Statistical Methods 2

Homework – Kernel methods for regression

Submit your work on Blackboard by 4pm on Friday 17/03/2023

Let $\{(y_i^0, x_i^0)\}_{i=1}^n$ be n observations such that $y_i^0 \in \mathbb{S}$ and $x_i^0 \in \mathbb{R}^p$ for all $i \in \{1, \dots, n\}$ and for some $p \geq 1$. Below we assume the following model for the observations:

$$Y_i^0 \sim f(y; \mu_i, \phi)\mathrm{d}y, \quad g(\mu_i) = \alpha + f(x_i^0), \quad i = 1, \dots n, \quad \alpha \in \mathbb{R}, \quad \phi \in (0, \infty), \quad f \in \mathcal{F} \tag{1}$$

where $\mathcal{F} \subset \{f : \mathbb{R}^p \to \mathbb{R}\}$ and where, for all $\mu \in \mathbb{R}$ and $\phi \in (0, \infty)$, the distribution $f(y; \mu, \phi)\mathrm{d}y$ belongs to an exponential family of distributions.

Remark: For $\mathcal{F} = \{\sum_{j=1}^p f_j,\ f_j \in \mathcal{C}^2(\mathbb{R}), \forall j\}$ the model (1) reduces to the generalized additive model discussed in Chapter 10.

Let $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ be a positive semi-definite kernel (see Chapter 4, Definition 4.2) and recall that, by the Moore-Aronszajn theorem, there exists a unique reproducing kernel Hilbert space (RKHS) $(\mathcal{H}_k, < \cdot, \cdot >_k)$ for which k is the reproducing kernel. Here, $\mathcal{H}_k$ is a vector space of functions and $< \cdot, \cdot >_k$ is an inner product on $\mathcal{H}_k$, so that $(\mathcal{H}_k, < \cdot, \cdot >_k)$ is a Hilbert space. In addition, the space $(\mathcal{H}_k, < \cdot, \cdot >_k)$ being an RKHS it follows that

$$f(x) =< f, k(x, \cdot) >_k, \quad \forall x \in \mathcal{X},\ \forall f \in \mathcal{H}_k \quad \text{[Reproducing property]}.$$

Below we denote by $\| \cdot \|_k$ the norm on $\mathcal{H}_k$ induced by the inner product $< \cdot, \cdot >_k$, i.e. $\| \cdot \|_k$ is defined by

$$\|f\|_k = \sqrt{< f, f >_k}, \quad f \in \mathcal{H}_k.$$

In this assignment we consider the model (1) with $\mathcal{F} = \mathcal{H}_k$ and, for a given $\lambda \in (0, \infty)$, we want to estimate model (1) using:

$$\big(\hat{\alpha}_\lambda, \hat{\phi}_\lambda, \hat{f}_\lambda\big) \in \mathrm{argmax}_{\alpha \in \mathbb{R},\, \phi \in (0,\infty),\, f \in \mathcal{H}_k} \frac{1}{2n} \sum_{i=1}^n \log f\Big(y_i; g^{-1}\big(\alpha + f(x_i^0)\big), \phi\Big) - \lambda \|f\|_{\mathcal{H}_k}^2. \tag{2}$$

Answer the following questions:

Part 1

1. **(5 marks)** Give an example of kernel for which the model (1) is identifiable, and an example of kernel for which the model (1) is not identifiable.

2. **(10 marks)** Let
$$\tilde{\mathcal{H}}_n = \mathrm{span}\big\{k(x_1^0, \cdot), \dots, k(x_n^0, \cdot)\big\}$$
and assume that we can write $\mathcal{H}_k$ as $\mathcal{H}_k = \tilde{\mathcal{H}}_n \oplus \tilde{\mathcal{H}}_n^\perp$, that is that for all $f \in \mathcal{H}_k$ there exist a function $f_1 \in \tilde{\mathcal{H}}_n$ and $f_2 \in \tilde{\mathcal{H}}_n^\perp$ such that $f(x) = f_1(x) + f_2(x)$ for all $x \in \mathbb{R}^p$, and we have
$$< f_1, f_2 >_k = 0, \quad \forall f_1 \in \tilde{\mathcal{H}}_n, \quad f_2 \in \tilde{\mathcal{H}}_n^\perp.$$

Show that, for any $\lambda > 0$, if the function $\hat{f}_\lambda$ is as in (2) then

$$\hat{f}_\lambda = \sum_{i=1}^{n} \hat{\beta}_{\lambda,i} k(x_i^0, \cdot) \tag{3}$$

for some $\hat{\beta}_\lambda \in \mathbb{R}^n$.

3. **(5 marks)** Letting $\hat{\beta}_\lambda$ be as defined in (3), write down the optimization problem that needs to be solved in order to compute $\hat{\gamma}_\lambda := \left(\hat{\alpha}_\lambda, \hat{\phi}_\lambda, \hat{\beta}_\lambda\right)$ for a given $\lambda > 0$.

4. **(10 marks)** Computing, for a given $\lambda > 0$, the parameter $\hat{\gamma}_\lambda$ defined in Question 3 requires to solve an $n+2$ dimensional optimization problem, which can be computationally expensive in practice. Given an integer $m < n+2$, explain how the Nyström method can be used to compute an approximate solution to (2) which only requires to solve an m-dimensional optimization problem. Write down the m-dimensional optimization problem that needs to be solved in order to compute this approximate solution to (2).

5. **(10 marks)** Reformulate the m-dimensional optimization defined in Question 4 in such a way that the `R` package `glmnet` can be used to solve it.

Part 2

Choose a dataset with $p \geq 2$ variables on which you can fit model (1) with $f(y;\mu,\phi)\mathrm{d}y \neq \mathcal{N}_1(\mu,\phi)$. For instance, you can use the `wesdr` dataset [1] in which case $y_i^0 \in \{0,1\}$ for all $i \in \{1,\ldots,n\}$. Choose a collection $\{\tilde{k}_c,\, c \in \mathcal{C}\}$ of kernels so that model (1) is identifiable for all $c \in \mathcal{C}$, and split the dataset into a training and a test set.

6. **(10 marks)** Using your answer to Question 5 and the training set, write an `R` function that computes, for any $\lambda > 0$, $c \in \mathcal{C}$ and integer $m \leq n+2$, the approximate solution to (2) defined in Question 4.

7. **(10 marks)** When running the code of Question 6, do you obtain some numerical errors for some values of (λ, c, m)? If yes propose a solution to fix the problem.

8. **(10 marks)** As for the penalized regression methods studied in Chapters 6-10, the performance of kernel regression methods depends crucially on the choice of λ.

   Using your answer to Question 5 and the training set, write an `R` function that computes, for any $c \in \mathcal{C}$ and integer $m \leq n+2$, the approximate solution to (2) defined in Question 4 obtained when λ is chosen using cross-validation. Clearly mention which cross-validation procedure you use.

9. **(10 marks)** The performance of kernel regression depends also crucially on the choice of the kernel k, which determines the space $\mathcal{H}_k$ the estimated function $\hat{f}_\lambda$ belongs to.

   Using your answer to Question 5 and the training set, write an `R` function that computes, for any integer $m \leq n+2$, the approximate solution to (2) defined in Question 4 obtained when λ and c are chosen using cross-validation. Clearly mention which cross-validation procedure you use.

[1] available from the `R` package `gss`.

10. **(10 marks)** Using the code you wrote in Question 9, fit the model for different values of m and, for each fitted model, predict the response variable for the observations in the test set and compute the prediction error. Comment on your results.

11. **(10 marks)** Compute the "GAM" estimate of the model (1) using the `R` package `mcv` and cross-validation for choosing the penalty parameters. Using this model, predict the response variable for the observations in the test set and compute the prediction error. Compare the results you obtain with those obtained in Questions 10.