Exercise: Decision Tree Construction

August 31, 2016

1 Task

Construct a decision tree given the following training data set.

Outlook	Temp.	Humidity	Windy	Play
sunny	85	85	false	No
sunny	80	90	true	No
overcast	83	78	false	Yes
rain	70	96	false	Yes
rain	68	80	false	Yes
rain	65	70	true	No
overcast	64	65	true	Yes
sunny	72	95	false	No
sunny	69	70	false	Yes
rain	75	80	false	Yes
sunny	75	70	true	Yes
overcast	72	90	true	Yes
overcast	81	75	false	Yes
rain	71	80	true	No

Outlook	categorical (sunny, overcast, rain)
Temperature	continuous
Humidity	continuous
Windy	categorical (true, false)
Play	categorical target (Yes, No)

Table 1: Attributes

2 Building a Decision Tree

2.1 Basics

- Each node corresponds to an attribute and each edge to a possible value of that attribute. A leaf of the tree specifies the expected value of the target attribute for the records described by the path from the root to that leaf.
- Each node should be associated with the attribute which is the *most informative* among the attributes not yet considered in the path from the root.
- Entropy is used to measure how informative a node is.

2.2 Algorithm (ID3)

`function ID3` (R: a set of attributes, C: the target attribute, S: a training set) returns a decision tree

- If S is empty, return a leaf node with the default class (majority class in the entire training set).
- If S consists of records all with the same value for the target attribute, return a single node with that value (this will be a leaf node).
- If R is empty, then return a single node with as value the most frequent of the values of the target attribute that are found in records of S (this will be a leaf node; note that then there will be errors, that is, records that will be improperly classified).
- Otherwise (if none of the previous conditions are met): Let D be the attribute with largest $Gain(D, S)$ among the attributes in R.[1]

$$Gain = \text{Entropy}(p) - \sum_{j=1}^{k} \frac{N(v_j)}{N} \text{Entropy}(v_j), \tag{1}$$

where k is the number of attribute values, N is the total number of records at the parent node ($= |S|$), $N(v_j)$ is the number of records associated with the child node v_j.
The Entropy for two classes ($C = No$, $C = Yes$):

$$\text{Entropy}(t) = -P(C = No|t) \cdot log_2 P(C = No|t) - P(C = Yes|t) \cdot log_2 P(C = Yes|t) \tag{2}$$

- Let $\{d_j | j = 1, 2, \ldots, m\}$ be the values of attribute D. Let $\{S_j | j = 1, 2, \ldots, m\}$ be the subsets of S consisting respectively of records with value d_j for attribute D.
- Return a tree with root labeled D and edges labeled $d_1, d_2, \ldots, d_m$ going respectively to the trees $ID3(R - \{D\}, C, S_1), ID3(R - \{D\}, C, S_2), \ldots, ID3(R - \{D\}, C, S_m)$

[1] You can use Gain Ratio instead of Gain.

3 Solution