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

1. Let $Y \in \mathbb{R}^n$ be a vector of responses, $\Phi \in \mathbb{R}^{n\times p}$ 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 $\alpha \in \mathbb{R}^n$.

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 $\mathbb{E}\{\hat{\phi}(x)\hat{\phi}(x')\} = (x^T x')^2$.

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).

4. Consider the logistic regression model where we assume $Y_1, \dots, Y_n \in \{-1, 1\}$ 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 $\beta \in \mathbb{R}^p$.

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 $M \subseteq \{1, \dots, p\}$ 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.*

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

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 $1/2$.

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 $|\alpha| < 1/4$.

(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) $u_1, \ldots, u_n \in \mathbb{R}^p$ (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.

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.

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.

11. Show that if $\lambda \geq \lambda_{\max} := \|X^T Y\|_\infty / n$, then $\hat{\beta}^{\mathrm{L}}_\lambda = 0$.

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

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

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 $\|\beta\|_0 := |\{k : \beta_k \neq 0\}|$?