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. $\square$