MAP534 Introduction to machine learning

Linear regression, penalization, kernel regression

Gaussian vectors

1. Let X be a Gaussian vector with mean $\mu \in \mathbb{R}^n$ and definite positive covariance matrix Σ. Pove that the characteristic function of X is given, for all $t \in \mathbb{R}^n$, by

$$\mathbb{E}[\mathrm{e}^{i\langle t\,;\,X\rangle}] = \mathrm{e}^{i\langle t\,;\,\mu\rangle - t^T\Sigma t/2}\,.$$

   Only requires to compute the mean and variance of the Gaussian random variable $\langle t\,;\,X\rangle$.

2. Let ε be a random variable in $\{-1,1\}$ such that $\mathbb{P}(\varepsilon = 1) = 1/2$. If $(X,Y)^T \sim \mathcal{N}(0, I_2)$ explain why the following vectors are or are not Gaussian vectors.

   (a) $(X, \varepsilon X)$.
   Not Gaussian since the probability that $X + \varepsilon X = 0$ is $1/2$.

   (b) $(X, \varepsilon Y)$.
   Gaussian since coordinates are independent Gaussian random variables.

   (c) $(X, \varepsilon X + Y)$.
   Not Gaussian since the characteristic function of $(1+\varepsilon)X + Y$ is not the Gaussian characteristic function.

   (d) $(X, X + \varepsilon Y)$.
   Gaussian as a linear transform of (b).

3. Let X be a Gaussian vector in $\mathbb{R}^n$ with mean $\mu \in \mathbb{R}^n$ and covariance matrix $\sigma^2 I_n$. Prove that the random variables $\bar{X}_n$ and $\widehat{\sigma}^2_n$ defined as

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \quad \text{and} \quad \widehat{\sigma}^2_n = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$$

   are independent.

   Left as an exercise.

Regression: prediction of a new observation

Consider the regression model given, for all $1 \leqslant i \leqslant n$, by

$$Y_i = X\beta_\star + \xi_i\,,$$

where $X \in \mathbb{R}^{n\times d}$ the $(\xi_i)_{1\leqslant i\leqslant n}$ are i.i.d. centered Gaussian random variables with variance $\sigma^2_\star$. Assume that $X^T X$ has full rank and that $\beta_\star$ and $\sigma^2_\star$ are estimated by

$$\widehat{\beta}_n = (X^T X)^{-1} X^T Y \quad \text{and} \quad \widehat{\sigma}^2_n = \frac{\|Y - X\widehat{\beta}_n\|^2}{n-d}\,.$$

Let $x_\star \in \mathbb{R}^d$ and assume that its associated observation $Y_\star = x_\star^T\beta_\star + \varepsilon_\star$ is predicted by $\widehat{Y}_\star = x_\star^T\widehat{\beta}_n$.

1. Provide the expression of $\mathbb{E}[(\widehat{Y}_\star - x_\star^T\beta_\star)^2]$?

   Correction soon.

2. Provide a confidence interval for $x_\star^T\beta_\star$ with statistical significiance $1-\alpha$ for $\alpha \in (0,1)$?

   Correction soon.

Kernels

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 $(x,y) \in \mathsf{X} \times \mathsf{X}$,

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

   The proof follows from Cauchy-Schwarz inequality and the fact that $f(x) = \langle f, k(x,\cdot)\rangle$.

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 $x \in \mathsf{X}$,

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

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

   Left as an exercise.

Penalized kernel regression

Consider the regression model given, for all $1 \leqslant i \leqslant n$, by

$$Y_i = f^*(X_i) + \xi_i\,,$$

where for all $1 \leqslant i \leqslant n$, $X_i \in \mathsf{X}$, and the $(\xi_i)_{1\leqslant i\leqslant n}$ are i.i.d. centered Gaussian random variables with variance σ^2. In this exercise, f^* is estimated by

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

with $\lambda > 0$ and $\mathcal{H}$ a RKHS on X with symmetric positive definite kernel k.

1. Check that $\widehat{f}(x) = \sum_{j=1}^n \widehat{\beta}_{n,j} k(X_j, x)$ where $\widehat{\beta}_n$ is solution to

$$\widehat{\beta}_n = \underset{\beta\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \|y - K\beta\|^2 + \lambda\beta^T K\beta \right\}\,,$$

   with K defined, for all $1 \leqslant i,j \leqslant n$, by $K_{i,j} = k(X_i, X_j)$. Provide the explicit expression of $\widehat{\beta}_n$ when K is nonsingular.

   First, we prove that $\widehat{f}$ belongs to $V = Span(k(x_i,\cdot), i = 1,\dots,n)$. Take $f \in \mathcal{H}$ and set $f = f_V + f_{V^\perp}$ where $f_V \in V$ and $f_{V^\perp} \in V^\perp$. Therefore

$$\frac{1}{n}\sum_{i=1}^n \Big(y_i - f(x_i)\Big)^2 + \frac{\lambda}{n}|f|_{\mathcal{H}}^2 = \frac{1}{n}\sum_{i=1}^n \Big(y_i - f_V(x_i)\Big)^2 + \frac{\lambda}{n}\Big(|f_V|_{\mathcal{H}}^2 + |f_{V^\perp}|_{\mathcal{H}}^2\Big),$$

since, by definition of $V^\perp$, for all $1 \le i \le n$,

$$f_{V^\perp}(x_i) = \langle f_{V^\perp}, k(x_i, \cdot)\rangle = 0.$$

Thus the initial optimization problem can be written as

$$\widehat{f} = \underset{f \in V}{argmin} \left\{ \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \frac{\lambda}{n}|f|_{\mathcal{H}}^2 \right\}. \tag{1}$$

In other words, there exist β_j such that, for all x,

$$\widehat{f}(x) = \sum_{j=1}^{n} \widehat{\beta}_j k(x_j, x).$$

Injecting this expression into (1), we get

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \frac{\lambda}{n}|f|_{\mathcal{H}}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{n}\beta_j k(x_j, x_i))^2 + \frac{\lambda}{n}\langle \sum_{j=1}^{n}\beta_j k(x_j, \cdot), \sum_{i=1}^{n}\beta_i k(x_i, \cdot)\rangle,$$

which gives the result, since

$$\langle \sum_{j=1}^{n}\beta_j k(x_j, \cdot), \sum_{i=1}^{n}\beta_i k(x_i, \cdot)\rangle = \sum_{i,j=1}^{n}\beta_i\beta_j k(x_i, x_j).$$

Let

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

The gradient of L is then given by

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

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. Check that
$$K\widehat{\beta}_n = \sum_{i=1}^{n}\frac{\lambda_i}{\lambda_i + \lambda}\langle Y_i, u_i\rangle u_i.$$

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

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

3. Prove that
$$\mathbb{V}[K\widehat{\beta}_n] = \sum_{i=1}^{n} \left(\frac{\lambda_i \sigma}{\lambda_i + \lambda} \right)^2 u_i u_i' .$$

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

$$\begin{aligned} \mathcal{C}(K\widehat{\beta}) &= K\mathcal{C}\left((K + \lambda I)^{-1} y\right) K' \\ &= K(K + \lambda I)^{-1}\mathcal{C}(y)(K + \lambda I)^{-1} K \\ &= \sigma^2 K^2 (K + \lambda I)^{-2} \\ &= \sum_{i=1}^{n} \left(\frac{\lambda_i \sigma}{\lambda_i + \lambda} \right)^2 u_i u_i^T , \end{aligned}$$

using the eigenvector decomposition of K.