Assignment 2

Machine Learning

October 23, 2017

Instructions

- You may use any typesetting software you wish, but I would encourage you to use R-Markdown or LaTeX.
- Provide complete code for each of the relevant questions under separate headings as an appendix to your write-up. For example, for Question 2, provide all code pertaining to this question as a single script in the appendix under the subsection heading 'R-Code: Question 2'. Start each question on a new page.
- You may NOT provide R output interspersed between your answers!. Please typeset relevant elements in the output either in-line, or tabulate results formally. Plots are very useful, but use them sparingly — make sure that a given plot is relevant to the question and pertains to text in your answer. Figures are meant to enrich your analysis, not leave it to the reader to analyse. Provide captions for all figures and tables. Square figures only.
- When you typeset R code use `courier` or an equivalent 'typewriter'-like font.
- All assignments must be accompanied by a signed plagiarism declaration — a template is provided on Vula.
- Hand-in dates will be announced on Vula.

Problem Set

1. Consider a linear model of the form:

$$\begin{aligned} y_i &= 0.8x_i + \epsilon_i; \quad -1 \leq x \leq 1,\ i = 1,..,N \\ \text{where: } \epsilon_i &\sim \text{Normal}(0,1). \end{aligned} \tag{1}$$

   i. Suppose $x \sim \text{Uniform}(-1,1)$. Simulate a dataset of size $N = 30$ and fit the following two linear models to the dataset:

$$\begin{aligned} g_1(x) &= 0.5 + b_1 x \\ g_2(x) &= -0.5 + b_2 x. \end{aligned}$$

   Consequently plot the data, the true underlying model as well as the two fitted models $g_1(x)$ and $g_2(x)$. Comment of the expected relative performances of $g_1(x)$ and $g_2(x)$.

   Hint: To fit a linear model with fixed intercept a to dataset of size n use: lm(y ~ 0+x,offset=rep(a,n)).

   ii. Using model (1), simulate 10,000 datasets of size $N = 30$. Divide each dataset into a validation set of size i and a training set of size $30-i$ where $i = 5, 6, ..., 25$. For every value of i, fit $g_1(x)$ and $g_2(x)$ to each of the training sets and use the validation set to choose the best model $g^*(x)$. Consequently, calculate for each value of i the expected error for both $E_{out}(g^*)$ and $E_{val}(g^*)$. Plot the expected errors as a function of the size of the validation set i and comment on the behaviour of the curves.

2. i. Suppose $x \sim \text{Uniform}(-1,1)$. Simulate a dataset of size $N = 50$ for the following model:

$$\begin{aligned} y_i &= \sin(\pi x_i) + \epsilon_i; \quad -1 \leq x \leq 1,\ i = 1, ..., N \\ \text{where: } \epsilon_i &\sim \text{Normal}(0,1). \end{aligned}$$

   Plot the simulated data together with the model.

   ii. Consider the model:

$$y_i = \sum_{q=0}^{10} \beta_q L_q(x)$$

   where $L_q(x)$ is a Legendre function of order q. Fit this model to the simulated dataset in Question 2 part i. using regularisation parameter values of $\lambda = 0$ and $\lambda = 5$. Plot the two fitted curves together with the dataset and the true model. Comment on the plots.

   iii. Using 10 fold cross-validation, calculate the cross-validation error for the 10-th order polynomial for $0.1 \leq \lambda \leq 10$ and plot. Consequently, select the optimal value of λ and plot the resulting fitted model together with the data and the true model.

3. The file *faces.zip* on Vula contains 10 greyscale images for each of 40 different people. Download the file and use the package "pixmap" to read in the images to *R*.

i. Plot the mean image as well as the standard deviation image. Hence use these to plot the original and a scaled version of "*168.pgm*".

ii. Using all the images, derive the first ten eigenfaces and plot them.

iii. Plot the scaled version of image "*115.pgm*". Alongside it, plot the reconstructed versions of the image when the first 5, 50 and 200 eigenfaces are used. Comment on your results.