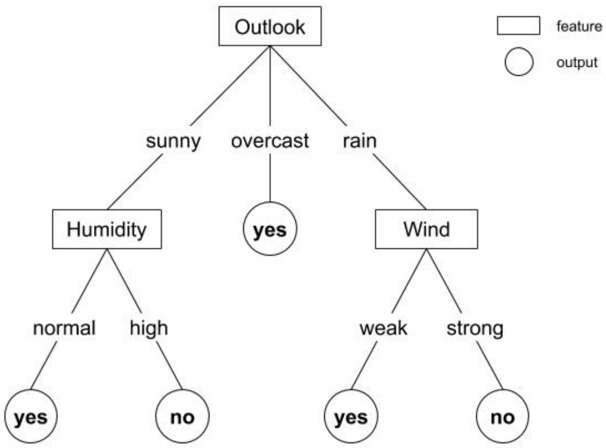

Figure 1: Decision tree generated using ID3 algorithm