Statistical Inference and Machine Learning

MGT-448

Homework 2

Feature Selection & Decision Trees

Students
Student1 Alexis Couturier
Student2 Erick Maraz Zuniga
Student3 Alon Tchelet
Student4 Mustafa Yildirim

Professor
Negar Kiyavash

Abstract

Second homework assignment of MGT-448: Statistical Inference and Machine Learning course covering the topics of Feature Selection and Decision Trees.

January 31, 2022

1 Feature Selection

Let (X, Y) be a pair of random variables with values over space $\mathcal{X} \times \mathcal{Y}$. If their joint distribution is $P_{(X,Y)}$ and the marginal distributions are P_X and P_Y, the mutual information is defined as :

$$I(X, Y) = H(Y) - H(Y|X) \tag{1}$$

where $H(\cdot)$ and $H(\cdot|\cdot)$ are respectively the entropy and conditional entropy defined as :

$$H(Y) = -\sum_{y \in \mathcal{Y}} P_Y(y) \log_2 P_Y(y) \tag{2}$$

$$H(Y|X) = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{(X,Y)}(x, y) \log_2 \frac{P_{(X,Y)}(x, y)}{P_X(x)} \tag{3}$$

1. Using Eq.1,2 and 3 the canonical expression of mutual information I is derived as :

$$\begin{aligned}
I(X, Y) &= -H(Y|X) + H(Y) \\
&\underset{(1)}{=} -\sum_{x \in \mathcal{X}} P_X(x) H(Y|X = x) - \sum_{y \in \mathcal{Y}} P_Y(y) \log_2 P_Y(y) \\
&\underset{(2)}{=} \sum_{x \in \mathcal{X}} P_X(x) \Big(\sum_{y \in \mathcal{Y}} P_{Y|X=x}(y) \log_2 P_{Y|X=x}(y) \Big) - \sum_{y \in \mathcal{Y}} \Big(\sum_{x \in \mathcal{X}} P_{(X,Y)}(x, y) \Big) \log_2 P_Y(y) \\
&= \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_X(x) P_{Y|X=x}(y) \log_2 P_{Y|X=x}(y) - \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{(X,Y)}(x, y) \log_2 P_Y(y) \\
&\underset{(3)}{=} \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{(X,Y)}(x, y) \log_2 \left(\frac{P_{(X,Y)}(x, y)}{P_X(x)} \right) - \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{(X,Y)}(x, y) \log_2 P_Y(y) \\
&= \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{(X,Y)}(x, y) \log_2 \left(\frac{P_{(X,Y)}(x, y)}{P_X(x) P_Y(y)} \right)
\end{aligned}$$

   (1) : The first part of the expression is obtained using the total law of probability. Indeed $(\{x\})_{x \in \mathcal{X}}$ is a set of mutually exclusive and exhaustive events. The second part comes from the definition of the entropy in Eq.2.

   (2) : The first part of the expression comes from Eq.3. The second part comes from the definition of the marginal law for a pair of random variables.

   (3) : Both transformations in the first summation come from the chain rule of events : $P(A \cap B) = P(A) \cdot P(B|A)$.

2. Using the above reasoning, the mutual information is shown to be symmetric :

$$\begin{aligned}
I(Y, X) &= -H(X|Y) + H(X) \\
&= -\sum_{y \in \mathcal{Y}} P_Y(y) H(X|Y = y) - \sum_{x \in \mathcal{X}} P_X(x) \log_2 P_X(x) \\
&= \sum_{y \in \mathcal{Y}} P_Y(y) \Big(\sum_{x \in \mathcal{X}} P_{X|Y=y}(x) \log_2 P_{X|Y=y}(x) \Big) - \sum_{x \in \mathcal{X}} \Big(\sum_{y_\in \mathcal{Y}} P_{(X,Y)}(x, y) \Big) \log_2 P_X(x) \\
&= \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_Y(y) P_{X|Y=y}(x) \log_2 P_{X|Y=y}(x) - \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{(X,Y)}(x, y) \log_2 P_X(x) \\
&= \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{(X,Y)}(x, y) \log_2 \left(\frac{P_{(X,Y)}(x, y)}{P_Y(y)} \right) - \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{(X,Y)}(x, y) \log_2 P_X(x) \\
&= \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{(X,Y)}(x, y) \log_2 \left(\frac{P_{(X,Y)}(x, y)}{P_X(x) P_Y(y)} \right) \\
&= I(X, Y)
\end{aligned}$$

3. From question 1, the mutual information of pair (X, Y) with values over space $\mathcal{X} \times \mathcal{Y}$ is given by :

$$I(X,Y) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{(X,Y)}(x,y) \log_2 \left(\frac{P_{(X,Y)}(x,y)}{P_X(x) P_Y(y)} \right) \tag{4}$$

where $P_{(X,Y)}$ is the joint distribution and P_X, P_Y the marginal distributions of variables X and Y.

In order to calculate mutual information $I(X_i, Y)$ for each feature X_i, these probability distributions need to be estimated.

Marginal distribution P_Y :

In this example, label Y is a random variable with values in $\mathcal{Y} = \{\text{yes}, \text{no}\}$ which decides whether or not to play tennis. Assuming a Bernoulli distribution $Y \sim \mathcal{B}(\phi)$, parameter ϕ is estimated as follows :

$$\phi = P(Y = \text{yes}) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{Y^{(i)} = \text{yes}\}} \quad \text{and} \quad P(Y = \text{no}) = 1 - \phi \tag{5}$$

where $\mathbb{1}_{\{\cdot\}}$ denotes the indicator function of set $\{\cdot\}$

Marginal distribution P_X :

Similarly, assuming a multinomial distribution for features X_i with values in $\mathcal{X} = \{\text{sunny, overcast, rain}\}$ or {hot, mild, cold} or {high, normal} or {weak, strong}, the parameters of the distribution are estimated as :

$$P(X_i = x) = \frac{1}{N} \sum_{j=1}^{N} \mathbb{1}_{\{X_i^{(j)} = x\}} \tag{6}$$

Joint distribution $P_{(X,Y)}$:

Finally, the joint distribution $P_{(X,Y)}$ can be obtained from $P_Y \cdot P_{X|Y=y}$ which requires the conditional distribution $P_{X|Y=y}$ to be calculated. Assuming a multinomial distribution, the parameters of the distribution can be estimated as :

$$P(X_i = x | Y = y) = \frac{\sum_{j=1}^{N} \mathbb{1}_{\{X_i^{(j)} = x\}} \cdot \mathbb{1}_{\{Y^{(j)} = y\}}}{\sum_{j=1}^{N} \mathbb{1}_{\{X_i^{(j)} = x\}}} \tag{7}$$

Using the above distributions and Eq.2 and 3, the entropy and conditional entropy on each feature is calculated. The values are shown in Table 1.

$H(Y)$	$H(Y\|\text{outlook})$	$H(Y\|\text{temperature})$	$H(Y\|\text{humidity})$	$H(Y\|\text{wind})$
0.9403	0.6935	0.9111	0.7885	0.8922

Table 1: Entropy of label vectors and conditional entropy of each feature vector

Given these values and Eq.1, the mutual information of each feature is computed and displayed in Table 2.

$I(\text{outlook}\|Y)$	$I(\text{temperature}\|Y)$	$I(\text{humidity}\|Y)$	$I(\text{wind}\|Y)$
0.2468	0.0292	0.1518	0.0481

Table 2: Mutual information of each feature vector to the labels vector

From Table 2, one can conclude that the most informative feature in the given dataset is outlook. Indeed, it has the largest mutual information with label Y.

2 Quadratic Programming

ID3 is an algorithm meant to construct decision trees. The principle behind the algorithm is as follows: Choose the most informative feature, create a branch for each possible value of this feature and repeat for further branching until a definitive label is reached for the branch. The label is sometimes referred to as a leaf. The algorithm is recursive, it progresses until it reaches a leaf and then collapse back up as described in Algorithm 1. The final tree shows sequential flow of the different possible values of the features (branches) that all terminate in a single output label (leaf).

Algorithm 1: ID3 pseudo-code algorithm

Function ID3(S, TargetAttribute)
Create RootNode for the tree;
if $\forall j \in S, \quad Y^{(j)} = C$ **then**
  RootNode = single-node tree with label C;
else if $S = \{\varnothing\}$ **then**
  RootNode = single-node tree with label most common value of TargetAttribute in Samples;
else
  $i = \arg\min_k H(Y|X_k)$;
end
i is decision attribute for RootNode;
for x_j *in* $X_i \quad \forall j$ **do**
  add a Branch below RootNode, testing for $X_i = x_j$;
  $S = \{j|X_i = x_j\}$;
  if $S = \{\varnothing\}$ **then**
    below Branch add Leaf with label most common value of TargetAttribute in Samples;
  else
    below Branch add Subtree ID3($S \setminus \{i\}$, TargetAttribute);
  end
end
return RootNode
end

where $H(Y|X)$ is a vector with the conditional entropy of each feature with the label vector, X_i is features vector i, x_j is a particular value within $\mathcal{X}_i$, Y is the label vector and S is the set of indices of the features of the data.

Running the algorithm, the following decision tree is obtained Fig. 1 :

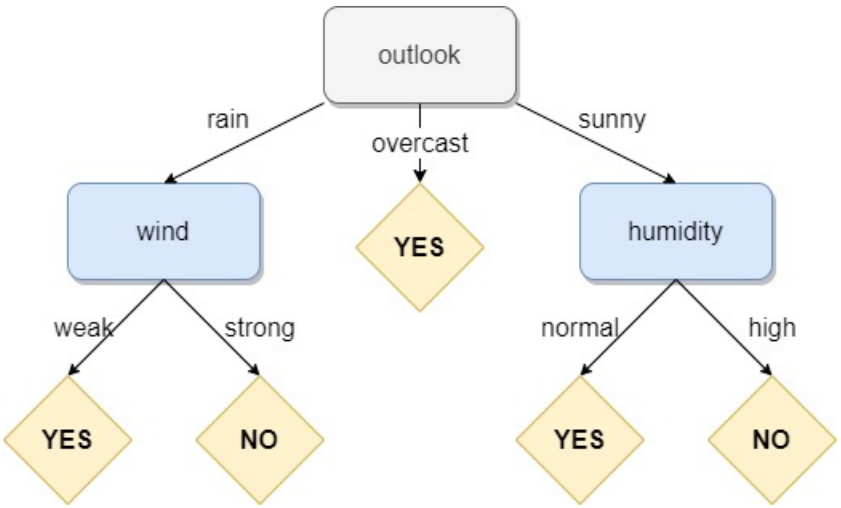

Figure 1: Decision tree from ID3 algorithm for the dataset