GUIDE TO EVALUATION AND METHODOLOGY (UNIT 4)

COURSE: MACHINE LEARNING I

MASTER IN BIG DATA ANALYTICS

Ricardo Aler Mur

- The main goals of this lecture are:
    - To describe why it is important to evaluate models, by explaining the concept of generalization and overfitting, both in classification and regression.
    - Then, several methods of evaluation are explained: train / test (holdout partition), repeated train/test, and crossvalidation, the latter one being the recommended method.
    - Crossvalidation works by splitting the dataset into K independent partitions. Then, for each partition P, a model is trained with all partitions but P, and tested with P. The final evaluation is the average for all test partitions.
    - Train / test can suffer from biases in the partitions, especially if data is scarce. Repeated train / test can suffer from overlaps in the test partitions. Crossvalidation is recommended because there is no overlap in the test partitions and each instance is used in two possible ways: as a training instance (in some partitions) and as a test instance (in other partitions).
    - There are many evaluation measures, both for classification and regression. Two of them are explained in more detail: success rate (for classification) and root mean squared error RMSE (for regression).
    - It is shown that RMSE depends on the scale of the output variable. Relative measures are thus introduced, where RMSE is normalized by dividing by the

error of the mean. Relative errors range from 0 to 1, where 1 means that our model has the same prediction error as the mean (and therefore, it is a trivial model).

- Finally, the process of hyper-parameter tuning is introduced.
- In the initial slides, it was shown how the complexity of the model and overfitting are related (the more complex the model, the more likely it is overfitting, although there is always the possibility of underfitting)
- It is explained that all machine learning algorithms have some hyper-parameter that determine the performance of the model, and specifically, its complexity. For instance, decision trees have the maximum depth and the minimum number of instances at the leaves. KNN have K, which is the number of neighbours to be considered.
- Those hyper-parameters need to be adjusted, or tuned. The process of grid-search, by which all possible combinations of parameters are tested, is explained.

Associated material:

- In addition to the lecture slides and some exercises (check the course guide), a new assignment with Scikit-learn is proposed. Now that both the training and evaluation of models have been covered, a more challenging assignment is required to be completed. This assignment is centered on digit classification (SEMEION dataset).