EPFL, Autumn 2024
Yoav Zemel
Guillaume Obozinski

Ho Yun
Shivang Sachar

Statistical Machine Learning

Exercise sheet 4

Practical exercise

Exercise 4.1 (Variable selection and regularization) You will perform ridge regression and lasso on the bodyfat dataset.

If you work on R, the data are in the library "mfp" and are then loaded using the command `data(bodyfat)`; be careful, you must remove the columns 1,2 and 4 when fitting your models. If you work on another software, you can find the bodyfat data in a csv file on Moodle. The objective is to predict `siri` using the other variables.

(a) Read §§6.6 from [ISL] to familiarize yourself with the packages and the functions in R you will need for the purpose of this exercise if you haven't done so already.

(b) Perform ridge and lasso regression for some sequences of λ. By using `plot` function from the `glmnet` library, plot a graph of the coefficient values over λ.

(c) We will try to find optimal values of λ for ridge and lasso, by evaluating the errors of the models on an independent validation set. Unfortunately, we don't have additional observations. Therefore, we shall split the original dataset into two parts: 152 observations used for the training set and 100 observations used for the validation set. This split should be done at random, but it is useful to fix the random seed of R first using for example `set.seed(1)`. You can use the command `sample(1:252,152)` to get indices of 152 observations to be used in the training set.

For ridge and lasso, identify a good value of λ that minimizes the validation error. You can plot a graph of the validation error over a sequence of values of λ.

(d) What happens when you fix a different random seed instead? i.e., repeat the procedure by first calling `set.seed(5)`, this will change the training and validation sets. Does your result from (c) change? Discuss.

(e) Write your own code to perform leave-one-out cross-validation, plot a graph of the validation error over a sequence of values of λ and use it to estimate the optimal value $\widehat{\lambda}_{\mathrm{CV}}$ of the regularization parameter λ.

(f) Write your own code to perform K-fold cross-validation, plot a graph of the validation error over a sequence of values of λ and use it to estimate the optimal value $\widehat{\lambda}_{\mathrm{CV}}$ of the regularization parameter λ.