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. □

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 $\mathcal{X} = \mathcal{P}(\{1, \ldots, p\})$ and $z, z' \in \mathcal{X}$. Let k be the Jaccard similarity kernel. Let π be a random permutation of $\{1, \ldots, p\}$. 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

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

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

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

and analogously

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

We can combine the above formulas as

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

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

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

and it also maximises the log-likelihood function

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

which is of course equivalent to minimising

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

□

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 replac eeach 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. content...

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

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

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

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

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

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

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

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 $(2k)! \ge 2^k k!$ for each $k \in \mathbb{N}$, that

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

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

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

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

for $t \in (0, 1)$. You may use the facts that the mgf of a χ^2_1 random variable is $(1 - 2\alpha)^{-1/2}$ for $\alpha < 1/2$, and $e^{-\alpha}(1 - 2\alpha)^{-1/2} \le e^{2\alpha^2}$ 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 $u_1, \ldots, u_n \in \mathbb{R}^p$, 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:

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

This is the famous Johnson-Lindenstrauss Lemma.

Proof. (a)

□