Statistical Methods 2

Portfolio 5 – Ridge regression, LASSO and smoothing

Complete the following tasks and submit your work on Blackboard by 4pm on Friday 03/03/2023

Task 1 (60 marks)

Choose a dataset on which you can fit a linear regression model of the form $Y_i^0 = \alpha + \beta^\top x_i^0 + \epsilon_i$ and split the dataset into a training and a test set.

In a first step, use the training set to estimate the model parameters with ridge regression and with LASSO regression. In particular,

- Plot the LASSO path.
- For the two regression methods choose the penalty parameter λ using cross-validation, making clear which cross-validation method you use.
- Compare the estimate of β obtained with ridge and LASSO regression.

Remark: Do not forget to normalize the variables if needed. And if you normalize the variables do not forget to correct the estimate of β that you compute (as explained in Chapter 6).

In a second step, use the test set to compare the out-of-sample prediction error of the model fitted with ridge regression with that of the model fitted with LASSO regression.

For this task you can for instance use the Communities and Crime dataset[1] and perform LASSO regression using the `R` package `glmnet`.

Optional: In addition to ridge and LASSO regression, fit your model on the training set using principal components regression. Choose the number $q < n$ of principals components to use using cross-validation, and compare the prediction error of the resulting fitted model on the test with that obtained with ridge and LASSO regression.

Task 2 (40 marks)

Choose a dataset $\{(y_i^0, x_i^0)\}_{i=1}^n$ where $x_i^0 \in \mathbb{R}$ and where the relationship between the y_i^0's and the x_i^0's is non-linear. Consider the model $Y_i^0 = f(x_i^0) + \epsilon_i$ where $f \in \mathcal{C}^2(\mathbb{R})$ and estimate f using the smoothing approach discussed in Chapter 8 and using generalized cross-validation to choose the penalty parameter λ. Make a plot showing both the estimated function and the observations, and comment your results.

For this task you can for instance use the Bone Mineral Density dataset[2] and try to reproduce Figure 5.6 of *The Elements of Statistical Learning*. You can also use a simulated dataset if you want. Note that smoothing methods are for instance implemented in the `R` package `mgcv`.

[1] Available from the UCI Machine Learning repository or from the `R` package `mogavs`.

[2] available at `https://hastie.su.domains/ElemStatLearn/`