INF264 - Homework 5

Natacha Galmiche

1 Boosting

Boosting classifiers are a type of ensemble classifiers whose main objective is to reduce the bias by training a sequence of weak learners where subsequent models compensate the errors made by the earlier ones. This approach is motivated by a theorem stating that weak learners (that is to say learners that are barely better than a random classifier) can in theory be 'boosted' to perform as well as 'strong learner' (that is to say, classifiers whose accuracy is almost 1).

Typically a weak learner could be a decision tree with a max depth set to 1 or 2.

The dataset we will use in this exercise is the Sonar dataset. This dataset describes sonar chirp returns bouncing off different surfaces. The 60 input variables are the strength of the returns at different angles. It is a binary classification problem with 208 observations that requires a model to differentiate rocks (labeled R) from metal (labeled M) cylinders.

1. Load the sonar dataset and store the features in a matrix X and labels in a vector y. Remember that in the adaboost algorithm, labels M and R should be converted into 1 and -1 labels. Split the dataset for a cross validation pipeline
2. Boosting training function. Implement the adaboost training algorithm as explained in the lecture 10: *ensemble*, slide 21.
3. Boosting predict function. Implement the adaboost predict algorithm as explained in the lecture 10: *ensemble*, slide 21.
4. Boosting pipeline, using cross validation on the sonar dataset using decision trees.:
   (a) Train different boosting classifiers based on different numbers `n_clfs` of Decision Trees classifiers whose max depth is set to 1 for different number of classifiers `n_clfs`. Select using cross validation.
   (b) Evaluate the best model.
   (c) What can we say about the variance and bias of a simple decision tree compared to the selected boosted classifier?
   (d) What are the disadvantages of a boosting classifier then?
5. Evaluate boosting classifiers based on different numbers of Logistic regression models.