Final project STK4030/9030 - Statistical learning: Advanced regression and classification Fall 2015

Available Monday November 2nd. Handed in: Monday November 23rd at 13.00

November 4, 2015

This is the problem set for the project part of the finals in STK4030 fall 2015. The reports shall be individually written.

Three copies marked with your candidate number shall be placed in Kristoffer Hellton's post box at room B700 at the seventh floor in Niels Henrik Abels hus. Handwritten reports are acceptable. Enclose the parts of the computer outputs which are necessary for answering the questions. The other parts can be collected in appendices. When you refer to material in these, be careful to indicate explicitly where.

Importantly, each student needs to submit a special extra page with her or his report. This page is the self-declaration form, properly signed, and with the appropriate course form STK4030 (master level) or STK 9030 (PhD level) clearly marked; it is available at the webpage as "Exam Project, declaration form".

All data sets and R scripts are available at the course webpage.

Kristoffer H. Hellton

1 Problem 1 - *Vino verde* wine quality

A Portugese wine seller wants your help in constructing a predictor for wine taste preferences. He has collected easily available physiochemical data and a quality assessment (between 0 and 10) by human wine experts of 600 *Vino verde* white wine samples. The data set `wine.RData` contains the response `quality` and 11 covariates[1] (`fixed acidity`, `volatile acidity`, `citric acid`, `residual sugar`, `chlorides`, `free sulfur dioxide`, `total sulfur dioxide`, `density`, `pH`, `sulphates`, `alcohol`).

The data set is divided into a training and a test set of 300 samples each, and the last column, `test`, indicates whether a sample is in the test set. Before the analysis, center the mean and scale the variance of the covariates in the whole data set using all 600 samples.

[1] Cortez et al. (2009) *Modeling wine preferences by data mining from physicochemical properties*

(*a*) Estimate a linear regression model with `quality` as the response with all 11 predictors by (ordinary) least squares (OLS), and predict the quality on the 300 test samples. Report the estimated coefficients and the test error. Which covariates have the strongest association with the wine quality?

(*b*) Estimate the same linear regression model by ridge regression, where the optimal tuning parameter λ is chosen by 10-fold cross-validation, and predict the quality on the 300 test samples.

Plot the cross-validated mean squared error against the logarithm with base 10 of the λ values. What is the optimal value of λ?

Report the estimated coefficients, and the mean squared error on the test data.

(*c*) Estimate the same linear regression model by lasso where the optimal tuning parameter λ is chosen by 10-fold cross-validation, and predict the quality on the 300 test samples.

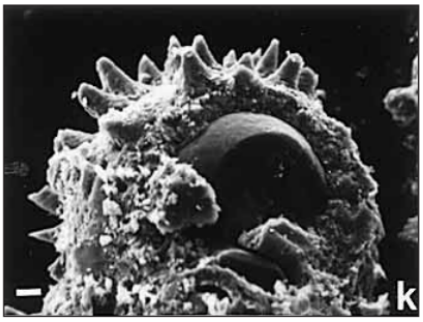
Plot the cross-validated mean squared error against the logarithm with base 10 of the λ values. What is the optimal value of λ?

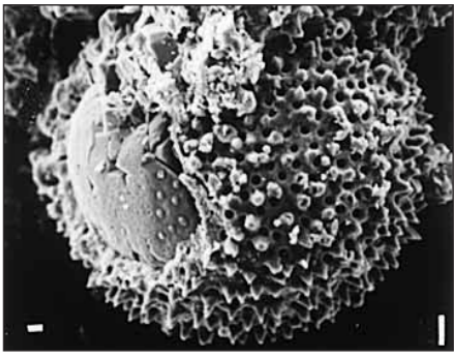
Report the estimated coefficients, and the the mean squared error on the test data.

(*d*) Use the test data set for validation and compare the OLS, ridge and lasso in terms of the mean squared error on the test data. Which predictor would you recommend to the wine seller? How will the estimated mean squared error on the test data for your chosen predictor compare to the true test error?

2 Problem 2 - Fossils

Bralower et al. (1997) *Mid-Cretaceous strontium-isotope stratigraphy of deep-sea sections* investigated how fossils of different age have varying ratios of strontium isotopes. The characterization of the strontium levels is important for understanding the production rate of the oceanic crust. The change of the ratio between strontium isotopes over time is highly non-linear.





The data set `fossils.RData` contains the ratio between Strontium-87 and Strontium-86 and the age (in millions of years) for 106 fossil samples. You will estimate a non-linear semi-parametric regression model with `age` as input and `strontium.ratio` as response.

(a) Construct a B-spline basis of order 4 for the input with external boundary knots at the range of `age`, and 40 internal knots located

i) equidistantly, with a distance 0.761 apart

ii) at equidistant quantiles of the distribution of `age`

Plot both these sets of basis functions.

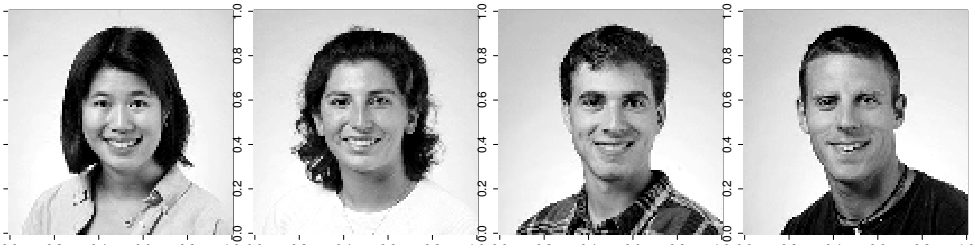
(b) Use the B-splines with quantile-based knots as a spline basis and estimate the smoothing splines regression curve for the relationship between `age` and `strontium.ratio`. Select the smoothing parameter both using leave-one-out cross-validation and the generalized cross-validation (GCV) criterion. Report the estimated smoothing parameters.

Plot the data and both smoothing spline curves. What do you predict the strontium ratio to be in a 113.5 million year old sample?

3 Problem 3 - Face recognition

The data set `faces.RData` (courtesy of Håvard Rue, NTNU) contains portraits or grey level images of $n = 200$ medical students at Stanford University, where $i = 1, \ldots, 100$ are male portraits and $i = 101, \ldots, 200$ are female portraits. Each portrait can be thought of as a matrix of 100×100 pixels and is recorded as a vector of length $p = 100^2 = 10000$, treating each pixel as a variable. Consider all

The function `display.matrix` (found on the course webpage) scales and maps each pixel to a grey value (black being '0' and white being '1') and can be used to display a 10000×1 vector as an image. The data matrix `faces` is therefore given on the form $p \times n$ and the images of the 101st, 102nd, 1st and 2nd column of `faces` are shown as



(a) Estimate the mean of the portraits for each gender, and display them as 100×100 images. Does the mean vary across the image? Which features are represented in the estimated mean portraits?

The portraits are a mix of images taken at the shoulders and at the chin, such that the relative size of the face varies significantly. The variable `shoulder` indicates whether a portrait includes the shoulders or not.

(b) Find the principal components of the whole data set and display the first three eigenvectors (the principal component directions) as 100×100 images. The images of the eigenvector are usually referred to as *eigenfaces* in context of face recognition. What do the first three components represent in terms of facial or image features?

How many principal components are needed to express 80% of the variation in the data set?

(c) Find the first three principal components ($n \times 1$ vectors) of the whole data set. [2] Plot the first and second principal components and the first and third principal components against each other and color the observations according to gender. Then color the plots according to the shoulders being present or not. What do you conclude? Display all four figures.

(d) Recode gender as an indicator variable

$$y_i = \begin{cases} -1, & \text{if } G_i = \text{male}, \\ 1, & \text{if } G_i = \text{female}. \end{cases}$$

Construct a classifier for gender using principal component regression (PCR) with y_i as the response, and classify an observation as female if $\hat{f}(x_i) > 0$. Use leave-one-out cross-validation to select the number of components, $m < 200$, and select the smallest m if the crossvalidation error has several minima. What is the misclassification rate (the error with a 1-0 loss function on the training data)?

(e) Construct a classifier for gender using partial least square (PLS) with y_i as the response, and classify an observation as female if $\hat{f}(x_i) > 0$. Use leave-one-out cross-validation to select the number of PLS directions, $m < 200$, and select the smallest m if the crossvalidation error has several minima. What is the misclassification rate (the error with a 1-0 loss function on the training data) now?

Why does the optimal m for PCR and PLS differ?

Ordinary linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) cannot be used in a high-dimensional setting with $p > n$, as the sample covariance matrix will be singular. You will now explore two ways to avoid this issue; reduce the dimension and regularizing the covariance matrix.

(f) Let the whole data set (10000×200) to be represented by the first five principal components (5×200). Use the reduced data as input for QDA and construct a classifier for gender. What is misclassification rate on the training data? Plot the quadratic decision boundaries in the figures from 3(c).

[2] The first three columns of $XV = UD$ in the notation of the singular value decomposition.

(g) Now consider the gender and the shoulder categories together as four different categories (`maleShoulder`, `maleNoShoulder`, `femaleShoulder` and `femaleNoShoulder`). Use the first five principal components (5×200) to construct a QDA classifier for the four gender-shoulder categories and classify the training data. Merge the classifications for the groups with and without shoulders within each gender, such that you end up with a classification for gender. Compare the result to the classification in (f). Does the misclassification rate decrease when taking into account the shoulders?

(h) Finally, return to the whole data set and use a version of regularized LDA to classify only gender. Motivate how you regularize the sample covariance matrix and argue for your choice of tuning (penalty) parameter. What is the misclassification rate?