MODERN STATISTICAL METHODS Part III
Example Sheet 2 (of 4) Michaelmas 2019

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

$$Q_1(\theta) := c(Y, \Phi\theta) + J(\|\theta\|_2^2)$$

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

$$Q_2(\alpha) := c(Y, K\alpha) + J(\alpha^T K\alpha)$$

over $\alpha \in \mathbb{R}^n$.
Solution: Suppose $\hat{\theta}$ minimises Q_1. Let $\Pi \in \mathbb{R}^{d\times d}$ be the orthogonal projection on to the row space of Φ. Then $\Phi\hat{\theta} = \Phi\Pi\hat{\theta}$ but $\|\hat{\theta}\|_2^2 = \|\Pi\hat{\theta}\|_2^2 + \|(I-\Pi)\hat{\theta}\|_2^2$. As J is strictly increasing, and by minimality of $\hat{\theta}$, we must have $(I-\Pi)\hat{\theta} = 0$. Thus $\hat{\theta}$ lies in the row space of Φ, so $\hat{\theta} = \Phi^T\alpha$, for some $\alpha \in \mathbb{R}^n$. We see then that

$$Q_1(\hat{\theta}) = c(Y, K\alpha) + J(\alpha^T K\alpha) = Q_2(\alpha).$$

Thus by minimality of $\hat{\theta}$, α must be such that it minimises Q_2. Now suppose $\hat{\alpha}$ minimises Q_2. Write $\hat{\theta} = \Phi^T\hat{\alpha}$, and note that $Q_1(\hat{\theta}) = Q_2(\hat{\alpha})$. Suppose $\tilde{\theta}$ has $Q_1(\tilde{\theta}) \le Q_1(\hat{\theta})$. Then from the argument above, $Q_1(\Pi\tilde{\theta}) \le Q_1(\tilde{\theta})$, and we may write $\Pi\tilde{\theta} = \Phi^T\tilde{\alpha}$ for some $\tilde{\alpha} \in \mathbb{R}^n$. But then we have

$$Q_2(\hat{\alpha}) = Q_1(\hat{\theta}) \ge Q_1(\tilde{\theta}) \ge Q_1(\Pi\tilde{\theta}) = Q_2(\tilde{\alpha}) \ge Q_2(\hat{\alpha}),$$

the last inequality following from minimality of $\hat{\alpha}$.

2. Let $x, x' \in \mathbb{R}^p$ and let $\psi \in \{-1,1\}^p$ be a random vector with independent components taking the values $-1, 1$ each with probability $1/2$. Show that $\mathbb{E}(\psi^T x \psi^T x') = x^T x'$. Construct a random feature map $\hat{\phi} : \mathbb{R}^p \to \mathbb{R}$ such that $\mathbb{E}\{\hat{\phi}(x)\hat{\phi}(x')\} = (x^T x')^2$.
Solution: $\mathbb{E}(x^T\psi\psi^T x') = x^T\mathbb{E}(\psi\psi^T)x' = x^T x'$. Given two vectors independent vectors $\psi^{(1)}, \psi^{(2)}$ each with the same distribution as ψ, set $\hat{\phi}(x) = x^T\psi^{(1)}x^T\psi^{(2)}$.

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

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

when $z, z' \neq \emptyset$. 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 that none of the data points are the empty set).
Solution: Let $H = \underset{k\in z\cup z'}{\operatorname{argmin}}\, \pi(k)$. Then

$$\mathbb{P}(M = M') = \mathbb{P}(M = M' = \pi(H)) = \mathbb{P}(H \in z \cap z') = \frac{|z \cap z'|}{|z \cup z'|}.$$

Now

$$\mathbb{E}(\psi_M \psi_{M'}) = \mathbb{E}(\psi_M \psi_{M'} | M = M')\mathbb{P}(M = M') + \mathbb{E}(\psi_M \psi_{M'} | M \neq M')\mathbb{P}(M \neq M') = \mathbb{P}(M = M').$$

Given $z_1, \ldots, z_n \in \mathcal{X} \setminus \{\emptyset\}$, let $M_i = \min\{\pi(k) : k \in z_i\}$ and define $S \in \{-1, 1\}^n$ by $S_i = \psi_{M_i}$. Let $K \in \mathbb{R}^{n \times n}$ have $K_{ij} = k(z_i, z_j)$. Then from the above we know $K = \mathbb{E}(SS^T)$ showing that it is positive semi-definite. If we have $z'_1, \ldots, z'_m = \emptyset$, the kernel matrix corresponding to $z_1, \ldots, z_n, z'_1, \ldots, z'_m$ would be block diagonal with one block corresponding to K above, and the other block being a matrix of ones. As both blocks are positive-semidefinite, k must be a kernel.

Now, suppose we have data $(z_1, Y_1), \ldots, (z_n, Y_n)$ with $z_i \in \mathcal{X}$, $Y_i \in \mathbb{R}$, and n is very large. We'd like to compute the fitted values in kernel ridge regression, which are given by $K(K + \lambda I)^{-1}Y$ where $K_{i,j} = K(z_i, z_j)$. This may be too expensive as inverting an $n \times n$ matrix costs $O(n^3)$ operations. We'd like to approximate the fitted values replacing K in the previous formula for an approximation $\hat{K} = XX^T$ where $X \in R^{n \times \ell}$ is a random feature matrix. Using the identity $\hat{K}(\hat{K} + \lambda I)^{-1}Y = X(X^TX + \lambda I)^{-1}X^TY$, we can obtain approximate fitted values in $O(\ell^3 + \ell^2 n)$ operations. We define X as follows. Let $\psi^{(1)}, \ldots, \psi^{(\ell)}$ be i.i.d. copies of ψ, and $\pi^{(1)}, \ldots, \pi^{(\ell)}$ be i.i.d. copies of π, then let

$$X_{i,k} = \frac{\psi^{(k)}_{M_i^{(k)}}}{\sqrt{\ell}}$$

where $M_i^{(k)} = \min\{\pi^{(k)}(j) : j \in z_i\}$. This definition ensures that

$$\hat{K}_{i,j} = (XX^T)_{i,j} = \frac{1}{\ell}\sum_{k=1}^{\ell} \psi^{(k)}_{M_i^{(k)}} \psi^{(k)}_{M_j^{(k)}} \approx K_{i,j}$$

as the left hand side is an average of ℓ i.i.d. variables with expectation $K_{i,j}$.

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

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

Show that the maximum likelihood estimate $\hat{\beta}$ minimises

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

over $\beta \in \mathbb{R}^p$.
Solution: Fix i and let $u = x_i^T \beta^0$. We have

$$\mathbb{P}(Y_i = 1) = \frac{e^u}{1 + e^u} = \frac{1}{1 + e^{-u}},$$

and $\mathbb{P}(Y_i = -1) = 1/(1 + e^u)$, so $\mathbb{P}(Y_i = y) = 1/(1 + e^{-uy})$, from which the result easily follows by noting that the Y_i are independent.

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

   (a) First centre Y and all the columns of X. Initialise the current model $M \subseteq \{1, \ldots, p\}$ to be $\emptyset$ 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 on X_{k^*}. Further replace each variable in M^c with the residual from regressing itself on X_{k^*}.

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

Show that this algorithm is equivalent to forward selection. *Hint: Use induction on the iteration m of the algorithm. Consider strengthening the natural inductive hypothesis that the model at iteration m is the same as that selected after m steps of forward selection.*
Solution: We use induction on the iteration of the algorithm. Write $X^{(m)}$ and $R^{(m)}$ for the transformed version of the matrix of predictors and residual at iteration m. Our inductive hypothesis will be that (i) at iteration m, the current model M is the same as that selected after m iterations of forward selection; (ii) that all variables in M^c are orthogonal to the column space of $X_M^{(m)}$ (i.e. ${X_M^{(m)}}^T X_{M^c}^{(m)} = 0$); (iii) that for $S \subseteq \{1, \ldots, p\}$, the column space of $X_{M \cup S}^{(m)}$ is the column space of $X_{M \cup S}$; and (iv) that $R^{(m)} = Y - X_M({X_M}^T X_M)^{-1}{X_M}^T Y$. First we show that X_{k^*} reduces the RSS by the greatest amount among all possible variables. Let P be the orthogonal projection onto X_M (or equivalently $X_M^{(m)}$). For $k \in M^c$,

$$\begin{aligned}
\min_\beta \|Y - X_{M\cup\{k\}}\beta\|_2^2 &= \min_\beta \|Y - X_{M\cup\{k\}}^{(m)}\beta\|_2^2 \\
&= \min_{\beta_k} \|(I-P)Y - X_k^{(m)}\beta_k\|_2^2 \\
&= \|R^{(m)}\|_2^2 - 2({X_k^{(m)}}^T R^{(m)})^2 / \|X_k^{(m)}\|_2^2.
\end{aligned}$$

using (ii) and then (iv). We see that if $X_{k^*}^{(m)}$ is the variable most correlated with $R^{(m)}$, it decreases the RSS the most. We have thus verified (i) at iteration $m+1$. Next, note that since each $X_k^{(m)}$ with $k \notin M \cup \{k^*\}$ has $X_k^{(m)}$ orthogonal to the column space of $X_M^{(m)}$, the residual from regressing $X_k^{(m)}$ on $X_{k^*}^{(m)}$ will still be orthogonal to the column space of $X_M^{(m)}$, and it will also be orthogonal to $X_{k^*}^{(m)}$. Thus, it will be orthogonal to the column space of $X_{M\cup\{k^*\}}^{(m)}$. This shows (ii) to hold true at iteration $m+1$. Next, we show that for $S \subseteq \{1, \ldots, p\}$, the column space of $X_{M\cup\{k^*\}\cup S}^{(m+1)}$ is that of $X_{M\cup\{k^*\}\cup S}$. This follows from noticing that each column of $X_S^{(m+1)}$ is simply a linear combination of the corresponding column of $X_S^{(m)}$ and $X_{k^*}^{(m)}$, so the column space of $X_{M\cup\{k^*\}\cup S}^{(m+1)}$ is the column space of $X_{M\cup\{k^*\}\cup S}^{(m)}$. Finally, as $X_{k^*}^{(m)}$ is orthogonal to the column space of $X_M^{(m)}$, writing Q for the orthogonal projection on to $X_{M\cup\{k^*\}}$ or equivalently $X_{M\cup\{k^*\}}^{(m)}$, we have

$$\begin{aligned}
Y - QY &= Y - (PY + X_{k^*}^{(m)}{X_{k^*}^{(m)}}^T Y / \|X_{k^*}^{(m)}\|_2^2) \\
&= R^{(m)} - X_{k^*}^{(m)}{X_{k^*}^{(m)}}^T R^{(m)} / \|X_{k^*}^{(m)}\|_2^2 = R^{(m+1)},
\end{aligned}$$

which shows (iv) at step $m+1$, thus completing the induction.

6. Show that if W is mean-zero and sub-Gaussian with parameter σ, then $\mathrm{Var}(W) \le \sigma^2$.
Solution: We know that $\mathbb{E}(e^{\alpha W}) \le e^{\alpha^2\sigma^2/2}$ for all $\alpha \in \mathbb{R}$. A power series expansion both sides of the above inequality yields

$$1 + \alpha\mathbb{E}(W) + \alpha^2\mathbb{E}(W^2)/2 + \alpha^2\mathbb{E}\left(\sum_{r=3}^{\infty} \frac{\alpha^{r-2}W^r}{r!}\right) \le 1 + \alpha^2\sigma^2/2 + O(\alpha^4).$$

Subtracting 1 and dividing by α^2 gives

$$\mathbb{E}(W^2)/2 + \mathbb{E}\left(\sum_{r=2}^{\infty} \frac{\alpha^{r-2} W^r}{r!}\right) \leq \sigma^2/2 + O(\alpha^2).$$

Now $e^W + e^{-W} \geq e^{|W|}$ so $e^{|W|}$ is integrable and $e^{|W|} \geq \sum_{r=3}^{\infty} \frac{\alpha^{r-2} W^r}{r!}$ for all $\alpha < 1$. Thus by dominated convergence theorem, $\mathbb{E}(\sum_{r=3}^{\infty} \frac{\alpha^{r-2} W^r}{r!}) \to 0$ as $\alpha \to 0$.

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 $1/2$.
Solution:

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

8. (a) Let $W \sim \chi^2_d$. 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^2_1 random variable is $1/\sqrt{1-2\alpha}$ for $\alpha < 1/2$, and $e^{-\alpha}/\sqrt{1-2\alpha} \leq e^{2\alpha^2}$ when $|\alpha| < 1/4$.
Solution: Note that W is the sum of d independent χ^2_1-distributed random variables. Thus

$$\mathbb{E}e^{\alpha(W-d)} = (1-2\alpha)^{-d/2} e^{-\alpha d}$$

for $\alpha < 1/2$. Suppose $t \in (0, 1)$. Using the Chernoff bound, we get

$$\begin{aligned}
\mathbb{P}(W - d \geq dt) &\leq \inf_{0<\alpha<1/2} \left(\frac{e^{-\alpha}}{\sqrt{1-2\alpha}} e^{-\alpha t}\right)^d \\
&\leq \inf_{0<\alpha<1/4} \exp\{d(2\alpha^2 - \alpha t)\} \quad \text{using the hint,} \\
&= e^{-dt^2/8}
\end{aligned}$$

setting $\alpha = t/4$ in the last line, which is permitted since $t < 1$. The argument to bound $\mathbb{P}(d - W \geq dt)$ is similar, and the result follows from a union bound.

(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

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

Solution: Let a_i be the ith row of A. Then $a_i^T u/\|u\|_2 \sim N(0,1)$ and thus by independence of the rows, $\|Au\|_2^2/\|u\|_2^2$ has a χ^2_d distribution. Now use the answer from part (a).

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

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

This is the famous Johnson–Lindenstrauss Lemma.
Solution: Apply the result from (b) with $u = u_i - u_j$. Using the union bound, we get

$$\mathbb{P}\left(\left|\frac{\|w_i - w_j\|_2^2}{\|u_i - u_j\|_2^2} - 1\right| \geq t\right) \leq \binom{n}{2} \times 2e^{-dt^2/8} \leq \exp\{-dt^2/8 + 2\log(n)\} \leq \epsilon.$$

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.

9. Show that any two Lasso solutions when $\lambda > 0$ must have the same ℓ_1-norm.
Solution: We know that for any two Lasso solutions $\hat{\beta}^{(1)}$, $\hat{\beta}^{(2)}$ we have that $X\hat{\beta}^{(1)} = X\hat{\beta}^{(2)}$. Thus the least squares part of their Lasso objectives will be equal. Since the two Lasso objectives must be equal (they are both Lasso solutions), the ℓ_1 parts must also be the same.

10. A *convex combination* of a set of points $S = \{v_1, \ldots, v_m\} \subseteq \mathbb{R}^{d'}$ is any point of the form

$$\alpha_1 v_1 + \cdots + \alpha_m v_m,$$

where $\alpha_j \in \mathbb{R}$ and $\alpha_j \geq 0$ for $j = 1, \ldots, m$, and $\sum_{j=1}^m \alpha_j = 1$. Carathéodory's Lemma states that if S 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 i.e. there exist $j_1, \ldots, j_{d+1} \in \{1, \ldots, m\}$ and non-negative reals $\alpha_1, \ldots, \alpha_{d+1}$ summing to 1 with

$$v = \alpha_1 v_{j_1} + \cdots + \alpha_{d+1} v_{j_{d+1}}.$$

With this knowledge, show that for any value of λ, there is always a Lasso solution with no more than n non-zero coefficients.
Solution: As the columns of X are centred, they live in an $n-1$-dimensional subspace of $\mathbb{R}^n$. $X\hat{\beta}_\lambda^L$ is a convex combination of points in

$$\|\hat{\beta}_\lambda^{\mathrm{L}}\|_1\{\pm X_1, \ldots, \pm X_p\} \subset \mathbb{R}^n.$$

By Carathéodory's Lemma, there exist $1 \leq k_1, \ldots, k_n \leq p$ and an $\alpha \in \mathbb{R}^n$ with $\|\alpha\|_1 = 1$ such that

$$X\hat{\beta}_\lambda^{\mathrm{L}} = \|\hat{\beta}_\lambda^{\mathrm{L}}\|_1 \sum_{j=1}^n \alpha_j X_{k_j}.$$

Since $\|\hat{\beta}_\lambda^{\mathrm{L}}\|_1 \|\alpha\|_1 = \|\hat{\beta}_\lambda^{\mathrm{L}}\|_1$, the expression on the RHS must also constitute a Lasso solution with only n non-zero components.

11. Show that if $\lambda \geq \lambda_{\max} := \|X^T Y\|_\infty / n$, then $\hat{\beta}_\lambda^{\mathrm{L}} = 0$.
Solution: We need only check that $\hat{\beta}_\lambda^{\mathrm{L}} = 0$ satisfies the KKT conditions and this is clear. The solution is unique by the answer to question 9.

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 kth component of the Lasso estimator is given by

$$\hat{\beta}_{\lambda,k}^L = (|\hat{\beta}_k^{\mathrm{OLS}}| - \lambda)_+ \mathrm{sgn}(\hat{\beta}_k^{\mathrm{OLS}})$$

where $(\cdot)_+ = \max(0, \cdot)$. What is the corresponding estimator if the ℓ_1 penalty $\|\beta\|_1$ in the Lasso objective is replaced by the ℓ_0 penalty $\|\beta\|_0 := |\{k : \beta_k \neq 0\}|$?
Solution: Note that

$$\frac{1}{2n}\|Y - X\beta\|_2^2 = \sum_{k=1}^{p} \frac{1}{2}(\hat{\beta}_k^{\mathrm{OLS}} - \beta_k)^2 + \frac{1}{2n}\|Y - X\hat{\beta}^{\mathrm{OLS}}\|_2^2.$$

Thus for the first part we need to find the minimiser of

$$\frac{1}{2}(\hat{\beta}_k^{\mathrm{OLS}} - \beta_k)^2 + \lambda|\beta_k|.$$

We write $\hat{\beta}$ for $\hat{\beta}_\lambda^{\mathrm{L}}$ for simplicity. Note by question 9, $|\hat{\beta}_k|$ is unique. By the KKT conditions,

$$\hat{\beta}_k^{\mathrm{OLS}} - \hat{\beta}_k = \lambda\hat{\nu}_k$$

where $|\hat{\nu}_k| \leq 1$ and $\hat{\nu}_k = \mathrm{sgn}(\hat{\beta}_k)$ if $\hat{\beta}_k \neq 0$. Thus $\hat{\beta}_k = 0$ when $|\hat{\beta}_k^{\mathrm{OLS}}| \leq \lambda$. If $\hat{\beta}_k^{\mathrm{OLS}} > \lambda$, $\hat{\beta}_k = \hat{\beta}_k^{\mathrm{OLS}} - \lambda$; if $\hat{\beta}_k^{\mathrm{OLS}} < -\lambda$, $\hat{\beta}_k = \hat{\beta}_k^{\mathrm{OLS}} + \lambda$.

Now let $\hat{\beta}$ be the optimising β with the ℓ_0 penalty. Clearly when $(\hat{\beta}_k^{\mathrm{OLS}})^2/2 < \lambda$, $\hat{\beta}_k = 0$ is optimal. When $(\hat{\beta}_k^{\mathrm{OLS}})^2/2 = \lambda$, two solutions $\hat{\beta}_k = \hat{\beta}_k^{\mathrm{OLS}}$ or $\hat{\beta}_k = 0$ both minimise the objective. When $(\hat{\beta}_k^{\mathrm{OLS}})^2/2 > \lambda$ then $\hat{\beta}_k = \hat{\beta}_k^{\mathrm{OLS}}$.