

Dat550-exercise2

Vinay Setty

30. January 2020

1 Entropy

Consider the following dataset with 8 documents d_1 to d_8 and features/attributes f_1 to f_3 .

Doc	f_1	f_2	f_3	Class
d1	2	0	0	c_1 (Algebra)
d2	2	0	0	c_1 (Algebra)
d3	0	0	0	c_2 (Calculus)
d4	0	1	0	c_2 (Calculus)
d5	0	2	0	c_3 (Stochastics)
d6	0	2	0	c_3 (Stochastics)
d7	0	1	1	c_3 (Stochastics)
d8	0	2	1	c_3 (Stochastics)

1. What is the entropy over the categories for these training instances d1 to d8? Recall that the entropy of a partition \mathcal{T} is given as $H(\mathcal{T}) = -\sum_j P(\mathcal{T}_j) \cdot \log_2 P(\mathcal{T}_j)$. Note the \log_2 has been chosen to make the calculations simple for you. You actually do not need a calculator.
2. Using the training set d1 to d8, suppose we want to construct a decision tree for the binary classification of the category c_3 (“Stochastics”), i.e., the tree decides whether a new document belongs to c_3 category or not, using binary splits. Determine the split with the highest information gain for binary split at the root level. Recall Information gain formula

$$G(k, k_1, k_2) = H(k) - \frac{|k_1|}{|k|} H(k_1) - \frac{|k_2|}{|k|} H(k_2)$$

$$f_1 \geq 1 \quad f_2 \geq 1 \quad f_3 \geq 1$$

$$f_1 \geq 2 \quad f_2 \geq 2 \quad f_3 \geq 2$$

$$f_1 \geq 3 \quad f_2 \geq 3 \quad f_3 \geq 3$$

2 Build a Decision Tree

Construct a decision tree given the following training data set.

Table 1: Data

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	false	Yes
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

Table 2: Attributes

Outlook	categorical (sunny, overcast, rain)
Temperature	continuous (in Fahrenheit)
Humidity	continuous (percentage)
Windy	categorical (true, false)
Play	categorical class/label/target (Yes, No)

For the above task use the ID3 Algorithm

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 (You may also use gain ratio instead of gain).

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

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 ($C = No, C = Yes$):

$$Entropy(t) = -P(C = No|t) \cdot \log_2 P(C = No|t) - P(C = Yes|t) \cdot \log_2 P(C = Yes|t)$$

- Let $\{d_j | j = 1, 2, \dots, m\}$ be the values of attribute D . Let $\{S_j | j = 1, 2, \dots, m\}$ 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, \dots, d_m going respectively to the trees $ID3(RD, C, S_1), ID3(RD, C, S_2), \dots, ID3(RD, C, S_m)$