EPFL, Autumn 2024
Yoav Zemel
Guillaume Obozinski
Ho Yun
Shivang Sachar

Statistical Machine Learning

Exercise sheet 9

Exercise 9.1 (Linear kernel) Consider the function $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ defined by $K(\mathbf{x}, \mathbf{y}) \mapsto \mathbf{x}^\top \mathbf{y}$.

(a) Show that K is a symmetric positive-definite function, and by Aronszajn's theorem, a reproducing kernel.

(b) Let $\mathcal{H}$ be the RKHS with reproducing kernel K defined above. Show that $f \in \mathcal{H}$ if and only if f is a linear function, that is, there exists $\tilde{\mathbf{f}} \in \mathbb{R}^p$ such that $f(\mathbf{x}) = \mathbf{x}^\top \tilde{\mathbf{f}} = K(\mathbf{x}, \tilde{\mathbf{f}})$.

(Hint: One direction is very easy. For the other, you can first show that all the functions $K(\mathbf{x}, \cdot)$ live a finite dimensional space and therefore have a canonical basis on which can we decompose them, and then use the kernel reproducing property to extend this to all functions in $\mathcal{H}$).

(c) Using only elementary linear algebra (that is, without using any facts about reproducing kernels), show that $\mathcal{H}$ forms a Hilbert space under the inner product $\langle K(\cdot, \mathbf{x}), K(\cdot, \mathbf{y}) \rangle = K(\mathbf{x}, \mathbf{y})$.

Exercise 9.2 (Squared loss regression in RKHS) Let $\mathcal{H}$ denote the RKHS associated to a Mercer kernel K.

(a) *Preliminary questions*

(i) Let $\mathbf{K}$ be a positive semi definite matrix. Show that $\mathbf{K}$ and $(\mathbf{K} + \lambda \mathbf{I})^{-1}$ commute.

(ii) Deduce from the previous question that if $\mathbf{h} \in \ker(\mathbf{K})$ then so does $(\mathbf{K} + \lambda \mathbf{I})^{-1}\mathbf{h}$.

(iii) Let $\mathbf{K} = \big(K(\mathbf{x}_i, \mathbf{x}_j)\big)_{1 \le i,j \le n}$ with K the above defined Mercer kernel. Show that if $\mathbf{h} \in \ker(\mathbf{K})$ then the function $\sum_{i=1}^n h_i K(x_i, \cdot)$ is constant and equal to 0.

(b) Show that the solution to the regression problem

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \{y_i - f(\mathbf{x}_i)\}^2 + \lambda \|f\|_{\mathcal{H}}^2$$

is $\widehat{f}(\mathbf{x}) = \sum_{i=1}^n \widehat{\alpha}_i K(\mathbf{x}, \mathbf{x}_i)$ with $\widehat{\boldsymbol{\alpha}} = (\mathbf{K} + n\lambda \mathbf{I})^{-1} \boldsymbol{y}$, where $\mathbf{K}$ is the Gram matrix associated to K.

(c) Using the above result show that the solution to the ridge regression problem with no intercept,

$$\widehat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}} \frac{1}{n} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_2^2,$$

where $\boldsymbol{y} \in \mathbb{R}^n$ and the design matrix $\mathbf{X}$ is $n \times p$ is given by $\widehat{\boldsymbol{\beta}} = \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top + n\lambda \mathbf{I})^{-1} \boldsymbol{y}$.

Exercise 9.3 (Ridge regression and kernel trick) Consider again, the ridge regression problem with no intercept,

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{n}\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_2^2,$$

where $\boldsymbol{y} \in \mathbb{R}^n$ and the design matrix $\mathbf{X}$ is $n \times p$.

(a) Using what you know about ridge regression and the identity, $(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)\mathbf{X}^\top = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_n)$, show that $\widehat{\boldsymbol{\beta}} = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + n\lambda\mathbf{I}_n)^{-1}\boldsymbol{y}$ as in the previous problem.

(b) Thus, there are two methods for computing $\widehat{\boldsymbol{\beta}}$: $\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + n\lambda\mathbf{I}_n)^{-1}\boldsymbol{y}$ and $(\mathbf{X}^\top\mathbf{X} + n\lambda\mathbf{I}_n)^{-1}\mathbf{X}^\top\boldsymbol{y}$. What is the computational complexity of applying each method? When should one be favored over the other?

Exercise 9.4 (RKHS and the representer theorem) Suppose that K has an eigen-expansion

$$K(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{\infty} \gamma_j \phi_j(\mathbf{x})\phi_j(\mathbf{y}), \tag{1}$$

where $\gamma_j \geq 0$ are eigenvalues that satisfy $\sum_{j=1}^{\infty} |\gamma_j|^2 < \infty$ and $\{\phi_j\}_{j=1}^{\infty}$ forms the orthogonal basis of the function space $\mathcal{H}$. The space $\mathcal{H}$ has the form

$$\mathcal{H} = \left\{ f : \mathbb{R}^p \to \mathbb{R} : f(\mathbf{x}) = \sum_{i=1}^{\infty} c_i\phi_i(\mathbf{x}) \text{ for all } \mathbf{x} \text{ and } \sum_{i=1}^{\infty} c_i^2/\gamma_i < \infty \right\}$$

For $f(\mathbf{x}) = \sum_{i=1}^{\infty} c_i\phi_i(\mathbf{x})$ and $g(\mathbf{x}) = \sum_{i=1}^{\infty} d_i\phi_i(\mathbf{x})$ in $\mathcal{H}$, define

$$\langle f, g\rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} \frac{c_i d_i}{\gamma_i}.$$

NOTE: In the following problems, do not use any results about reproducing kernels.

(a) Show that $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is an inner product.

(b) For any $f \in \mathcal{H}$ and $\mathbf{x} \in \mathbb{R}^p$, show that $\langle K(\cdot, \mathbf{x}), f\rangle_{\mathcal{H}} = f(\mathbf{x})$.

(c) For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$, show that $\langle K(\cdot, \mathbf{x}), K(\cdot, \mathbf{y})\rangle_{\mathcal{H}} = K(\mathbf{x}, \mathbf{y})$.

(d) If $f(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i K(\mathbf{x}, \mathbf{x}_i)$ and $g(\mathbf{x}) = \sum_{j=1}^{k} \beta_j K(\mathbf{x}, \mathbf{x}_j)$, show that

$$\langle f, g\rangle_{\mathcal{H}} = \sum_{i=1}^{m}\sum_{j=1}^{k} \alpha_i\beta_j K(\mathbf{x}_i, \mathbf{x}_j)$$

and in particular,

$$\|f\|_{\mathcal{H}}^2 = \sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j K(\mathbf{x}_i, \mathbf{x}_j).$$