Machine Learning

Kiar Fatah

February 22, 2020

Contents

If you can't explain something in simple terms, you don't understand it.

Richard Feynman

1 Lecture 2

1.1 Decision Tree

A decision tree is a map of the possible outcomes of a series of related choices.

1.2 How to ask the best questions?

The machine will automate the questions for the decision tree, however how does it obtain the best questions to ask? It does so by with the help of Gini impurity and information gain.

1.2.1 Gini impurity

Gini impurity is a measurement of uncertainty in a node. It is calculated by how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset.

$$1 - \sum_i p_i^2. \tag{1.1}$$

1.2.2 Entropy

Entropy is also a measurement of uncertainty in a node and can be applied instead of Gini impurity. It is calculated as

$$\sum_i -p_i log_2 p_i. \tag{1.2}$$

The information entropy is measured in bits, corresponding to the logarithmic base 2. Tossing a coin results in the entropy 1, hence it has 1 bit of information. For a real dice and a fake dice the value is 2.58 and 2.16 bits of information respectively.

1.2.3 Information gain

Information gain is how much the question reduces the uncertainty. The information is given by the first set of values for the internal node subtracted by the average value of the impurity of the split set.

1.2.4 Best question

The question with the highest value on the information gain will be the one to ask.

1.3 How to build a Tree

1. Choose the best question (according to the information gain), and split the input data into subsets.
2. Terminate: call branches with a unique class labels leaves (no need for further questions).
3. Grow: recursively extend other branches (with subsets bearing mixtures of labels).

1.4 Overfitting

Overfitting is when the learned models are overly specialised for the training samples. This occurs when the data is too noisy, not representative and when the model is too complex. This can be tackled by choosing a simpler model.

1.4.1 Pruning

The idea of reduced error pruning is to consider each node in the tree as a candidate for removal. A node is removed if the resulting pruned tree performs at least as well as the original tree over a separate validation dataset. The pruning is done on validation set, until it is harmful for the model.

1.4.2 Validation set

Due to pruning there is a need for validation dataset that is not training nor test data set. Therefore if there is access to a rich data set, it is separated to three categories.