Elements of Statistical Learning
Homework Sheet 1

Optional. Only if you want. You can hand in solutions to questions 7 and 10 for marking for formative FEEDBACK. Questions 7 and 10 are marked below in blue and with $*$. Hand in solutions on Blackboard on Tuesday February 7th. Ensure your name is CLEARLY written on them. If you submit more than one page, ensure that the pages uploaded as a single document and in the right order, please!

Guidance. These questions are partly practice and partly to help you think about what you have seen in lectures. You do not have to answer all of them, but obviously the more you answer the more practice you will get. Since this is a formative work, please feel to work in groups and talk to each other about the questions, but ensure YOU participate!

It is also important that you keep check of your overall workload, so do whatever you think is reasonable to ensure you maintain a healthy workload and healthy work/life balance: this is different for everybody.

1. For the data $(h_i, w_i)_{i=1}^n$, where n is the number of data points. Derive the formulae for the least squares estimates $(\tilde{a}, \tilde{b})$ as given in lectures.

2. To make calculations easier, let $\bar{h} = n^{-1} \sum_{i=1}^n h_i$ be the sample mean of the hs. Form a new, temporary variable $g_i = h_i - \bar{h}$. Show that the linear regression estimates of $(g_i, w_i)_{i=1}^n$ are given by

$$\tilde{b}^{(g)} = \sum_{i=1}^n g_i w_i / \sum_{i=1}^n g_i^2 \quad \text{and} \quad \tilde{a}^{(g)} = \bar{w}.$$

   Show how to obtain $(\tilde{a}, \tilde{b})$ from $(\tilde{a}^{(g)}, \tilde{b}^{(g)})$.

3. Show that the fitted values $\hat{w}_j$ can be written as a linear combination of the $\{w_i\}_{i=1}^n$ values. I.e. show that you can write $\hat{w}_j = \sum_{i=1}^n r_i w_i$ for some set of values $\{r_i\}$. This is another characterisation of linear methods.

4. Find out the definition of leverage and Cook's distance and what value of Cook's distance for a point would mean that you take a closer interest in that point.

5. Suppose A is a symmetric $p \times p$ matrix and β is a p-vector. Let $w = \beta^T A \beta$. Show that $\frac{\partial w}{\partial \beta} = 2A\beta$.

6. Compute the eigenvalues and eigenvectors of $\begin{pmatrix} 1+\lambda & \rho \\ \rho & 1+\lambda \end{pmatrix}$.

7. $*$ Show that if $\beta_j \sim N(0, \tau^2)$ is a prior distribution for the β parameters in a linear model and $Y_i \sim N(\beta_0 + x_i^T \beta, \sigma^2)$, then the ridge regression estimate is the posterior mode, work out the variance of the posterior distribution.

8. Let X be a $n \times p$ data matrix. Suppose we centre each value in the matrix by $x^c_{i,j} = x_{i,j} - \bar{x}_j$, where $\bar{x}_j$ is the sample mean of the jth variable, i.e. $\bar{x}_j = n^{-1} \sum_{i=1}^n x_{i,j}$.

   Show that the centred data matrix $X_C = (I_n - 11^T/n)X$, where 1 is the n-vector of ones. Show that the centering matrix $C = I_n - 11^T/n$ is idempotent and a projection matrix. Geometrically interpret the action of applying the centering matrix.

9. Work out the bias for the ridge regression estimate when $X^T X = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ for ridge parameter λ.

10. * Prove that $\hat{\beta}_j^{\text{lasso}} = \text{sgn}(\hat{\beta}^{\text{ls}})(|\hat{\beta}_j^{\text{ls}}| - \lambda)^+$ for the case $\hat{\beta}_j^{\text{ls}} < 0$ in the lecture notes.

11. Work out the mean squared error on the `swiss` data for different values of λ for ridge regression or lasso. In each case, look at the parameter values.

12. Use the `lars` package to carry out a least angle regression on the `swiss` data. Examine the associated plots and the fit.

[Last updated: 21st January 2021]