
**GUIDE TO NEAREST NEIGHBOUR AND TREE-BASED MODELS (UNITS 2 AND 3)**

**COURSE: MACHINE LEARNING I**

**MASTER IN BIG DATA ANALYTICS**

**Ricardo Aler Mur**

- The main goal of this lecture is threefold:
    - To describe the basic machine learning pipeline (the sequence of processes that are typically followed in Machine Learning).
    - To explain some basic concepts: instances, attributes, and instance space.
    - To describe two of the basic Machine Learning algorithms: K-nearest neighbours and decision trees.

**About the KNN algorithm:**

- KNN is explained both for classification and regression. KNN is an algorithm that does not build an explicit model from the data. Rather, it stores the training dataset, and classes are computed at testing time. In order to do so, distances are computed from the test instance to all training instances. The class of the closest instance is the answer of the KNN model. Larger values than 1 can be used for K. In that case, the majority class among the K closest training instances is the answer of the algorithm. For regression problems, a simple and common strategy is to compute the average of the outputs of the K closest training instances.

- Some of the advantages of KNN are explained (the main one being that it is not necessary to build an explicit model), and the inconveniences (slowness in classificating new instances, the problem with irrelevant attributes, the curse of dimensionality, and the problems with noise)
- It is shown what the K hyper-parameter means, its influence on classification, and how to select it. In particular, K>1 can be useful for noisy problems, as a way of averaging or smoothing out the noise.

**About the decision tree algorithm:**

- The algorithm that builds decision trees is explained via an example.
- It is shown that the algorithm is recursive and basically amounts to choosing the best attribute at each node, by minimizing a measure called entropy, although other measures such as gini could also be used. Given that attributes are evaluated according to the quality of the partitions they generate, any measure that is able to properly rank data partitions could also, in principle, be used for evaluating attributes.
- It is shown how to extend the basic algorithm to regression tasks, via model trees and regression trees. In this case, the measure to minimize is variance, instead of entropy. Model trees contain linear models in the leaves, while regression trees are simpler, as they contain constants in the leaves. The linear models are constructed from the instances that reach that particular leaf, while the constants are the average of the instances that reach the leaf.

**Associated material**:

In addition to the lecture slides and some exercises (check the course guide), some labs and one assignment are proposed. The labs introduce Scikit-learn, the Machine Learning library that will be used in this course. More specifically, the labs show how to train and evaluate decision trees. The training process includes the hyper-parameter tuning step. The assignment requires the student to apply this knowledge to the other basic model studied in this lesson: KNN.