Simple Decision Tree Summary

Bill Roland

March 4, 2018

1 Context

Decision tree is a easy-understanding way to do classification, meanwhile it has good performance.

2 Experiment

2.1 Basic Idea

In this experiment, I use a very simple dataset [1] (which has only 20 samples) to train a decision tree.And then test its performance.

2.2 Equation Derivation

I use information gain to select attribute for splitting dataset.The account form of information gain:

$$IG(D) = Ent(D) - \sum_{v=1}^{|V|} \frac{|D_v|}{|D|} Ent(D_v) \tag{1}$$

In equation (1),D_v is a subset in which all the samples has same value v on an attribute , V is a set of values the attribute can be,and the $Ent(D)$ stands for the *Shanno information entropy*:

$$Ent(D) = -\sum_{k=1}^{|K|} p_k log_2 p_k \tag{2}$$

In equation (2) K is a set of class lables. p_k is the probability that class k appears in the dataset.

[1] http://archive.ics.uci.edu/ml/machine-learning-databases/balloons/adult-stretch.data

2.3 Conslusion

As a result, depth of this decision tree is 2(2 attributes is useless).I haven't implement puring on this simple emperiment.This decision tree can classify the dataset perfectly(i.e. has 0 error).