Modern Statistical Methods — Example Sheet 2

Lucas Riedstra

...

Question 1. *Let $Y \in \mathbb{R}^n$ be a vector of responses, $\Phi \in \mathbb{R}^{n\times p}$ a design matrix, $J\colon [0,\infty) \to [0,\infty)$ a strictly increasing function and $c\colon \mathbb{R}^n \to \mathbb{R}^n$ some cost function. Set $K = \Phi\Phi^\top$. Show, without using the representer theorem, that $\hat{\vartheta}$ minimises*

$$Q_1(\vartheta) \coloneqq c(Y, \Phi\vartheta) + J(\|\vartheta\|_2^2)$$

over $\vartheta \in \mathbb{R}^p$ if and only if $\Phi\hat{\vartheta} = K\hat{\alpha}$ and $\hat{\alpha}$ minimises

$$Q_2(\alpha) \coloneqq c(Y, K\alpha) + J(\alpha^\top K\alpha)$$

over $\alpha \in \mathbb{R}^n$.

Proof. Let $\hat{\vartheta}$ be a minimiser of Q_1, and write $\hat{\vartheta} = \Phi^\top\hat{\alpha} + \hat{\beta}$ with $\Phi^\top\hat{\alpha} \in \mathcal{N}(\Phi)^\perp = \mathcal{R}(\Phi^\top)$, $\hat{\beta} \in \mathcal{N}(\Phi)$.

Noting that $K\hat{\alpha} = \Phi\Phi^\top\hat{\alpha} = \Phi\hat{\vartheta}$ and $\left\|\Phi^\top\hat{\alpha}\right\| = \alpha^\top K\alpha$ we see

$$Q_1(\vartheta) = c(Y, K\hat{\alpha}) + J(\alpha^\top K\alpha + \|\hat{\beta}\|^2),$$

and therefore it is necessary that $\hat{\beta} = 0$. The claim follows. $\square$

Question 2. *Let $x, x' \in \mathbb{R}^p$ and let $\psi \in \{-1, 1\}^p$ be a random vector with independent components taking values $-1, 1$ each with probability $1/2$. Show that $\mathbb{E}(\psi^\top x\psi^\top x') = x^\top x'$. Construct a random feature map $\hat{\varphi}\colon \mathbb{R}^p \to \mathbb{R}$ such that $\mathbb{E}\{\hat{\varphi}(x)\hat{\varphi}(x')\} = (x^\top x)^2$.*

Solution. We have

$$\psi^\top x\psi^\top x' = \left(\sum_i \psi_i x_i\right)\left(\sum_j \psi_j x'_j\right) = \sum_i x_i x'_i + 2\sum_{i<j} \psi_i\psi_j x_i x'_j.$$

Noting that for $i \neq j$ we have $\mathbb{E}[\psi_i\psi_j] = \mathbb{E}[\psi_i]\mathbb{E}[\psi_j] = 0$ it follows that $\mathbb{E}[\psi^\top x\psi^\top x'] = \sum_i x_i x'_i = x^\top x'$.

Let ψ_* be an identical independent copy of ψ and define $\hat{\varphi}(x) = \psi^\top x\psi_*^\top x$. Then we find

$$\mathbb{E}[\hat{\varphi}(x)\hat{\varphi}(x')] = \mathbb{E}[\psi^\top x\psi^\top x']\mathbb{E}[\psi_*^\top x\psi_*^\top x'] = (x^\top x')^2.$$

Question 3. *Let $\mathcal{X} = \mathcal{P}(\{1,\dots,p\})$ and $z, z' \in \mathcal{X}$. Let k be the Jaccard similarity kernel. Let π be a random permutation of $\{1,\dots,p\}$. Let $M = \min\{\pi(j) \mid j \in z\}$, $M' = \min\{\pi(j) \mid j \in z'\}$. Show that*

$$\mathbb{P}(M = M') = k(z, z'),$$

when $z, z' \neq \varnothing$. Now let $\psi \in \{-1, 1\}^p$ be a random vector with i.i.d. components taking the values -1 or 1, each with probability 1/2. By considering $\mathbb{E}[\psi_M\psi_{M'}]$ show that the Jaccard similarity kernel is indeed a kernel. Explain how we can use the ideas above to approximate kernel ridge regression with Jaccard similarity, when n is very large (you may assume none of the data points are the empty set).

Proof. We have

$$\mathbb{P}(M = M') = \mathbb{P}\left(\operatorname*{arg\,min}_{j \in z \cup z'} \pi(j) \in z \cap z'\right) = \frac{|z \cap z'|}{|z \cup z'|} = k(z, z') \quad \text{since } \pi \text{ is random.}$$

Furthermore, we have

$$\mathbb{E}[\psi_M \psi_{M'}] = \mathbb{P}(M = M')\mathbb{E}[\psi_M^2] + \mathbb{P}(M \neq M')\mathbb{E}[\psi_M \psi_{M'}] = k(z, z'),$$

since for $M \neq M'$ we have $\mathbb{E}[\psi_M \psi_M'] = \mathbb{E}[\psi_M]\mathbb{E}[\psi_{M'}] = 0$. Let $z_1, \ldots, z_n \in \mathcal{X}$ with corresponding $M_1, \ldots, M_n$, and write $\hat{\psi} = (\psi_{M_1}, \ldots, \psi_{M_n})^\top$, then the kernel matrix K is given by $\mathbb{E}[\hat{\psi}\hat{\psi}^\top]$ which is positive semidefinite.

Using the random feature map $\hat{\varphi}(z) = \psi_{M_z}$ we can approximate kernel ridge regression using the random feature map method. □

Question 4. *Consider the logistic regression model where we assume $Y_1, \ldots, Y_n \in \{-1, 1\}$ are independent and*

$$\log\left(\frac{\mathbb{P}(Y_i = 1)}{\mathbb{P}(Y_i = -1)}\right) = x_i^\top \beta^0.$$

Show that the maximum likelihood estimate β minimises

$$\sum_{i=1}^n \log\big(1 + \exp\big(-Y_i x_i^\top \beta\big)\big)$$

over $\beta \in \mathbb{R}^p$.

Proof. Let $(y_1, \ldots, y_n)$ be the responses, then the likelihood function becomes

$$L(\beta) = \mathbb{P}(Y_1, \ldots, Y_n \mid \beta)$$

Note that

$$\frac{\mathbb{P}(Y_i = 1)}{1 - \mathbb{P}(Y_i = 1)} = \exp\big(x_i^\top \beta\big) \implies \mathbb{P}(Y_i = 1) = \frac{1}{1 + \exp\big(-x_i^\top \beta\big)},$$

and analogously

$$\mathbb{P}(Y_i = -1) = \frac{1}{1 + \exp\big(x_i^\top \beta\big)}.$$

We can combine the above formulas as

$$\mathbb{P}(Y_i = y_i) = \frac{1}{1 + \exp\big(-y_i x_i^\top \beta\big)}.$$

Therefore our the MLE $\hat{\beta}$ maximises

$$L(\beta) = \prod_{i=1}^n \frac{1}{1 + \exp\big(-Y_i x_i^\top \beta\big)},$$

and it also maximises the log-likelihood function

$$\log(L(\beta)) = -\sum_{i=1}^n \log\big(1 + \exp\big(-Y_i x_i^\top \beta\big)\big)$$

which is of course equivalent to minimising

$$-\log(L(\beta)) = \sum_{i=1}^n \log\big(1 + \exp\big(-Y_i x_i^\top \beta\big)\big).$$

□

Question 5. *Consider the following algorithm for model selection when we have a response $Y \in \mathbb{R}^n$ and a matrix of predictors $X \in \mathbb{R}^{n\times p}$.*

(a) First centre Y and all the columns of X. Initiale the current model $M \subseteq \{1, \dots, p\}$ to be $\varnothing$ and set the current residual R to be Y.

(b) Find the variable k^* in M^c most correlated with the current residual R. Set M to be $M \cup \{k^*\}$. Replace R with the residual from regressing R onto X_{k^*}. Further replace each variable in M^c with the residual from regressing itself onto X_{k^*}.

(c) Continue the previous step unto $R = 0$.

Show that this algorithm is equivalent to forward selection.

Hint: *Use induction on the iteration m of the algorithm. Consider strengthening the natural induction hypothesis that the model at iteration m is the same as that selected after m steps of forward selection.*

Solution. We follow the hint and use induction on the iteration m of the algorithm. For the base case, note that centering Y is equivalent to fitting an intercept-only model. ???

Question 6. *Show that if W is mean-zero and sub-Gaussian with parameter σ, then* $\mathrm{Var}(W) \leq \sigma^2$*.*

Proof. If W is mean-zero, then $\mathrm{Var}(W) = \mathbb{E}(W^2)$. By the proof of lemma 15 we have $\mathbb{E}(W^2) \leq 2\sigma^2$, but this bound is too loose.

Since W is sub-Gaussian, we have for all $\alpha \in \mathbb{R}$ that

$$\begin{aligned}
\mathbb{E}[e\alpha W] &\leq e^{\alpha^2\sigma^2/2} \\
\sum_{k=0}^{\infty} \frac{\mathbb{E}[W^k]}{k!}\alpha^k &\leq \sum_{k=0}^{\infty} \frac{\sigma^{2k}}{2^k k!}\alpha^{2k} \\
\frac{\mathbb{E}[W^2]}{2}\alpha^2 + \sum_{k=3}^{\infty} \frac{\mathbb{E}[W^k]}{k!}\alpha^k &\leq \frac{\sigma^2}{2}\alpha^2 + \sum_{k=2}^{\infty} \frac{\sigma^{2k}}{2^k k!}\alpha^{2k} \\
\frac{1}{2}\alpha^2\left(\sigma^2 - \mathbb{E}[W^2]\right) &\geq \alpha^3 P(\alpha),
\end{aligned}$$

where P is a power series in α. Rescaling we find

$$\frac{1}{2}(\sigma^2 - \mathbb{E}[W^2]) \geq \alpha P(\alpha),$$

and letting $\alpha \to 0$ also lets $\alpha P(\alpha) \to 0$, and therefore $\sigma^2 - \mathbb{E}[W^2] \geq 0$ so $\mathrm{Var}(W) \leq \sigma^2$. □

Question 7. *Verify Hoeffding's lemma for the special case where W is a Rademacher random variable, so W takes the values $-1, 1$ each with probability $1/2$.*

Proof. If W is a Rademacher random variable then we have, using $(2k)! \geq 2^k k!$ for each $k \in \mathbb{N}$, that

$$\mathbb{E}[e^{\alpha W}] = \frac{1}{2}e^{-\alpha} + \frac{1}{2}e^{\alpha} = \sum_{k=0}^{\infty} \frac{\alpha^{2k}}{(2k)!} \leq \sum_{k=0}^{\infty} \frac{\alpha^{2k}}{2^k k!} = e^{\alpha^2/2}$$

which shows that W is sub-Gaussian with parameter $1 = (1-(-1))/2$, so Hoeffding's lemma holds indeed. □

Question 8. *(a) Let $W \sim \chi_d^2$. Show that*

$$\mathbb{P}(|W/d - 1| \geq t) \leq 2e^{-dt^2/8}$$

for $t \in (0, 1)$. You may use the facts that the mgf of a χ_1^2 random variable is $(1-2\alpha)^{-1/2}$ for $\alpha < 1/2$, and $e^{-\alpha}(1-2\alpha)^{-1/2} \leq e^{2\alpha^2}$ when $|\alpha| < 1/4$.

(b) *Let $A \in \mathbb{R}^{d\times p}$ have i.i.d. standard normal entries. Fix $u \in \mathbb{R}^p$. Use the result above to conclude that*

$$\mathbb{P}\left(\left|\frac{\|Au\|_2^2}{d\|u\|_2^2} - 1\right| \geq t\right) \leq 2e^{-dt^2/8}.$$

(c) *Suppose we have data $u_1, \dots, u_n \in \mathbb{R}^p$, with p large and $n \geq 2$. Show that for a given $\varepsilon \in (0,1)$ and $d > 16\log(n/\sqrt{\varepsilon})/t^2$, each data point may be compressed down to $u_i \mapsto Au_i/\sqrt{d} = w_i$, whilst approximately preserving the distance between the points:*

$$\mathbb{P}\left(1 - t \leq \frac{\|w_i - w_j\|_2^2}{\|u_i - u_j\|_2^2} \leq 1 + t \text{ for all } i \neq j \in \{1, \dots, n\}\right) \geq 1 - \varepsilon.$$

This is the famous Johnson-Lindenstrauss Lemma.

Proof. (a) We have

$$\begin{aligned}\mathbb{P}(|W/d - 1| \geq t) &= \mathbb{P}(W/d - 1 \geq t) + \mathbb{P}(W/d - 1 \leq -t)\\ &= \mathbb{P}(W \geq d(1+t)) + \mathbb{P}(W \leq d(1-t)).\end{aligned}$$

Now we use Chernoff bounds: we have

$$\begin{aligned}\mathbb{P}(W \geq d(1+t)) &\leq \inf_{\alpha>0} e^{-\alpha d(1+t)}\mathbb{E}[e^{\alpha W}] \leq \inf_{0<\alpha<\frac{1}{2}} e^{-\alpha dt}\left(\frac{e^{-\alpha}}{\sqrt{1-2\alpha}}\right)^d\\ &\leq \inf_{0<\alpha<\frac{1}{4}} e^{d(2\alpha^2-\alpha t)} \overset{\star}{=} e^{dt^2/8},\end{aligned}$$

where $\star$ is a consequence of plugging in the minimum $\alpha = t/4 \in (0, 1/4)$.

Analogously, we have for $\alpha > 0$ that

$$\begin{aligned}\mathbb{P}(W \leq d(1-t)) &\leq \inf_{\alpha>0} e^{\alpha d(1-t)}\mathbb{E}[e^{-\alpha W}] \leq \inf_{-\frac{1}{2}<\alpha<0} e^{\alpha dt}\left(\frac{e^{-\alpha}}{\sqrt{1-2\alpha}}\right)^d\\ &\leq \inf_{-\frac{1}{4}<\alpha<0} e^{d(2\alpha^2+\alpha t)} \overset{\star}{=} e^{dt^2/8},\end{aligned}$$

where $\star$ is now a consequence of plugging in the minimum $\alpha = -t/4 \in (-1/4, 0)$.

Plugging what we found back into our original equation we obtain for $t \in (0,1)$ indeed

$$\mathbb{P}(|W/d - 1| \geq t) \leq 2e^{dt^2/8}.$$

(b) We wish to show that $\|Au\|^2/\|u\|^2 \sim \chi_d^2$. For this, let $Z_1, \dots, Z_p$ be i.i.d. $N(0,1)$ random variables, then

$$(Au)_i = \sum_{j=1}^p A_{ij}u_j \sim \sum_{j=1}^p u_j Z_j \sim N(0, \sum_i u_i^2) = N(0, \|u\|_2^2).$$

Therfore, we have

$$\frac{\|Au\|^2}{\|u^2\|} = \frac{\sum_i (Au)_i^2}{\|u\|^2} = \sum_i \left(\frac{(Au)_i}{\|u\|}\right)^2 \sim \sum_i Z_i^2 \sim \chi_d^2.$$

Combining this with (a) proves the claim.

(c) We have

$$\begin{aligned}
&\mathbb{P}\left(1-t \le \frac{\|w_i-w_j\|_2^2}{\|u_i-u_j\|_2^2} \le 1+t \text{ for all } i \neq j \in \{1,\dots,n\}\right)\\
&= \mathbb{P}\left(\left|\frac{\|A(u_i-u_j)\|^2}{d\|u_i-u_j\|^2}-1\right| \le t \text{ for all } i \neq j \in \{1,\dots,n\}\right)\\
&= 1-\mathbb{P}\left(\left|\frac{\|A(u_i-u_j)\|^2}{d\|u_i-u_j\|^2}-1\right| \ge t \text{ for some } i \neq j \in \{1,\dots,n\}\right)\\
&\overset{\star}{\ge} -\frac{n^2}{2}\cdot 2e^{-dt^2/8} \ge 1-\varepsilon,
\end{aligned}$$

where $\star$ follows from the fact that there are $\frac{n(n-1)}{2} \le \frac{n^2}{2}$ pairs $i \neq j$, and the last inequality follows from

$$e^{-dt^2/8} \le e^{-2\log\left(n/\sqrt{\varepsilon}\right)} = \left(\frac{n}{\sqrt{\varepsilon}}\right)^{-2} = \frac{\varepsilon}{n^2}.$$

□

In the following questions, assume that X has had its columns centred and scaled to have ℓ_2-norm $\sqrt{n}$, and that Y is also centred.

Question 9. *Show that any two Lasso solutions when $\lambda > 0$ must have the same ℓ_1-norm.*

Proof. By proposition 21, $X\hat{\beta}^L_\lambda$ is unique, so if $\hat{\beta}$, $\hat{\gamma}$ are Lasso solutions then $X\hat{\beta} = X\hat{\gamma}$ so

$$Q_\lambda(\hat{\beta}) = \frac{1}{2n}\left\|Y - X\hat{\beta}\right\|_2^2 + \lambda\left\|\hat{\beta}\right\|_1 = \frac{1}{2n}\left\|Y - X\hat{\beta}\right\|_2^2 + \lambda\|\hat{\gamma}\|_1 = Q_\lambda(\hat{\gamma}),$$

which implies $\lambda\|\hat{\beta}\|_1 = \lambda\|\hat{\gamma}\|_1$. For $\lambda > 0$ this shows $\left\|\hat{\beta}\right\|_1 = \|\hat{\gamma}_1\|$. □

Question 10. *Carathéodory's Lemma states that if $S \subseteq \mathbb{R}^{d'}$ is in a subspace of dimension d, any v that is a convex combination of points in S can be expressed as a convex combination of $d+1$ points from S.*

With this knowledge, show that for any value of λ, there is always a Lasso solution with no more than n non-zero coefficients.

Proof. Let $v_1, \dots, v_p \subseteq \mathbb{R}^n$ be the columns of X. Because X has centred columns, we know that $v_1, \dots, v_p$ lie in the $(n-1)$-dimensional subspace of vectors in $\mathbb{R}^n$ with mean zero.

Now, let $\hat{\beta}$ be any lasso solution. If $\hat{\beta} = 0$, we are done, so suppose $\hat{\beta} \neq 0$, and write

$$X\hat{\beta} = \begin{bmatrix} v_1 & \cdots & v_p \end{bmatrix}\begin{bmatrix}\hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p\end{bmatrix} = \left\|\hat{\beta}\right\|_1 \begin{bmatrix}\operatorname{sgn}(\hat{\beta}_1)v_1 & \cdots & \operatorname{sgn}(\hat{\beta}_p)v_p\end{bmatrix}\begin{bmatrix}\left|\hat{\beta}_1\right|/\left\|\hat{\beta}\right\|_1 \\ \vdots \\ \left|\hat{\beta}_p\right|/\left\|\hat{\beta}\right\|_1\end{bmatrix}.$$

If we define $S = \left\{\left\|\hat{\beta}\right\|_1 \operatorname{sgn}(\hat{\beta}_1)v_1, \dots, \left\|\hat{\beta}\right\|_1 \operatorname{sgn}(\hat{\beta}_p)v_p\right\}$, then we see that $X\hat{\beta}$ can be written as a convex combination of points in S. Since the points in S also all have mean zero, by Caratheodory's lemma there exists a vector $\gamma \in \mathbb{R}^p$ with at most n nonzero coefficients, nonnegative entries, and $\|\gamma\|_1 \ge 0$ such that

$$X\hat{\beta} = \left\|\hat{\beta}\right\|_1 \begin{bmatrix}\operatorname{sgn}(\hat{\beta}_1)v_1 & \cdots & \operatorname{sgn}(\hat{\beta}_p)v_p\end{bmatrix}\gamma = X\left(\left\|\hat{\beta}\right\|_1 \begin{bmatrix}\operatorname{sgn}(\hat{\beta}_1)\gamma_1 \\ \vdots \\ \operatorname{sgn}(\hat{\beta}_p)\gamma_p\end{bmatrix}\right) =: X\hat{\beta}'.$$

Now, we have $X\hat{\beta} = X\hat{\beta}'$ and $\left\|\hat{\beta}\right\|_1' = \left\|\hat{\beta}\right\|_1$, which shows that $\hat{\beta}'$ is a Lasso solution with no more than n nonzero coefficients. $\square$

Question 11. *Show that if $\lambda \geq \lambda_{\max} := \left\|X^\top Y\right\|_\infty / n$ then $\hat{\beta}^L_\lambda = 0$.*

Proof. We first show that 0 is a lasso solution: we have

$$\frac{1}{n} X^\top (Y - X0) = \frac{1}{n} X^\top Y = \lambda \cdot \frac{\frac{1}{n} X^\top Y}{\lambda} =: \lambda \hat{\nu},$$

and since

$$\|\hat{\nu}\|_\infty = \frac{\left\|X^\top Y\right\|_\infty / n}{\lambda} \leq \frac{\left\|X^\top Y\right\|_\infty / n}{\left\|X^\top Y\right\|_\infty / n} \leq 1$$

it follows that 0 is a Lasso solution. By question 9, it follows that any other Lasso solution must also have norm 0, and therefore 0 is the only Lasso solution. $\square$

Question 12. *Show that when the columns of X are orthogonal (so necessarily $p \leq n$) and scaled to have ℓ_2-norm $\sqrt{n}$, the k-th component of the Lasso estimator is given by*

$$\hat{\beta}^L_{\lambda,k} = (\left|\hat{\beta}^{\text{OLS}}_k\right| - \lambda)_+ \operatorname{sgn}(\hat{\beta}^{\text{OLS}}_k).$$

What is the corresponding estimator if the ℓ_1 penalty $\|\beta\|_1$ in the Lasso objective is replaced by the ℓ_0 penalty $\|\beta\|_0 := \#\{k \mid \beta_k \neq 0\}$?

Proof. For simplicity, let $\hat{\beta}$ denote the proposed Lasso estimator and $\tilde{\beta}$ the OLS estimator. Since the columns of X are orthogonal, X is injective, so the Lasso estimator is unique (since $X\hat{\beta}^L_\lambda$ is unique).

Now, define $\gamma \in \mathbb{R}^p$ by

$$\gamma_i = \begin{cases} \tilde{\beta}_i, & \left|\tilde{\beta}_i\right| \leq \lambda, \\ \lambda, & \left|\tilde{\beta}_i\right| > \lambda, \\ -\lambda, & \left|\tilde{\beta}_i\right| < -\lambda \end{cases},$$

so that $\hat{\beta} = \tilde{\beta} - \gamma$. Then we find

$$\frac{1}{n} X^\top (Y - X\hat{\beta}) = \frac{1}{n} X^\top (Y - X(\tilde{\beta} - \gamma)) = \frac{1}{n} X^\top (Y - X\tilde{\beta}) + \frac{1}{n} X^\top X \gamma = \frac{1}{n}\gamma = \lambda \frac{1}{n\lambda}\gamma,$$

since $Y - X\tilde{\beta}$ is orthogonal to the columns of X so $X^\top (Y - X\tilde{\beta}) = 0$.

Now, we have $\left\|\frac{1}{n\lambda}\gamma\right\|_\infty = \frac{1}{n\lambda}\|\gamma\|_\infty \leq \frac{1}{n} \leq 1$, and furthermore, if $\hat{\beta}_i \neq 0$, then it is clear that $\operatorname{sgn}(\hat{\beta}_i) = \operatorname{sgn}(\gamma_i)$. This shows that $\hat{\beta}$ is indeed a Lasso solution.

Suppose we replace the Lasso objective by the ℓ_0 penalty. Let $\{X_1, \ldots, X_p\}$ be the columns of X and expand them to an orthonormal basis $\{X_1, \ldots, X_n\}$ of $\mathbb{R}^n$. Then we can write

$$\|Y - X\beta\|_2^2 = \left\|\sum_{i=1}^n P_{X_i} Y - \sum_{i=1}^p \beta_i X_i\right\|_2^2 = \sum_{i=1}^p (\langle X_i, Y\rangle - \beta_i)^2 + \sum_{j=p+1}^n \langle X_j, Y\rangle^2.$$

Since the latter sum does not depend on β, our objective function is equivalent to minimising

$$\sum_{i=1}^p \frac{1}{2n} (\langle X_i, Y\rangle - \beta_i)^2 + \lambda \mathbb{1}_{\beta_i \neq 0}.$$

Clearly, this is minimised by choosing $\beta_i = \tilde{\beta}_i$ if $\lambda \leq \frac{1}{2n}\langle X_i, Y\rangle^2$ and $\beta_i = 0$ otherwise (note, if $\lambda = \frac{1}{2n}\langle X_i, Y\rangle^2$, then choosing $\beta_i = 0$ or $\beta_i = \tilde{\beta}_i$ gives the same outcome). $\square$