Modern Statistical Methods — Example Sheet 1

Lucas Riedstra

November 2, 2020

Question 7. *Suppose we have a matrix of predictors $X \in \mathbb{R}^{n\times p}$ where $p \gg n$. Explain how to obtain the fitted values of the following ridge regression using the kernel trick:*

$$\text{Minimise over } \beta \in \mathbb{R}^p,\ \vartheta \in \mathbb{R}^{p(p-1)/2},\ \gamma \in \mathbb{R}^p,$$

$$\sum_{i=1}^n \left(Y_i - \sum_{k=1}^p X_{ik}\beta_k - \sum_{k=1}^p \sum_{j=1}^{k-1} X_{ik}X_{ij}\vartheta_{jk} - \sum_{k=1}^p X_{ik}^2 \gamma_k \right)^2 + \lambda_1 \|\beta\|^2 + \lambda_2 \|\vartheta\|^2 + \lambda_3 \|\gamma\|^2.$$

Note that we have indexed ϑ with two numbers for convenience.

Proof. Our linear model has predictors

$$\{X_{ik} \mid 1 \leq k \leq p\} \cup \{X_{ij}X_{ik} \mid 1 \leq j < k \leq p\} \cup \left\{X_{ik}^2 \mid 1 \leq k \leq p\right\}$$

for $i = 1, \ldots, k$. We want to assume that $\lambda_1 = \lambda_2 = \lambda_3$: to this end, define $\xi = \frac{\lambda_1}{\lambda_2}$ and $\eta = \frac{\lambda_3}{\lambda_1}$, and replace ϑ by $\xi\vartheta$ and γ by $\eta\gamma$. This means we will have to replace our predictors by

$$Z_i \coloneqq (X_{ik} \mid 1 \leq k \leq p)^\top \cup (\xi X_{ij}X_{ik} \mid 1 \leq j < k \leq p)^\top \cup (\gamma X_{ik}^2 \mid 1 \leq k \leq p)^\top$$

We use the kernel trick: note that

$$\begin{aligned} \langle Z_i, Z_j \rangle &= \sum_k X_{ik}X_{jk} + \xi^2 \sum_{k<\ell} X_{ik}X_{i\ell}X_{jk}X_{j\ell} + \gamma^2 \sum_k X_{ik}^2 X_{jk}^2 \\ &= \left(\frac{1}{\sqrt{2}\xi} + \frac{\xi}{\sqrt{2}} \sum_k X_{ik}X_{jk} \right)^2 + (\gamma^2 - \frac{\xi^2}{2}) \sum_k X_{ik}^2 X_{jk}^2 - \frac{1}{2\xi^2}. \end{aligned}$$

We have therefore rewritten $\langle Z_i, Z_j \rangle$ into something that can be computed in $O(p)$, which is exactly the aim of the kernel trick, since this allows computation of the kernel matrix K in $O(n^2 p)$. $\square$