Machine Learning

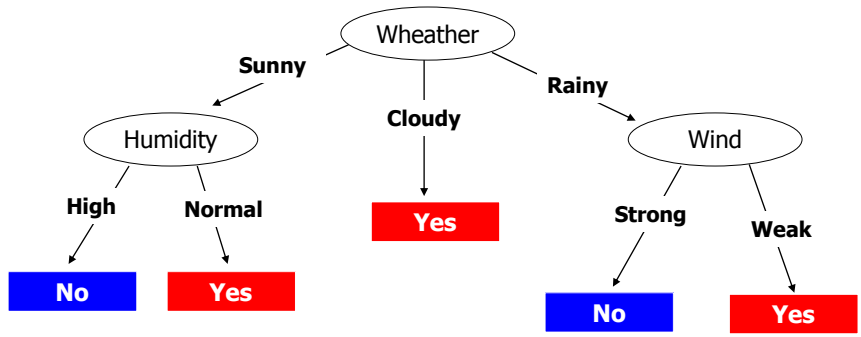
3. week

- Entropy
- Decision Trees
  - ID3
  - C4.5
- Classification and Regression Trees (CART)

What is Decision Tree

As a short description, decision tree is a data classification procedure which is made by asking questions in right order.

Decision Tree Sample



Decision Trees

It start with entropy based classification tree methods such as

- ID3
- C4.5.

But first, we should learn something about entropy, uncertainty and information.

Entropy and Uncertainty

In probability theory, **entropy** is a measure of the **uncertainty** about a random variable.

In information theory, **entropy** (Shannon) is the expected value (average) of **information** in each situation of the random variable.

According to Shannon, the proper choice for "function to measure **information**" must be logarithmic.

Uncertainty and Information

The **entropy** and **uncertainty** are maximized if all situations of the random variable have the same probability.

The difference between prior and posterior probability distributions shows the amount of **information** gained or lost (Kullback–Leibler divergence).

Accordingly, posteriors of situations with maximum **uncertainty** causes likely maximum **information** gain.

Information

In fact, both information and uncertainty are like antonyms to each other. But since maximum uncertainty can cause maximum information, information gain and uncertainty focus on the same concept. Self information is represented by

$$I(x) = \log \frac{1}{P(x)} = -\log P(x)$$

Entropy

Shannon Entropy (*H*) term is the expected value of all a_i situations with P_i probabilities.

$$H(X) = E(I(X)) = \sum_{1 \le i \le n} P(x_i).I(x_i)$$

$$= \sum_{i=1}^{n} P(x_i) \log_2 \frac{1}{P(x_i)} = -\sum_{i=1}^{n} P_i \log_2 P_i$$

Entropy

Let X be random process of a coin toss. Since its two situations have the same probability, the entropy of X is found 1 as below.

$$H(X) = -\sum_{i=1}^{2} p_i \log_2 p_i$$
$$= -(0.5 \log_2 0.5 + 0.5 \log_2 0.5) = 1$$

We can comment it as that after the process, we will win 1 bit of information.

Entropy in Decision Tree

A decision tree divides multidimensional data into several sets with a condition on specified feature.

In each case, to decide working on which feature of data and based on which condition is needed solving a big combinational problem.

Entropy in Decision Tree

Even only in a data with 20 samples and 5 features, we can construct more than 106 decision trees.

Therefore, each branching should be performed methodologically.

Entropy in Decision Tree

According to Quinlan, we should branch the tree by gaining maximum information.

For this reason, it must be found the feature which causes the minimum uncertainty.

In the ID3, all branches are examined individually, and feature causing the minimum uncertainty is preferred to make the branching on the tree.

ID3 Algorithm

It works with only categorical data.

In the first step of each iteration, entropy of the data set is computed for all features.

Then entropy of each feature depending on the class is computed, and it is subtracted from the entropy calculated in the first step.

The feature is chosen, which supports maximum information gain.

ID3 Sample

V1	V2	S
A	C	E
B	C	F
B	D	E
B	D	F

We have a dataset with 2 features (V1 and V2), a class vector (S), and 4 samples.

Find the first branching feature?

H(S) - H(V1,S)=?
H(S) - H(V2,S)=?

ID3 Sample

V1	V2	S
A	C	E
B	C	F
B	D	E
B	D	F

At first, general entropy:

$$H(S) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1$$

Entropy of V1

$$H(V1) = \frac{1}{4}H(A) + \frac{3}{4}H(B)$$

$$= \frac{1}{4}0 - \frac{3}{4}\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right)$$

$$= 0 + \frac{3}{4}0{,}9183 = 0{,}6887$$

Entropy of V2

$$H(V2) = \frac{1}{2}H(C) + \frac{1}{2}H(D) = \frac{1}{2} + \frac{1}{2} = 1$$

We choose V1.

C4.5 Algorithm

It is a form of ID3 algorithm that can be applied to numerical data.

It transforms numerical properties into categorical form by a thresholding method.

Different threshold values are tested.

The threshold value that supports the best information gain is selected.

After categorizing the feature vector with the selected threshold, ID3 can be directly applied.

C4.5 Algorithm

To determine the best threshold,

1. All numerical data are sorted (a_1, a_2,..., a_n)
2. An average value of each sorted pair is computed by $(a_i + a_{i+1})/2$
3. We have n-1 average values as thresholds
4. Each average is tested to branch data
5. The best average should support the maximum information gain.

Lost Data

If some samples are lost, there are two ways to be followed:

1. Samples with lost features are completely removed from the dataset.
2. The algorithm is arranged to operate with lost data.

If the number of lost samples is too much, then the second option should be applied.

Lost Data

When calculating the information gain for the feature vector with lost data, information gain is calculated by excluding the lost samples, and then it is multiplied by F coefficient. F coefficient is the ratio of lost data in dataset.

$$IG(X) = F.(H(X) - H(V, X))$$

Lost Data

Writing the most frequent value in feature vector with lost data into the lost cells is another one of the proposed methods.

Overfitting

As in other machine learning methods, we should avoid overfitting also in the decision tree.

If precautions are not taken, all decision trees make overfitting.

Therefore the trees should be pruned during or after creation.

Pruning

Pruning is process of removing needless parts to the classification from decision tree.

In this way, the decision tree becomes both simple and understandable.

There are two types of pruning methods;

- pre-pruning
- post-pruning

Pre-Pruning

Pre-pruning is done during creation of trees.

If divided features values is less than a threshold value (fault tolerance), the partitioning process is stopped.

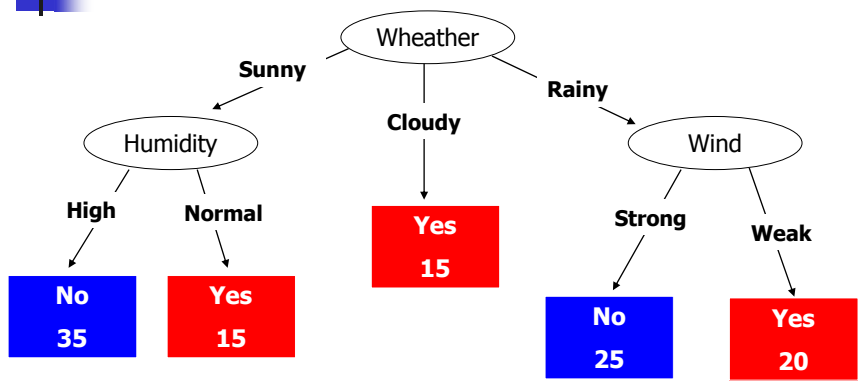
At that moment a leaf is created, which has the label of the dominant class of the available data set.

Post-Pruning

Post-pruning is done after creation of tree by

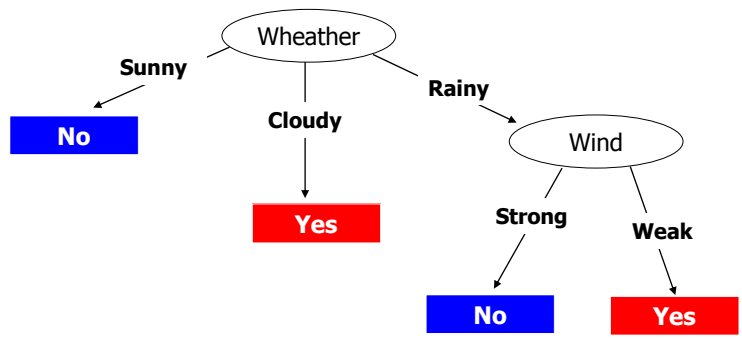
- creating a leaf instead of a sub-tree,
- rising a sub-tree,
- and removing some branches.

Pruning Sample



If fault tolerance is 33%, "Yes" ratio at sub-tree of "Humidity" node is computed as 30%. Therefore instead of "Humidity" node, a "No" leaf is created.

Pruning Sample



Classification and Regression Trees (CART)

Briefly, CART trees divides each node into 2 branches. The most known methods are:

- Twoing algorithm
- Gini algorithm

MATLAB Application

```
>edit C4_5_ornek.m
```

Student should study with this sample by using given datasets.

Presentation Task

You can choose one of two CART algorithms listed below.

- Twoing
- Gini