Note Title

Choosing attributes in a decision tree

In section 18.3.4 (p703), the textbook describes how to choose which attribute to split on, but only for a 2-class classification problem (i.e. with only positive and negative examples). Here we describe a generalization of that procedure for multi-class problems.

Suppose we have:

- N training samples $x_1, x_2, \ldots x_N$
- an attribute A with D distinct values, divide the training set into subsets $S_1, S_2, \ldots S_D$. The number of elements in S_d is n_d.

  $$\left(\text{So } \sum_{d=1}^{D} n_d = N \right)$$

- The proportion of training samples in S_d is π_d, so

  $$\pi_d = \frac{n_d}{N} \qquad d = 1, 2, \ldots D$$

- There are C classes : $1, 2, \ldots C$.

- The number of elements from the set S_d in class c is denoted $n_{d,c}$.

  Thus, $$\sum_{c=1}^{C} n_{d,c} = n_d \quad , \text{ for } d = 1, 2, \ldots D.$$

- The proportion of elements from the set S_d in class c is denoted $\pi_{d,c}$
Thus,
$$\pi_{d,c} = \frac{n_{d,c}}{n_d}$$

- The entropy of the distribution of classes in S_d, written H_d, can be computed as
$$H_d = \sum_{c=1}^{C} -\pi_{d,c} \log_2 \pi_{d,c}$$

- The <u>expected entropy</u> for attribute A, denoted $E(A)$ is given by
$$E(A) = \sum_{d=1}^{D} \pi_d H_d$$

We want to choose the attribute A^* with the highest <u>information gain</u>. But info gain = (current entropy) − (expected entropy), so this is equivalent to choosing the attribute with lowest expected entropy.

Thus, we choose
$$A^* = \underset{A}{\operatorname{argmin}}\ E(A)$$