Introduction to Statistical Machine Learning for Bioinformaticians and Medical Informaticians, SS 2022
Dr. Ralf Eggeling
Dana Petracek
Methods in Medical Informatics
University of Tuebingen

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Assignment 6

Deadline: Monday, June 20, 2:00 p.m.

This problem set is worth 35 points. You can submit in groups of two people or alone. Submit your solutions by uploading them to moodle (none of the other students can see the files you upload). Name the files

`Assignment_6_[lastname].pdf` and `Assignment_6_[lastname].R`
for individual submissions and
`Assignment_6_[lastname1]_[lastname2].pdf` and `Assignment_6_[lastname1]_[lastname2].R`
for team submissions.

In the latter case, include names of both students at the top of both files. Both students must upload the identical files to moodle in time.

Moodle allows to upload **drafts**, which can then be further edited until the deadline. Use this for submitting finished tasks in case you might run out of time. Once you formally submit, the upload cannot be edited any longer and is ready for grading. If you never formally submit, the uploaded draft at the submission deadline will be graded.

Task 1 (10 Points)

Derive the variance formula:

$$\mathrm{Var}\left(\frac{1}{k}\sum_{i=1}^{k} X_i\right) = \rho\sigma^2 + \frac{1-\rho}{k}\sigma^2$$

where X_i , $i = 1,\ldots,k$, are identically distributed random variables with positive pairwise correlation ρ and $\mathrm{Var}(X_i) = \sigma^2$ for $i = 1,\ldots,k$.

Task 2 (25 Points)

Go through **5.3 Lab: Cross-Validation and the Bootstrap** (ISLR p.212–219), **6.5 Lab: Linear Models and Regularization** (ISLR p. 267–279). The objective of this programming exercise is to predict the logarithm of the prostate specific antigen (PSA) level based on the other predictors. You find the dataset *prostate.txt* on moodle. Note: for reproducibility, set the seed of the random number generator to one by invoking *set.seed(1)* everytime you call a function that involves randomness.

(a) Read and normalize the data: use *read.table()* to load the data; column 9 is the output lpsa for the regression and column 10 determines whether this data entry belongs to the training set. Column 1 is just an index and should not be used for prediction. Normalize each input feature to a mean of 0 and a variance of 1. Split up the data set into training and test set respectively. For this subtask, you may include the code (with proper formatting) into the pdf report. Useful functions: mean(), sd(), and the MASS library. (**2 Points**)

(b) Use LOOCV, 5- and 10-fold cross-validation on the training data set to estimate the test error of using linear regression to predict lpsa from all other features. Use the full training data set to train a linear regression model and compute the test error. Compare your estimates obtained from cross-validation to the error obtained from the test set and argue about your findings. Which of the methods is (theoretically) the fastest? (**4 Points**)

(c) Use the training set to fit ridge regression models and generate a plot showing the values of the coefficients in relation to the parameter λ (cf. Figure 6.4, p. 238, ISLR). What can you observe? (**3 Points**)

(d) Perform 10-fold cross-validation on the training set to determine the optimal value for λ for the ridge regression model. Report train and test error measured in MSE for this λ. (**3 Points**)

(e) Use the training set to fit lasso models and generate a plot showing the values of the coefficients in relation to the parameter λ (cf. Figure 6.6, p. 242, ISLR). What can you observe in comparison to the plot generated in (c)? (**3 Points**)

(f) Perform 10-fold cross-validation on the training set to determine the optimal value for λ in the lasso. Report train and test error measured in MSE for this λ. How many and which features are used? Compare this to the coefficients determined for ridge regression in (d). (**3 Points**)

(g) Compare the models generated in (d) and (f) to the model generated in (a). Which model would you choose? What alternative model could have been used? (**2 Points**)

Note: 5 of the 25 points are awarded for presentation style, rewarding **clarity**. The maximum number of style points you can get depends to the number of solved tasks, but not (necessarily) on the correctness of the solution.