Modern Statistical Methods — Example Sheet 1

Lucas Riedstra

...

Question 1. *Consider minimising the following objective involving response $Y \in \mathbb{R}^n$ and design matrix $X \in \mathbb{R}^{n\times p}$ over $(\mu, \beta) \in \mathbb{R} \times \mathbb{R}^p$:*

$$\|Y - \mu\mathbf{1} - X\beta\|_2^2 + J(\beta).$$

Here $J\colon \mathbb{R}^p \to \mathbb{R}$ is an arbitrary penalty function. Suppose $\bar{X}_k = 0$ for $k = 1, \dots, p$. Assuming that a minimiser $(\hat{\mu}, \hat{\beta})$ exists, show that $\hat{\mu} = \bar{Y}$. Now take $J(\beta) = \lambda\|\beta\|_2^2$ so we have the ridge regression objective. Show that

$$\hat{\beta} = (X^\top X + \lambda I)^{-1} X^\top Y.$$

From here onwards, whenever we refer to ridge regression, we will assume X has had its columns mean-centred.

Solution. If a minimiser $\hat{\mu}$ exists, then the derivative of the objective must be 0. We expand the objective function as

$$\begin{aligned}\|Y - \mu\mathbf{1} - X\beta\|_2^2 + J(\beta) &= \sum_{i=1}^n \left(Y_i - \mu - (X\beta)_i\right)^2 + J(\beta) \\ &= \sum_{i=1}^n \left(Y_i - \mu - \sum_{j=1}^p X_{ij}\beta_j\right)^2.\end{aligned}$$

Now, differentiating w.r.t. μ yields

$$\begin{aligned}\frac{\partial}{\partial\mu}\|Y - \mu\mathbf{1} - X\beta\|_2^2 + J(\beta) &= -2\sum_{i=1}^n \left(Y_i - \mu - \sum_{j=1}^p X_{ij}\beta_j\right) \\ &= -2\left(\sum_{i=1}^n Y_i - n\mu - \sum_{i=1}^n\sum_{j=1}^p X_{ij}\beta_j\right) \\ &= -2\left(\sum_{i=1}^n Y_i - n\mu - \sum_{j=1}^p \beta_j(\sum_{i=1}^p X_{ij})\right) \\ &= -2\left(\sum_{i=1}^n Y_i - n\mu\right),\end{aligned}$$

since $\sum_{i=1}^p X_{ij} = 0$ for all i by assumption. Now, setting the derivative to 0 shows that $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n Y_i = \bar{Y}$ is the minimiser. Since the second derivative of the objective function is constantly $+2$, this $\hat{\mu}$ is indeed a minimum.

To compute $\hat{\beta}$, we differentiate the objective function w.r.t. β:

$$\nabla_\beta \|Y - \mu\mathbf{1} - X\beta\|_2^2 + \lambda\|\beta\|_2^2 = ...$$

TODO: Finish

Question 2. *Consider performing ridge regression when $Y = X\beta^0 + \varepsilon$, where $X \in \mathbb{R}^{n\times p}$ has full column rank, and $\mathrm{Var}(\varepsilon) = \sigma^2 I$. Let the SVD of X be $UDV^\top$ and write $U^\top X\beta^0 = \gamma$. Show that*

$$\frac{1}{n}\mathbb{E}\left\|X\beta^0 - X\hat{\beta}_\lambda^R\right\|_2^2 = \frac{1}{n}\sum_{j=1}^{p}\left(\frac{\lambda}{\lambda + D_{jj}^2}\right)^2\gamma_j^2 + \frac{\sigma^2}{n}\sum_{j=1}^{p}\frac{D_{jj}^4}{(\lambda + D_{jj}^2)^2}.$$

Solution. We write $\hat{\beta} \coloneqq \hat{\beta}_\lambda^R$. We write out

$$X\hat{\beta} = UDV^\top(V^\top D^2 V + \lambda I)^{-1}V^\top DU^\top Y = UD^2(D^2 + \lambda I)^{-1}U^\top Y = \sum_j \frac{D_{jj}^2}{\lambda + D_{jj}^2}U_j U_j^\top Y.$$

Since $X\beta^0 = UDV^\top\beta^0$ lies in the range of U, we have $X\beta^0 = \sum_j U_j U_j^\top X\beta^0$.

$$\begin{aligned}
X\beta^0 - X\hat{\beta} &= \sum_j U_j U_j^\top X\beta^0 - \sum_j \frac{D_{jj}^2}{\lambda + D_{jj}^2}U_j U_j^\top Y \\
&= \sum_j \left(U_j^\top X\beta^0 - \frac{D_{jj}^2 U_j^\top Y}{\lambda + D_{jj}^2}\right)U_j \\
&= \sum_j \left(\frac{D_{jj}^2 U_j^\top(X\beta^0 - Y) + \lambda U_j^\top X\beta^0}{\lambda + D_{jj}^2}\right)
\end{aligned}$$

and therefore

$$\begin{aligned}
\frac{1}{n}\mathbb{E}\left\|X\beta^0 - X\hat{\beta}\right\|_2^2 &= \frac{1}{n}\sum_j \mathbb{E}\left(\frac{D_{jj}^2 U_j^\top(X\beta^0 - Y) + \lambda U_j^\top X\beta^0}{\lambda + D_{jj}^2}\right)^2 \\
&\overset{*}{=} \frac{1}{n}\sum_j \frac{D_{jj}^4}{(\lambda + D_{jj}^2)^2}\mathbb{E}[(U_j^\top(X\beta^0 - Y))^2] + \frac{1}{n}\sum_j\left(\frac{\lambda}{\lambda + D_{jj}^2}\right)^2(U_j^\top X\beta^0)^2 \\
&\overset{**}{=} \frac{\sigma^2}{n}\sum_j \frac{D_{jj}^4}{(\lambda + D_{jj}^2)^2} + \frac{1}{n}\sum_j\left(\frac{\lambda}{\lambda + D_{jj}^2}\right)^2\gamma_j^2.
\end{aligned}$$

Note that the cross-terms in $(*)$ disappear since they are linear combinations of $\mathbb{E}(X\beta^0 - Y) = 0$, while the equality $**$ holds since

$$\mathbb{E}[(U_j^\top(X\beta^0 - Y))^2] = \mathrm{Var}(U_j^\top(X\beta^0 - Y)) = U_j^\top\,\mathrm{Var}(X\beta^0 - Y)U_j = \sigma^2 U_j^\top U_j = \sigma^2.$$

Question 3. *Show that the ridge regression estimates can be otained by ordinary least squares regression on an augmented data set with $\sqrt{\lambda}I$ added to the botten of X, and p zeroes added to the end of the response Y.*

Solution. Let $\tilde{X}, \tilde{Y}$ be the new data set, then our ordinary least squares objective function is

$$\begin{aligned} L(\beta) &:= \left\| \tilde{Y} - \tilde{X}\beta \right\|^2 \\ &= \|Y - X\beta\|^2 + \left\| \sqrt{\lambda}\beta \right\|^2 \\ &= \|Y - X\beta\|^2 + \lambda \|\beta\|^2, \end{aligned}$$

which is exactly our ridge regression objective function.

Question 4. *In the following, assume that forming AB where $A \in \mathbb{R}^{a\times b}, B \in \mathbb{R}^{b\times c}$ requires $O(abc)$ computational operations, and that if $M \in \mathbb{R}^{d\times d}$ is invertible, then forming M^{-1} requires $O(d^3)$ operations.*

(a) Suppose we wish to apply ridge regression to data $(Y, X) \in \mathbb{R}^n \times \mathbb{R}^{n\times p}$ with $n \gg p$. A complication is that the data is split into m separate datasets of size $n/m \in \mathbb{N}$,

$$Y = \begin{pmatrix} Y^{(1)} \\ \vdots \\ Y^{(m)} \end{pmatrix}, \quad X = \begin{pmatrix} X^{(1)} \\ \vdots \\ X^{(m)} \end{pmatrix},$$

with each dataset located on a different server. Moving large amounts of data between servers is expensive. Explain how one can produce ridge estimates $\hat{\beta}_\lambda$ by communicating only $O(p^2)$ numbers from each server to some central server. What is the total order of the computation time required at each server, and at the central server for your approach?

(b) Now suppose instead that $p \gg n$ and it is instead the variable sthat are split across m servers, so each server has only a subset of $p/m \in \mathbb{N}$ variables for each observation, and some central server stores Y. Explain how one can obtain the fitted values $X\hat{\beta}_\lambda$ communicating only $O(n^2)$ numbers from each server to the central server. What is the total order of the computation time required at each server, and at the central server for your approach?

Solution. (a) Note that

$$X^\top Y = \begin{pmatrix} X^{(1)\top} \dots X^{(m)\top} \end{pmatrix} \begin{pmatrix} Y^{(1)} \\ \vdots \\ Y^{(m)} \end{pmatrix} = \sum_{i=1}^m X^{(i)\top} Y^{(i)}.$$

Each of the $X^{(i)\top}Y^{(i)}$ can be computed at the data server (requiring $O(np)$ computation) and produces p numbers per data server.

Analogously, we have

$$X^\top X = \sum_{i=1}^m X^{(i)\top} X^{(i)},$$

which can also be computed at the data server (requiring $O(np^2)$ computation) and produces p^2 numbers per data server. In the end, each server will therefore have $O(p^2)$ numbers to send, having done $O(np^2)$ computations.

When each server sends their numbers to the central data server, the central data server can then compute $(X^\top X + \lambda I)^{-1}$ using $O(p^3)$ computations and then multiply that with $X^\top Y$ which will require $O(np)$ computations.

(b) In this case, we have

$$X = \begin{pmatrix} X^{(1)} \cdots X^{(m)} \end{pmatrix}.$$

We use the alternative form of $\hat{\beta}_\lambda$, namely

$$X\hat{\beta}^R_\lambda = XX^\top (XX^\top + \lambda I)^{-1} Y.$$

Now, we have $XX^\top = \sum_{i=1}^m X^{(i)} X^{(i)\top}$, and each $X^{(i)} X^{(i)\top}$ can be computed at the data server, requiring $O(n^2 p)$ computations and producing $O(n^2)$ numbers to send.

After sending these numbers over to the central server, the inverse $(XX^\top + \lambda I)^{-1}$ can be computed in $O(n^3)$ time, then multiplied from the left with $XX^\top$ in $O(n^3)$ time, and multiplied with Y in $O(n^2)$ time. The total amount of computation at the central server will therefore be $O(n^3)$ as well.