Problem sheet 3: Model reduction

Analisi dei Dati – CdS Matematica – 2022/23

When suitable, please provide summmarising and explanatory pictures.

Exercise 3.1. Consider the Yale face dataset. Apply a dimension reduction with PCA. Verify that it is possible to recognise individuals on the principal plane.

Exercise 3.2. Experiment with *subset selection* by evaluating all possible regressive models on the dataset `Ames` (with output of your choice). Then compare with the results of forward and backward selection. Produce a picture as in Fig. 6.1 of the book *An introduction to statistical learning*, highlighting also the "paths" of forward and backward selection.

Exercise 3.3. Analyse training and test error in dependence of the parameter λ with ridge regression and lasso on the dataset `Ames`. Produce a picture of coefficients values with respect to λ (estimate the test error with CV).

Exercise 3.4. Same as Exercise 3.3, but for a dataset generated ad-hoc, with non-trivial correlations among input factors and with the output. Obtain statistical estimates.

Exercise 3.5. Generate a dataset with a large number of factors ($\gg 100$), but with only a small number of them ($\sim 10 - 20$), randomly positioned, correlated with the output. Analyse the behaviour of λ, of the coefficients and of the training and test error in lasso and ridge regression.

Exercise 3.6. Implement principal components regression and evaluate the behaviour of training and test error with respect to the number of principal components, by analysing in particular the explained variance, on the dataset `Ames` and on a ad-hoc generated dataset with a large number of features ($\gg 100$), but with much smaller effective dimension.

Exercise 3.7. Consider the MNIST dataset. Use a classification algorithm of your choice and compare accuracy based on the original images and on the images obtained by PCA dimension reduction.