MSHD Projects

Below you will find the descriptions of the projects for this year. Each group must choose one project, different from the ones chosen by the others, and prepare a presentation either on the 14th or on the 16th of May during the hours usually devoted to the course lectures and seminars. Each group has also the opportunity to propose another topic for their final work that needs first to be presented to the professor and then accepted as an admissible project.

1 Criteria based on MSE: the FIC

In this project you should work with the Malnutrition in Zambia dataset. Consider the *breastfeeding duration* as the most important variable, develop an estimator for the Focused Information Criterion (FIC), build a model that optimizes the FIC and compare the resulting model with the one obtained with other methods (at least two) presented in the course with the Mean Squared Prediction Error.

2 Selection by Hypothesis Testing

In this project you should program the *StrongStop* rule and the *Benjamini & Hochberg* rule in the framework of the *selectiveInference* package. Check the performance of each rule, together with the already present *ForwardStop*, as α is varying and plot the power of the test reached by each rule with a simulation experiment. Finally apply each rule on a real high dimensional dataset and compare the results with a given criterion seen in class (e.g. prediction error).

3 Streamwise regression: SIS

In this project you should choose three existing model selection criteria to select the γ parameter of the *Sure Independece Screening* algorithm. Then you should evaluate the performance of each criterion both by means of simulations and on a real high dimensional dataset.

4 Pruning CART

In this project you should first adapt several (at least two) model selection criteria seen in the course to prune a CART object. Then you should compare the results with the standard method used in the *rpart* package, which is based on cost complexity, in light of the prediction error. The comparison should be done both in a simulated setting and with a real high dimensional dataset.

5 Penalized Regression: LASSO

In this project you should test different methods (at least two) to select the parameter λ and compare them with the default one of the *glmnet* package which is 10-fold CV. In order to do so, you may employ the ordering already given by the package *glmnet* and apply a given criterion to stop the search. Then you should evaluate the performance of each method both by means of simulations and on a real high dimensional dataset.

6 Analysis of a high dimensional dataset

In this project, you should analyze a high dimensional dataset (e.g. $p > 10000$) that you will find, using several methods (at least three) seen during the course or indeed others that you judge suitable for the task. Perform a clear comparison on the results of the chosen methods in light of different criteria: either the ones already used during the course or others of your choice that nevertheless need to be correctly justified.