KERNELS

1 Warm-up

Let $\mathcal{H}$ be a RKHS associated with a positive definite kernel $k : \mathsf{X} \times \mathsf{X} \to \mathbb{R}$.

1. Prove that for all $(x, y) \in \mathsf{X} \times \mathsf{X}$ and $f \in \mathcal{H}$,

$$|f(x) - f(y)| \leqslant \|f\|_{\mathcal{H}} \|k(x, \cdot) - k(y, \cdot)\|_{\mathcal{H}} \,.$$

The proof follows from Cauchy-Schwarz inequality as, for all $(x, y) \in \mathsf{X}^2$,

$$|f(x) - f(y)| = |\langle f, k(x, \cdot)\rangle_{\mathcal{H}} - \langle f, k(x, \cdot)\rangle_{\mathcal{H}}| = |\langle f, k(x, \cdot) - k(y, \cdot)\rangle_{\mathcal{H}}| \,.$$

2. Prove that the kernel k associated with $\mathcal{H}$ is unique, i.e. if $\widetilde{k}$ is another potitive definite kernel satisfying the RKHS properties for $\mathcal{H}$, then $k = \widetilde{k}$.

Write, for all $x \in \mathsf{X}$,

$$\|k(x,\cdot) - \widetilde{k}(x,\cdot)\|^2_{\mathcal{H}} = \langle k(x,\cdot) - \widetilde{k}(x,\cdot), k(x,\cdot) - \widetilde{k}(x,\cdot)\rangle = k(x,x) - \widetilde{k}(x,x) + \widetilde{k}(x,x) - k(x,x) = 0 \,.$$

Thus, by Cauchy-Schwarz and the reproducing property, for all $x, y \in \mathsf{X}$,

$$|k(x,y) - \widetilde{k}(x,y)| = |\langle k(x,\cdot), k(y,\cdot)\rangle_{\mathcal{H}} - \langle \widetilde{k}(x,\cdot), k(y,\cdot)\rangle_{\mathcal{H}}| \leq \|k(x,\cdot) - \widetilde{k}(x,\cdot)\|_{\mathcal{H}} \|k(y,\cdot)\|_{\mathcal{H}} = 0.$$

3. Prove that for all $x \in \mathsf{X}$, the function defined on $\mathcal{H}$ by $\delta_x : f \mapsto f(x)$ is continuous.

Let $x \in \mathsf{X}$. Let $f \in \mathcal{H}$ and $(f_n)_n \subset \mathcal{H}$ such that $\|f_n - f\|_{\mathcal{H}} \to 0$. Then, by Cauchy-Schwarz,

$$|\delta_x(f) - \delta_x(f_n)| = |f(x) - f_n(x)| = |\langle f_n - f, k(x,\cdot)\rangle_{\mathcal{H}}| \leq \|f_n - f\|_{\mathcal{H}} \|k(x,\cdot)\|_{\mathcal{H}} \to 0.$$

2 Kernel Ridge regression

Let $\mathcal{H}$ be a RKHS on $\mathcal{X}$ with kernel k. We consider the regression model $Y_i = f^*(X_i) + \xi_i$, $i \in \{1, \ldots, n\}$, with ξ_i, $1 \leq i \leq n$, independent centered noise with finite variance. The unknown function f^* is estimated by the solution $\widehat{f}$ of the convex minimization problem

$$\widehat{f} = \underset{f \in \mathcal{H}}{\mathrm{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \frac{\lambda}{n} \|f\|^2_{\mathcal{H}} \right\} ,$$

with $\lambda > 0$.

2.1 Solving Kernel ridge regression

1. Check that $\widehat{f} : x \mapsto \sum_{j=1}^{n} \widehat{\beta}_j k(X_j, x)$ where $\widehat{\beta} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_n)^\top$ is solution to

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^n}{\mathrm{argmin}} \left\{ \|Y - K\beta\|^2 + \lambda \beta^\top K \beta \right\}$$

with K defined by $K = (k(X_i, X_j))_{1 \leq i,j \leq n}$. Comment on this result.

There exists β such that, for all x,

$$\widehat{f}(x) = \sum_{j=1}^{n} \beta_j k(X_j, x).$$

This yields

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2 + \frac{\lambda}{n}\|f\|_{\mathcal{H}}^2 &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{n}\beta_j k(X_j, X_i))^2 \\ &\quad + \frac{\lambda}{n}\langle\sum_{j=1}^{n}\beta_j k(X_j, \cdot), \sum_{i=1}^{n}\beta_i k(X_i, \cdot)\rangle,\end{aligned}$$

which gives the result, since

$$\langle\sum_{j=1}^{n}\beta_j k(X_j, \cdot), \sum_{i=1}^{n}\beta_i k(X_i, \cdot)\rangle = \sum_{i,j=1}^{n}\beta_i\beta_j k(X_i, X_j) = \beta^\top K\beta.$$

2. Assume that K is non-singular. Give an explicit expression for $\widehat{\beta}$.

   Write, for all β,

   $$L(\beta) = \|Y - K\beta\|_2^2 + \lambda\beta^\top K\beta.$$

   The gradient of L is then given by

   $$\begin{aligned}\nabla L(\beta) &= -2K^\top(Y - K\beta) + \lambda(K\beta + K^\top\beta) \\ &= -2K(Y - K\beta) + 2\lambda K\beta.\end{aligned}$$

   The minimum $\widehat{\beta}$ of L satisfies

   $$\begin{aligned}&\Leftrightarrow -2K(Y - K\widehat{\beta}) + 2\lambda K\widehat{\beta} = 0 \\ &\Leftrightarrow \widehat{\beta} = (K + \lambda I)^{-1}Y.\end{aligned}$$

2.2 Bias and variance

We assume that $f^* \in \mathcal{H}$ and we write

$$f_V^* : x \mapsto \sum_{i=1}^{n} \beta_i^* k(X_i, x)$$

for the projection of f^* onto the linear span $V = \text{span}\{k(X_i, .) : i = 1, \dots, n\}$, with respect to the Hilbert norm $\|\cdot\|_{\mathcal{H}}$. We write $K = \sum_{i=1}^{n}\lambda_i u_i u_i^\top$ for an eigenvalue decomposition of K.

1. Check that

   $$K\widehat{\beta} = \sum_{i=1}^{n}\frac{\lambda_i}{\lambda_i + \lambda}\langle Y, u_i\rangle u_i \quad \text{with} \quad Y = (Y_1, \dots, Y_n)^\top.$$

Since $(u_i)_{1\leq i\leq n}$ is an orthonormal basis of $\mathbb{R}^n$,

$$\begin{aligned}
K\widehat{\beta} &= \sum_{i=1}^n \langle K\widehat{\beta}, u_i\rangle u_i \\
&= \sum_{i=1}^n \langle K(K+\lambda I)^{-1}Y, u_i\rangle u_i \\
&= \sum_{i=1}^n \langle Y, (K+\lambda I)^{-1}Ku_i\rangle u_i \\
&= \sum_{i=1}^n \frac{\lambda_i}{\lambda+\lambda_i}\langle Y, u_i\rangle u_i .
\end{aligned}$$

2. Check that
$$\|\mathbb{E}[K\widehat{\beta}] - K\beta^*\|_2^2 = \sum_{i=1}^n \left(\frac{\lambda\lambda_i}{\lambda_i+\lambda}\right)^2 \langle \beta^*, u_i\rangle^2 .$$

First, note that, for all $1 \leq i \leq n$, $f^*(X_i) = f_V^*(X_i) = K_i^\top \beta^*$ and

$$\langle \mathbb{E}[Y], u_i\rangle = \langle K\beta^*, u_i\rangle = \langle \beta^*, Ku_i\rangle = \lambda_i\langle \beta^*, u_i\rangle .$$

Consequently,

$$\begin{aligned}
\|\mathbb{E}[K\widehat{\beta}] - K\beta^*\|^2 &= \left\|\sum_{i=1}^n \frac{\lambda_i}{\lambda_i+\lambda}\langle \mathbb{E}[Y], u_i\rangle u_i - \sum_{i=1}^n \langle K\beta^*, u_i\rangle u_i\right\|_2^2 \\
&= \left\|\sum_{i=1}^n \left(\frac{\lambda_i^2}{\lambda_i+\lambda} - \lambda_i\right)\langle \beta^*, u_i\rangle u_i\right\|_2^2 \\
&= \sum_{i=1}^n \left(\frac{\lambda\lambda_i}{\lambda_i+\lambda}\right)^2 \langle \beta^*, u_i\rangle^2 .
\end{aligned}$$

3. We assume henceforth that the $\xi_i = Y_i - f^*(X_i)$, $i = 1, \ldots, n$, have a covariance $\mathbb{V}[\xi] = \sigma^2 I_n$. Check that the covariance matrix of $K\widehat{\beta}$ is equal to
$$\mathbb{V}[K\widehat{\beta}] = \sum_{i=1}^n \left(\frac{\lambda_i\sigma}{\lambda_i+\lambda}\right)^2 u_i u_i^\top .$$

Since $\widehat{\beta} = (K+\lambda I)^{-1}y$,

$$\begin{aligned}
\mathbb{V}[K\widehat{\beta}] &= K\mathbb{V}[(K+\lambda I)^{-1}Y]K^\top \\
&= K(K+\lambda I)^{-1}\mathbb{V}[Y](K+\lambda I)^{-1}K \\
&= \sigma^2 K^2 (K+\lambda I)^{-2} \\
&= \sum_{i=1}^n \left(\frac{\lambda_i\sigma}{\lambda_i+\lambda}\right)^2 u_i u_i^\top ,
\end{aligned}$$

using the eigenvector decomposition of K.

4. We define $\|f\|_n^2 := \frac{1}{n}\sum_{i=1}^n f(X_i)^2$. Prove that
$$\mathbb{E}\left[\|\widehat{f} - f^*\|_n^2\right] = \frac{1}{n}\sum_{i=1}^n \left(\frac{\lambda_i}{\lambda+\lambda_i}\right)^2 \left(\lambda^2\langle \beta^*, u_i\rangle^2 + \sigma^2\right).$$

By the biais-variance tradeoff,

$$\mathbb{E}\left[\|\widehat{f}-f^*\|_n^2\right] = \frac{1}{n}\mathbb{E}\left[\|K\widehat{\beta}-K\beta^*\|^2\right] = \frac{1}{n}\mathrm{Tr}(\mathbb{V}(K\widehat{\beta})) + \frac{1}{n}\|\mathbb{E}[K\widehat{\beta}]-K\beta^*\|^2.$$