Case 1

02582 Computational Data Analysis

s182324 & s181740

March 19, 2019

Introduction/Data description

For this assignment, we were provided with a dataset of 1100 rows and 100 columns. The first column contains the first 100 true values (Y) while the rest of the column is empty and must be fulfilled with our best prediction. The remaining columns are the features(X). The features are numerical values except for the last 5 which are categorical values. For all the above, we had to decide how we should approach the missing values in the dataset, also choose how we are were going to choose the features needed for training and which models we should implement in order to make our predictions.

Model and method

The methods we decided to implement were the Ridge regression, LARS and KNN regression.The reason for those choices is that most of the regression models suffer when they have to deal with multi-dimensional data, so we decided to go through those models and see which one gives us the best prediction with the least overfitting. Also, for the categorical features, we encoded them using "One hot encoding" method and perform "binarization" and then include them as a new feature to our model. As for the feature selection we decided to use the "Recursive Feature Elimination" (RFE) method because it uses the model's accuracy to identify which attributes (and combination of attributes) contribute the most to predicting the target attribute.

Missing values

At first, we decided to remove the rows with missing values, but only 35 rows were complete. Then, we decided to fill the missing values. We filled the missing values in 2 manner:

1. For the numerical values we chose the mean of the column and for the categorical values we put the mode.

2. We used the k-nearest neighbors(knn) method and by cross validation we found the suitable k which was 8. For the categorical we put the mode.

For our final model we used knn, as the accuracy was higher when we test it.

Factor handling and Model validation

We used 2-layer cross-validation for factor handling and model validation. In the first layer, we find the best factor value for each model and in the second layer, we find the accuracy of the models with the factor which was founded for them. After finding the best model with the best factor, we used the whole 100 data with y for training our final model.

Model selection

Since the number of features was large and even larger than the number of observations (before feature selection), we decided to use models which can handle this situation. Thus, we chose Ridge Regression, Lasso, Lars and KNN regression.

As we had lots of features, Random forest would not be a good choice, Since, the probability that it chooses not related feature is very high, so, we did not implement it.

Results

Ridge Regression had the best performance with around 84% accuracy, after finding this fact with 2 layer cross-validation, we fit the whole 100 data points -with y values- for training our final model with Ridge Regression. Then we predict the y value of 1000 data point by this model.

For estimating the prediction error, we used One-leave-out cross-validation which is a k fold validation when the number of k is equal to data points. Since we only had 100 data point it was not expensive to compute and had less bias in compare to k fold cross-validation. In summary, we fit our Ridge Regression model with our best alpha (which was 1) for 99 data points and test the model for the remained point and calculate the MSE. then, we repeat the process for all the data points and find the mean of the MSE.