TMA4268 Statistical Learning V2020

Module 6: Recommended exercises

Thiago G. Martins, Department of Mathematical Sciences, NTNU

Spring 2020

Recommended exercise 1

1. Show that the least square estimator of a standard linear model is given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

2. Show that the maximum likelihood estimator is equal to the least square estimator for the standard linear model.

Recommended exercise 2

Write R code to create a similar representation of the Credit data figure 1 shown below.

Recommended exercise 3

1. For the Credit Dataset, pick the best model using Best Subset Selection according to C_p, BIC and Adjusted R^2

- Hint: Use the `regsubsets()` of the `leaps` library, similar to what was done in Lab 1 of the book.

2. For the Credit Dataset, pick the best model using Best Subset Selection according to a 10-fold CV

- Hint: Use the output obtained in the previous step and build your own CV function to pick the best model.

3. Compare the result obtained in Step 1 and Step 2.

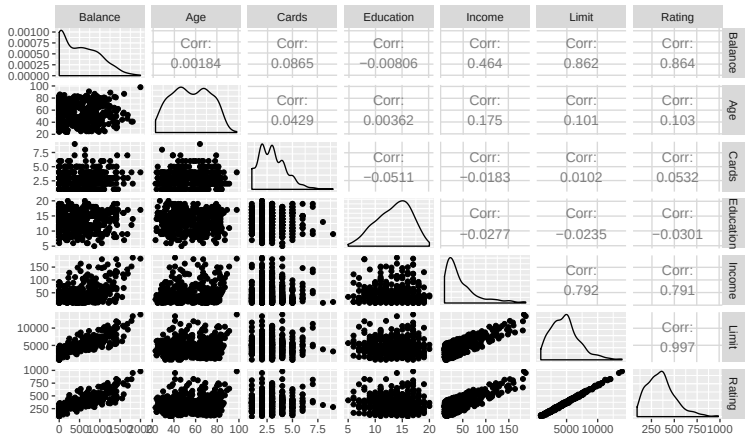

Figure 1: Credit data

Recommended exercise 4

1. Select the best model for the Credit Data using Forward, Backward and Hybrid (sequential replacement) Stepwise Selection.

- Hint: Use the `regsubsets()` of the `leaps` library

2. Compare with the results obtained with Best Subset Selection.

Recommended exercise 5

1. Apply Ridge regression to the Credit Dataset.

- Hint: Use the `glmnet()` function from the `glmnet` package, with parameter `alpha` $= 0$.

2. Compare the results with the standard linear regression.

Recommended exercise 6

1. Apply Lasso regression to the Credit Dataset.

- Hint: Use the `glmnet()` function with parameter `alpha` $= 1$.

2. Compare the results with the standard linear regression and the Ridge regression.

Recommended exercise 7

How many principal components should we use for the Credit Dataset? Justify?

- Hint: Use the `prcomp()` function (`stats` library) to calculate the PCs and use `plot(yourModel, type = "l")` to plot the the proportion of variance explained by each PC.

Recommended exercise 8

Apply PCR on the Credit dataset and compare the results with the previous methods used in this module.

- Hint: Use the `pcr()` function from the `pls` package, and use the argument `validation="CV"`.
- Hint: Use the function `validationplot(yourModel,val.type="MSEP")` to plot the mean squared error of prediction for number of components.

Recommended exercise 9

Apply PLS on the Credit dataset and compare the results with the previous methods used in this module.

- Hint: Use the `plsr()` function from the `pls` package, and use the argument `validation="CV"`.
- Hint: Use the function `validationplot(yourModel,val.type="MSEP")` to plot the mean squared error of prediction for number of components.