ANÁLISIS DE DATOS

Ricardo Aler Mur

SELFASSESSMENT OF DECISION TREES WITH QUESTIONS AND ANSWERS

1) What does the entropy measure?

Answer: The amount of information (measured in bits) that needs to be used to encode optimally the dataset. For example, a problem with 10 classes, where one class is very frequent and the other is rare, you could assign shorter codes (in bits) to the most frequent class, and longer codes to the less frequent. If the ten classes have a similar frequency, it would be necessary to encode all with the same number of bits (log2 (10)).

2) What is more likely to overfit, a tree with with many nodes or a tree with few nodes?

Answer: the one with many nodes, because it provides more degrees of freedom to the model. In the specific case of decision trees, we know that near the leaves, nodes are built with very little data, which may memorize the data and not generalize well. That is the reason why pruning techniques are used. Conversely, a tree with few nodes can underfit.

3) Why the entropy (or gain information or information gain) does not work well when an attribute has many values?

Answer: Consider an extreme case with an attribute that has as many values as data instances. In that case, you could put an instance on each leaf of the tree and the data would be perfectly classified (with a zero entropy). But obviously, this is an arbitrary classification that will not generalize beyond the training data.

4) What is a simple way to learn rules from trees?

Answer: You can create a rule for each path from the root to each leaf.

5) What types of trees for regression do exist?

Answer: Regression Trees where the leaves contain the means of the data coming into these leaves and tree models, where each leave contains a linear regression model.

6) How is it possible to build regression models with categorical variables?

Answer: Regression models require numerical variables and categorical variables are not. But from each categorical variable with n values, we can create n binary variables with value 1 if the variable takes the value and 0 otherwise. These variables can be used in regression models. Those binary variables are called dummy variables.