Evaluation

1. A classifier was trained and tuned on a training set until the error was 0. Then a new dataset, a test set, was used to evaluate the performance of the learned model, giving a significant error. How can the error on the test set be high, if the error on the training set was 0? What measures can be taken to prevent this phenomenon and improve generalization to unseen data?

   The phenomenon described is known as overfitting, and occurs when a model adjusts to the training data too closely, losing the ability to generalize. To the extreme, the supervised learning method can memorize the training set, and not be able to classify (or, in regression, predict the output of) new data points. It arises when the complexity of the model (e.g., number of parameters) exceeds the regularity of the underlining process generating the data. One way to prevent or alleviate overfitting is to use a simpler model. Given a model, it can be alleviated by measuring the error, during learning, on a different data set (a *validation* set). When the error on the new data begins to increase, it indicates that the model is starting to overfit the training data.

2. What is the role of the validation set as opposed to the test set?

   The validation set is used to test the error on a set different from the training set, in order to detect overfitting and do model selection. It is different from the test set, because by stopping training when the error is the lowest on the validation set, the model might be overfitting the validation set itself, and another data set is required to test the generalisation of the model.

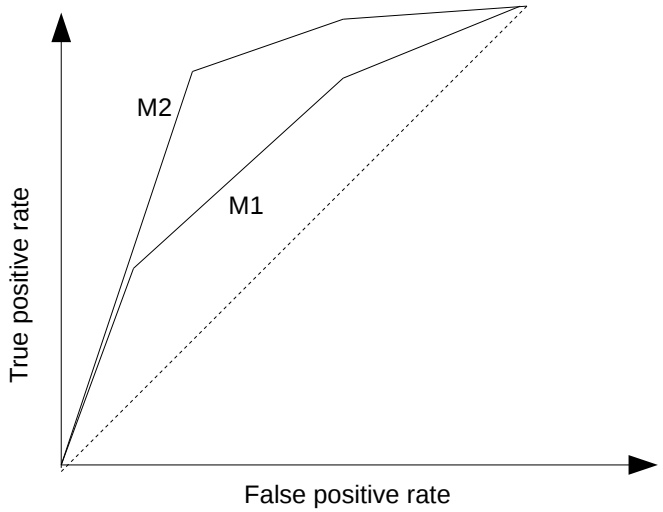
3. What is cross-validation? What is it used for?

   Cross-validation is a technique used to evaluate a model, by splitting the dataset into several batches, and then use one for testing and all the others for training. This is repeated for every batch, and the results are averaged across all batches at the end.

4. A model can distinguish between points of 5 classes. What is the most appropriate single representation of its accuracy?

   A confusion matrix is the most appropriate and compact representation of the accuracy for a multi-class (that is, non binary) classifier.

5. Given the two ROC curves below, which model is to be preferred? Justify your answer:



The model M2 has a higher ratio of true positives and lower ratio of false positives with respect to model M1, and its area under the curve in the ROC graph is higher. Therefore, it is the one to prefer.

6. A dataset has 500 points of class A, and 25 points of class B, which metric is most appropriate to represent its accuracy? Justify your answer.

   Since the dataset is strongly unbalanced, the best metric in this case is Matthew's Correlation Coefficient, which takes the total number of elements per class into account.