Decision Trees

Oliver Zhao

1 Description

Decision Trees operate on the notion of entropy, defined as

$$H(T) = -p_{pos}log_2p_{pos} - p_{neg}log_2p_{neg}. \tag{1.1}$$

Decision Trees consist of different nodes that contain decision thresholds for a given attribute at. Based on this threshold, the data set T is divided into subsets T_i. Each sequential layer of nodes leads to more divisions, ultimately resulting in leaves. All training examples in a given leaf are classified as the same class.

In order to calculated the weighted average of the entropies of the subsets T_i for a given attribute at, we first need to calculate the probability that a randomly drawn training sample is in T_i, defined as

$$P_i = \frac{|T_i|}{|T|}. \tag{1.2}$$

Then the weighted average of the entropies of the subsets can be calculated as $H(T, at)$, where

$$H(T, at) = \sum_i P_i \cdot H(T_i). \tag{1.3}$$

In order to examine how much information is *gained* $I(T, at)$ through the addition of an attribute, we compare the entropy before and after the attribute has been considered, where

$$I(T, at) = H(T) - H(T, at). \tag{1.4}$$

While the decision threshold for a given attribute is simple for discrete attributes, an infinite number of thresholds may be chosen for continuous attributes. The number of thresholds for continuous attributes may be narrowed down by ordering the values of a given attribute at in ascending order. Candidate thresholds $\theta_1, \theta_2, \ldots,$ are located between attribute values with opposite class labels, as any other additional thresholds between these candidate thresholds yield the same outcomes.

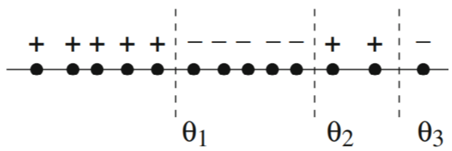

Figure 1: Example of attribute values sorted in ascending order. Note that candidate thresholds are drawn only at class boundaries.

The information gain $I(T, \theta_i)$ is then calculated for each of the candidate thresholds, and the threshold with the highest information gain is chosen as the threshold used at the node in mind.

2 Algorithm

Description: Consider a training set T divided into subsets T_i, that are characterized by a different value of an attribute at.

---

1. Find the attribute at that contributes the maximum information regarding class labels.
   (a) Calculate the entropy $H(T)$ of the training set T or subset T_i, where p_{pos} and p_{neg} are the percentages of positive and negative examples, respectively
   $$H(T) = -p_{pos} log_2 p_{pos} - p_{neg} log_2 p_{neg}. \tag{2.1}$$
   (b) For each attribute that divides T into subsets T_i, with relative sizes P_i, calculate the entropy of each subset T_i.
   (c) For each attribute, calculate the average entropy $H(T, at) = \sum_i P_i \cdot H(T_i)$.
   (d) For each attribute, calculate the information gain $I(T, at) = H(T) - H(T, at)$.
   (e) Identify the attribute with the highest value of information gain.
2. Divide T into subsets, T_i, each characterized by a different value of at.
3. For each T_i, find all examples in T_i that belong to the same class and create a leaf labeled with this class. Otherwise, apply procedures 1-3 on each training subset T_i.