**2020-03-04**

**Examination in 806640S Statistical Learning**

**Grades:**

5=29-30 points

4=26-28 points

3=23-25 points

2=20-22 points

1=18-19 points

F=0-17 points

Provide R-code and motivations for all answers.

***Assignment 1. Regression***

The dataset **lakesurvey.csv** contains 16 chemical and physical measurements from 2782 lakes in Scandinavia.

**a. Variable selection**

Import the data into R. Set up a model where pH is the response and the rest of the variables are predictors. Perform Best subset selection and present the best models (selected variables and their regression coefficients) based on both BIC and Mallows Cp. The next task is to perform variable selection with the LASSO using the glmnet package. Tune the regularization parameter with 500-fold CV. Present a plot that shows that the minimum MSE has been obtained. Also present the selected variables and their regression coefficients. Compare and discuss the results from both analyses. (**5p)**

**b. Prediction**

In the next task you will investigate the correlations between the predictors. As you can see there are some highly correlated variables. Extract the CV error from the LASSO in a. Find a method that is known to produce low prediction error on correlated data. The method should give at least as good prediction error as the LASSO. Motivate your choice and present details on how you have tuned the method. (**5p)**

***Assignment 2. Nonlinear regression***

File **Concrete\_Data\_2.xls** contains three variables. The Concrete Strength which is a possibly nonlinear function of Age and amount of Cement.

**a. GAM**

Import the data to R and fit two GAM models (with both Spline and Loess smoothers) using Concrete Strength as response variable, and Age and Cement as predictors. Use leave-one-out-cross-validation as model selection criterion to determine which model fits best. Explain the results and provide plots of the regression functions for the best model. **(5p)**

**b. Thin-plate spline**

Fit a thin-plate spline model to the Concrete data using the same response and predictors as in a. Provide model results and a contour fit plot of the best fitting model. What is the methodological difference compared to a. and how did you tune the model? **(5p)**

***Assignment 3. Classification***

The file ‘**olives.csv**’ contains data about the content of fatty acids in samples of olive oil from three regions and nine different areas in Italy.

**a. Linear Discriminant Analysis**

Load the data and check for missing values and correlated predictors. Delete the Sample and Area variables. Scale the numeric variables. Divide the data randomly into 70% train and 30% test using a seed of 3. Make a linear discriminant analysis (LDA) using the lda() function. Perform predictions based on the test data and calculate the accuracy. Present a plot of the classification result that clearly shows how the test predictions relate to the training predictions including decision boundaries. (**5p)**

**b. Random Forest**

Use the same data as in a. and implement a Random Forest analysis via the caret package. Use LOOCV for tuning of the mtry parameter with values c(3,5,7) while you fix the number of trees to 500 and node size to 3. Present a Variable Importance plot and the test accuracy that should be calculated on the test data. Compare with a. and discuss your results. (**5p)**