Universidade do Minho
Departamento de Informática
Mestrado [integrado] em Engenharia Informática

Dados e Aprendizagem Automática
1°/4° Ano, 1° Semestre
Ano letivo 2024/2025

Practical Exercise no. 2

Theme

Quality Metrics and Model Validation

Exercise

In the telecommunications sector, *churn* is a measure of the number of customers who are leaving an operator. Customers may be leaving because they have found lower prices in the competition or because they are unhappy with the service provided, among other reasons. Therefore, for a telecom's operator, as in any other business area, it is imperative to have models capable of predicting the possibility of customer's *churn*, i.e., the possibility of a customer leaving. This will allow the operator to try to hold on to the customer before they choose to leave, by offering better services or more attractive prices.

Tasks

Download two datasets containing data from a telecommunications operator. The first (https://goo.gl/BSUhZ3) contains call data from a customer while the second contains contract data (https://goo.gl/YZLDPf).

A value of *churn = 0* means that the customer has remained with the operator; *churn = 1* represent a customer who has left the operator.

The aim is to:

T1. Load the datasets and merge the two by "*Area Code*" and "*Phone*". Then transform the *Churn* attribute into a nominal attribute;

T2. Using a Decision Tree as a classifier (*sklearn.tree.DecisionTreeClassifier*), evaluate the model's accuracy in prediction the *churn*. Also evaluate the model using the metric *f1_macro*. Use *10-fold cross validation*:

Note: Define X and y. Pay attention to the type of attributes that are part of the X;

T3. Obtain the model's confusion matrices and analyse them critically. What conclusions can be drawn?

T4. Change hyperparameters of the Decision Tree (*criterion* and *max_depth*). What is the variation in model performance underlying these changes?