# PROJECT WORK IN SF2935: MODERN METHODS OF STATISTICAL LEARNING (2019)
# PART 2

## General information

The course has a project part that counts for 3 ECTS (out of the full 7.5 for the course). This part is an essential learning activity in the course and is mandatory in the sense that you cannot obtain a passing grade for the full course without completing the project part.

Regarding programming, although the book focuses on R, you can use other languages if you prefer. When writing the report, you should not include all your code. Instead, focus on describing the methods/techniques used and any active choices you make (the use of a regularisation technique, choice of learning method etc.). Moreover, be sure to describe what any existing libraries/functions that you are using do - check the documentation.

Guidelines for group organisation and grading are the same as for Project 1.

## Project 2

The project is divided into two parts, A and B. As in the first project, some formulations are vague and you should use your own judgement in deciding what to do and include in your report.

To complete the assignment you are free to use any software and corresponding packages of your own choice - you are *not* required to use R. Please state explicitly what you are using and what the specific function does.

You should *not* describe the theory underlying the tools you chose for your analysis in full detail, unless you make use of techniques not covered in the course. For the latter, a brief review of the technique and the most relevant theoretical results should be included in the report. Make sure to explicitly state what different variables, results etc. refer to.

### Part A

This part is partly an extension of the first project and deals with support vector machines (SVMs) for synthetic data and the data set *Auto* in the R library ISLR[1].

a) Follow the instructions for Part A in Project 1 and repeat parts (c) and (d) but with SMV as your method for classification in the *Auto* data set. Try different kernels, parameters and costs - which combination yields the best result?

b) Create a synthetic data set of 1000 Gaussian random variables in $\mathbb{R}^2$ - mean vector 0 and independent components (also independent between samples)[2]. Shift 500 of the data points some distance from the origin, i.e., pick a distance and shift the points radially. Assign these points to calls

[1] If you prefer to work in another programming language you can simply export the relevant parts of the data set

[2] You could also try this in $\mathbb{R}^3$ instead of $\mathbb{R}^2$.

1 and the remaining, unmoved, points to class $-1$. Select 100 points from each class and create a test set. Train an SVM to classify the points, using different kernels and parameters to find a model that performs reasonably well on the test set. How is this affected by the choice of distance $r$ you move the points in class 1 and the variance of your Gaussian random variables? Try to create a feature map and use a support vector classifier, with that feature map, to classify the points. What is the result? Can you explain the outcome (intuitively)? Lastly, change the ratio between points in class 1 and $-1$: rather than 50/50, let $100 * k$ % belong to class 1 and remaining to class $-1$, with $k > 0.5$. Keep the same ratio in the test set (i.e., $k * 200$ are from class 1, remaining from class $-1$). How does this affect the results of SVM when $k$ moves in $(0.5, 1)$?

**Part B**

In this part you will explore tree-based methods.

a) Follow the instructions for Part A in Project 1 and (a) above: repeat parts (c) and (d) from Part A in Project 1 but with classification trees as your method for classification in the *Auto* data set. Report your findings using both pruned and unpruned trees.

b) Similar to (a), use a regression tree to predict the *mpg* variable in the *Auto* data set. Repeat the steps of Part A in Project 1 and (a) above only now *do not* remove the *mpg* variable but rather use this as the response. Consider both pruned and unpruned trees. What are your results? Can you compare them to part (a), which method is most easily interpreted when considering whether a car will have a high mpg or not?

## Schedule and deadlines - Part 2

- October 30: Deadline for submission of final report (23:59).

The final report should be submitted on Canvas as a PDF, titled:

**SF2935Project2-Name1-Name2-Name3.pdf**.

Good luck!