Kernel Methods AMMI 2023
Final Exam

Juliette Marrie and Jean-Philippe Vert

July 2023

1 Dual of ridge regression (8.5 points)

Given $x_1, \ldots, x_n \in \mathbb{R}^p$ and $y_1, \ldots, y_n \in \mathbb{R}$, we consider the ridge regression problem

$$\min_{w \in \mathbb{R}^p} \quad \frac{1}{n} \sum_{i=1}^{n} (w^\top x_i - y_i)^2 + \lambda ||w||^2$$

Question 1: .... 2 points

(a) (.5 points) Rewrite this problem in matrix notation, with the matrix $X \in \mathbb{R}^{n \times p}$ and vector $y \in \mathbb{R}^n$ defined by $X = [x_1, ..., x_n]^\top$ and $y = (y_1, .., y_n)^\top$.

Solution:

$$\min_{w \in \mathbb{R}^p} \frac{1}{n}(Xw - y)^T (Xw - y) + \lambda w^T w$$

Note: Those who wrote wX instead of Xw did non get the points.

(b) (1.5 points) Find $w^\star \in \mathbb{R}^p$ that solves the problem.

Solution:

$$w = (X^T X + \lambda n I)^{-1} X^T y \quad \text{(proof in slides}$$

Question 2: .... 6 points

Let $u = Xw$. We rewrite the ridge regression problem as

$$\begin{aligned} \min_{u \in \mathbb{R}^n, w \in \mathbb{R}^p} \quad & \frac{1}{n} ||u - y||^2 + \lambda ||w||^2 \\ s.t. \quad & u = Xw \end{aligned}$$

(a) (.5 points) Write the Lagrangian of this primal problem.

Solution:

$$\mathcal{L}(w, u, \alpha) = \frac{1}{n} ||u - y||^2 + \lambda ||w||^2 + \alpha^T (u - Xw), \ \alpha \in \mathbb{R}^n$$

(b) (3 points) Show that the dual problem can be written as

$$\max_{\alpha \in \mathbb{R}^n} \quad -\frac{1}{4\lambda} \left[\alpha^T (X X^T + \lambda n I) \alpha - 4 \lambda \alpha^T y \right]$$

Solution: The dual problem is

$$\max_{\alpha \in \mathbb{R}^n} q(\alpha) \quad \text{with} \quad q(\alpha) = \inf_{w,u} \mathcal{L}(w, u, \alpha)$$

$$\nabla_w \mathcal{L} = 2\lambda w - X^T \alpha = 0 \implies w = \frac{X^T \alpha}{2\lambda}$$

$$\nabla_u \mathcal{L} = \frac{2}{n}(u - y) + \alpha = 0 \implies u = y - \frac{n\alpha}{2}$$

Therefore,

$$\begin{aligned} q(a) &= \frac{1}{n}\left|\left|y - \frac{n\alpha}{2} - y\right|\right|^2 + \lambda\left|\left|\frac{X^T\alpha}{2\lambda}\right|\right|^2 + \alpha^T\left(y - \frac{n\alpha}{2} - \frac{XX^T\alpha}{2\lambda}\right) \\ &= \frac{n}{4}\alpha^T\alpha + \frac{1}{4\lambda}\alpha^T XX^T\alpha + \alpha^T y - \frac{n}{2}\alpha^T\alpha - \frac{1}{2\lambda}\alpha^T XX^T\alpha \\ &= -\frac{n}{4}\alpha^T\alpha - \frac{1}{4\lambda}\alpha^T XX^T\alpha + \alpha^T y \\ &= -\frac{1}{4\lambda}\left[\alpha^T\left(XX^T + n\lambda I\right)\alpha - 4\lambda\alpha^T y\right] \end{aligned}$$

(c) (1 point) Find $\alpha^\star$ that solves this problem.

Solution:

$$\nabla_\alpha q = 2(XX^T + n\lambda I)\alpha - 4\lambda y = 0 \implies \alpha = 2\lambda(XX^T + n\lambda I)^{-1}y$$

(d) (.5 points) Deduce the value of $w^\star$ that solves the primal problem.

Solution:

$$w = \frac{X^T\alpha}{2\lambda} = X^T(XX^T + \lambda nI)^{-1}y$$

(e) (1.5 points) Show that solutions w^* found in Question 1(b) and in Question 2(d) are equal.

Solution: We want to show that

$$X^T(XX^T + \lambda nI)^{-1}y = (X^TX + \lambda nI)^{-1}X^Ty$$

We have

$$X^TXX^T + \lambda nI_pX^T = X^TXX^T + \lambda nX^TI_n$$

Factorizing by X^T,

$$(X^TX + \lambda nI_p)X^T = X^T(XX^T + \lambda nI_n)$$

Multiplying by $(X^TX + \lambda nI_p)^{-1}$ on the left and $(XX^T + \lambda nI_n)^{-1}$ on the right, on both sides of the equality, we get

$$X^T(XX^T + \lambda nI_n)^{-1} = (X^TX + \lambda nI_p)^{-1}X^T$$

2 Some p.d. kernels (3 points)

Show that the following kernels are p.d. (hint: write them as inner products):

a) (1.5 point) $K : \left\{ \begin{array}{ccc} \mathbb{R}^2 & \to & \mathbb{R} \\ (x, y) & \mapsto & K(x, y) = 2^{x+y} + 3^{x+y} \end{array} \right.$

Solution:

$$K(x, y) = \phi(x)^T \phi(y) \quad \text{with} \quad \phi(x) = (2^x, 3^x)^T$$

b) (1.5 point) $K : \left\{ \begin{array}{ccc} \mathbb{R}^2 & \to & \mathbb{R} \\ (x, y) & \mapsto & K(x, y) = \cos(x - y) \end{array} \right.$

Solution: We have $\cos(x - y) = \cos(x)\cos(y) + \sin(x)\sin(y)$, therefore

$$K(x, y) = \phi(x)^T \phi(y) \quad \text{with} \quad \phi(x) = (\cos(x), \sin(x))^T$$

3 Support Vector Regression (SVR) (5.5 pointw)

Given $x_1, \ldots, x_n \in \mathbb{R}^p$ and $y_1, \ldots, y_n \in \mathbb{R}$, linear SVR finds a linear model $f(x) = w^\top x + b$ with the maximum number of points such as the prediction $f(x_i)$ is within $\pm\epsilon$ of the output y_i. Samples with prediction error at least ϵ penalize the objective by $|f(x_i) - y_i| - \epsilon$.

The SVR problem can be formulated as the following optimization problem:

$$\begin{aligned} \min_{w \in \mathbb{R}^p, b \in \mathbb{R}, \xi \in \mathbb{R}^n} \quad & \frac{1}{2} w^\top w + C \sum_{i=1}^{n} \xi_i \\ s.t. \quad & y_i - w^\top x_i - b \leqslant \epsilon + \xi_i \\ & y_i - w^\top x_i - b \geqslant -\epsilon - \xi_i \\ & \xi_i \geqslant 0, \ i = 1, ..., n \end{aligned}$$

Rewrite this problem as a quadratic program, i.e., find z, P, q, G and h such that the problem can be formulated as

$$\begin{aligned} \min_{z} \quad & z^\top P z + q^\top z \\ s.t. \quad & Gz \leqslant h \end{aligned}$$

In other words,

- (.5 points) define the vector z and give its dimension
- (1 point) give the dimensions of P, q, G and h
- (1 + .5 + 1.5 + 1 = 4 points) give their value: you can write them as block vectors or block matrices, indicating the dimension of each block

Solution: We rewrite the SVR problrm in the QP form:

$$\begin{aligned} \min_{w\in\mathbb{R}^p, b\in\mathbb{R}, \xi\in\mathbb{R}^n} \quad & \frac{1}{2}w^\top w + C\sum_{i=1}^n \xi_i \\ s.t. \quad & -w^\top x_i - b - \xi_i \leqslant \epsilon - y_i \\ & w^\top x_i + b - \xi_i \leqslant \epsilon + y_i \\ & -\xi_i \leqslant 0,\ i = 1, ..., n \end{aligned}$$

We can define $z = (w_1, ..., w_p, b, \xi_1, ..., \xi_n)^T \in \mathbb{R}^{p+1+n}$.

Let $m = p + 1 + n$. The shapes of P, q, G and h are respectively (m, m), $(m,)$, $(3n, m)$ and $(3n,)$.

- $P = \text{diag}(1, ..., 1, 0, 0, ..., 0)$ with p ones and $1 + n$ zeros.
- $q = (0, ..., 0, 0, C, ..., C)^T$ with $p + 1$ zeros and n C.
- $G = \left(\begin{array}{c|c|ccc} & -1 & -1 & & \\ \textbf{-X} & \vdots & & \ddots & \\ & -1 & & & -1 \\ \hline & 1 & -1 & & \\ \textbf{X} & \vdots & & \ddots & \\ & 1 & & & -1 \\ \hline & 0 & -1 & & \\ \mathbf{0} & \vdots & & \ddots & \\ & 0 & & & -1 \end{array}\right)$
- $h = (\epsilon - y_1, ..., \epsilon - y_n, \epsilon + y_1, ..., \epsilon + y_n, 0, ..., 0)^T$.

4 Maximum Mean Discrepency (3 points)

Let $\mathcal{X}$ be a set and $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ a RKHS with reproducing kernel K.

Given n points $X = (x_1, ..., x_n) \in \mathcal{X}^n$ we define

$$\hat{\mu}_X = \frac{1}{n}\sum_{i=1}^n K_{x_i}\ ,$$

where for any $x \in \mathcal{X}$, $K_x \in \mathcal{H}$ denotes the function $t \in \mathcal{X} \mapsto K(x, t) \in \mathbb{R}$.

Question 1: .... 1.5 points

We consider two sets of points $X = (x_1, ..., x_n) \in \mathcal{X}^n$ and $Y = (y_1, ..., y_m) \in \mathcal{X}^m$. Express $||\hat{\mu}_X - \hat{\mu}_Y||^2_{\mathcal{H}}$ as a function of K.

Solution:

$$
\begin{aligned}
||\hat{\mu}_X - \hat{\mu}_Y||^2_{\mathcal{H}} &= \langle \hat{\mu}_X, \hat{\mu}_X \rangle_{\mathcal{H}} - 2\langle \hat{\mu}_X, \hat{\mu}_Y \rangle_{\mathcal{H}} + \langle \hat{\mu}_Y, \hat{\mu}_Y \rangle_{\mathcal{H}} \\
&= \left\langle \frac{1}{n}\sum_{i=1}^{n} K_{x_i}, \frac{1}{n}\sum_{j=1}^{n} K_{x_j} \right\rangle_{\mathcal{H}} - 2\left\langle \frac{1}{n}\sum_{i=1}^{n} K_{x_i}, \frac{1}{m}\sum_{j=1}^{m} K_{y_i} \right\rangle_{\mathcal{H}} + \left\langle \frac{1}{m}\sum_{i=1}^{m} K_{y_i}, \frac{1}{m}\sum_{j=1}^{m} K_{y_j} \right\rangle_{\mathcal{H}} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \langle K_{x_i}, K_{x_j} \rangle_{\mathcal{H}} - 2\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} \langle K_{x_i}, K_{y_j} \rangle_{\mathcal{H}} + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} \langle K_{y_i}, K_{y_j} \rangle_{\mathcal{H}} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K(x_i, x_j) - 2\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} K(x_i, y_j) + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} K(y_i, y_j)
\end{aligned}
$$

Question 2: .... 1.5 points

For any $f \in \mathcal{H}$, express $\langle f, \hat{\mu}_X \rangle_{\mathcal{H}}$ as a function of $f(x_1), ..., f(x_n)$.

Solution:

$$
\langle f, \hat{\mu}_X \rangle_{\mathcal{H}} = \left\langle f, \frac{1}{n}\sum_{i=1}^{n} K_{x_i} \right\rangle_{\mathcal{H}} = \frac{1}{n}\sum_{i=1}^{n} \langle f, K_{x_i} \rangle_{\mathcal{H}} = \sum_{i=1}^{n} f(x_i) \quad \text{since} \quad \langle f, K_x \rangle_{\mathcal{H}} = f(x)
$$

Note: f is not the solution of an optimization problem depending on $x_1, ..., x_n$. There is no supposed link between f and $x_1, ..., x_n$, therefore f does not necessarily lie the span of the K_{x_i}.

End of exam