Prof. Dr. Rudolf Mathar, Dr. Arash Behboodi, Emilio Balda

Exercise 12
- Proposed Solution -

Friday, February 2, 2018

Solution of Problem 1

From Mercer's Theorem we know that a kernel $K : \mathbb{R}^p \to \mathbb{R}$ is a valid kernel if and only if for any $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ the kernel matrix $\mathbf{K} = (K(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1,\dots,n}$ is non-negative definite.

a) For $K(\mathbf{x}, \mathbf{y}) = K_1(\mathbf{x}, \mathbf{y})$ we have that

$$\mathbf{K} = (K(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1,\dots,n} = (\alpha K_1(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1,\dots,n} = \alpha \mathbf{K}_1 \,.$$

Then, since $\mathbf{K}_1$ is non-negative definite and $\alpha > 0$, it holds

$$\mathbf{z}^T\mathbf{K}\mathbf{z} = \underbrace{\alpha}_{>0} \underbrace{z^T\mathbf{K}\mathbf{z}}_{\geq 0} \geq 0\,, \quad \forall \mathbf{z} \in \mathbb{R}^p \quad \Rightarrow \quad \mathbf{K}\text{is non-negative definite.}$$

b) For $K(\mathbf{x}, \mathbf{y}) = K_1(\mathbf{x}, \mathbf{y}) + K_2(\mathbf{x}, \mathbf{y})$ we have that $\mathbf{K} = \mathbf{K}_1 + \mathbf{K}_2$, thus

$$\mathbf{z}^T\mathbf{K}\mathbf{z} == \mathbf{z}^T(\mathbf{K}_1+\mathbf{K}_2)\mathbf{z} = \underbrace{z^T\mathbf{K}_1\mathbf{z}}_{\geq 0} + \underbrace{z^T\mathbf{K}_2\mathbf{z}}_{\geq 0} \geq 0\,, \quad \forall \mathbf{z} \in \mathbb{R}^p \Rightarrow \mathbf{K}\text{is non-negative definite.}$$

c) For $K(x, y) = K_1(x, y)K_2(x, y)$ we have that $\mathbf{K} = K_1 \odot \mathbf{K}2$, where $\odot$ denotes the-Hadamard product (i.e., point-wise multiplication) between two matrices. Let $\mathbf{K}_1 = \sum_{i=1}^n \lambda_i \mathbf{v}_i \mathbf{v}_i^T$ and $\mathbf{K}_2 = \sum_{i=1}^n \rho_i \mathbf{u}_i \mathbf{u}_i^T$ be the spectral decompositions of $\mathbf{K}_1$ and $\mathbf{K}_2$ respectively. Note that, since these matrices non-negative definite we have that their eigenvalues λ_i and ρ_i are non-negative for all $i = 1, \dots, n$. Therefore, we get

$$K = (\sum_{i=1}^n \lambda_i \mathbf{v}_i \mathbf{v}_i^T) \odot (\sum_{j=1}^n \rho_j \mathbf{u}_j \mathbf{u}_j^T) = \sum_{i=1}^n \sum_{j=1}^n \rho_i \lambda_j \mathbf{v}_i \mathbf{v}_i^T \odot \mathbf{u}_j \mathbf{u}_j^T = \sum_{i=1}^n \sum_{j=1}^n \rho_i \lambda_j (\mathbf{v}_i \odot \mathbf{u}_j)(\mathbf{v}_i \odot \mathbf{u}_j)^T \,.$$

Note that for any $i, j = 1, \dots, n$ the matrix $\rho_j \lambda_i (\mathbf{v}_i \odot \mathbf{u}_j)(\mathbf{v}_i \odot \mathbf{u}_j)^T$ is a rank-1 matrix with eigenvalue $\rho_j \lambda_i \geq 0$, thus a non-negative definite matrix. This means that $\mathbf{K}$ is a sum of non-negative definite matrices, therefore, as shown in **b)**, it is also non-negative definite.

Solution of Problem 2

(25P)

a) (3P) b is given as $-\frac{1}{2}\mathbf{a}^T(\mathbf{x}_1 + \mathbf{x}_2) = 3$.

b) (4P) Supporting vectors are those with $\lambda_i \neq 0$, which are $\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5$.

$\mathbf{x}_i$	y_i	λ_i	$\mathbf{x}_i$	y_i	λ_i
$\mathbf{x}_1 = \begin{pmatrix}1\\1\end{pmatrix}$	$y_1 = -1$	$\lambda_1 = 0$	$\mathbf{x}_4 = \begin{pmatrix}0.5\\-0.5\end{pmatrix}$	$y_4 = 1$	$\lambda_4 = 4.73$
$\mathbf{x}_2 = \begin{pmatrix}2\\0\end{pmatrix}$	$y_2 = -1$	$\lambda_2 = 0.67$	$\mathbf{x}_5 = \begin{pmatrix}-2\\1\end{pmatrix}$	$y_5 = 1$	$\lambda_5 = 0.94$
$\mathbf{x}_3 = \begin{pmatrix}0\\0\end{pmatrix}$	$y_3 = -1$	$\lambda_3 = 5$	$\mathbf{x}_6 = \begin{pmatrix}0\\-1\end{pmatrix}$	$y_6 = 1$	$\lambda_1 = 0$

c) (6P) For those vectors, the normal vector of the hyperplane is obtained as:

$$\mathbf{a} = \sum_{i=1}^{6} \lambda_i y_i \mathbf{x}_i = \lambda_2 y_2 \mathbf{x}_2 + \lambda_3 y_3 \mathbf{x}_3 + \lambda_4 y_4 \mathbf{x}_4 + \lambda_5 y_5 \mathbf{x}_5$$

$$\mathbf{a} = -0.67 \begin{pmatrix}2\\0\end{pmatrix} - 5 \begin{pmatrix}0\\0\end{pmatrix} + 4.73 \begin{pmatrix}0.5\\-0.5\end{pmatrix} + 0.94 \begin{pmatrix}-2\\1\end{pmatrix} = \begin{pmatrix}-0.86\\-1.43\end{pmatrix}$$

To find b, take two support vectors $\mathbf{x}_k$ and $\mathbf{x}_l$ with $y_k = 1$ and $y_l = -1$ with $0 < \lambda < 5$. For these support vectors, we have $y_i(\mathbf{a}^T \mathbf{x}_i + b) = 1$. Hence:

$$b^\star = \frac{-1}{2} \mathbf{a}^{\star T} (\mathbf{x}_k + \mathbf{x}_l) = -\frac{1}{2} \begin{pmatrix}-0.86 & -1.43\end{pmatrix} \left(\begin{pmatrix}2\\0\end{pmatrix} + \begin{pmatrix}-2\\1\end{pmatrix}\right) = \frac{1.43}{2} = 0.715. \quad (1)$$

d) (6P)

Suppose that a kernel is given by $K(\mathbf{x}, \mathbf{y}) = (2\mathbf{x}^T \mathbf{y} + 1)^2$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$. Write the kernel as

$$\begin{aligned} K(\mathbf{x}, \mathbf{y}) = (2\mathbf{x}^T \mathbf{y} + 1)^2 &= (2 \sum_{i=1}^{p} x_i y_i + 1)^2 \\ &= 4 \sum_{i=1}^{p} x_i^2 y_i^2 + 8 \sum_{1 \le i < j \le p} x_i x_j y_i y_j + 4 \sum_{i=1}^{p} x_i y_i + 1, \end{aligned}$$

therefore $\phi(\mathbf{x})$ can be written as:

$$\phi(\mathbf{x}) = (2x_1^2, \ldots, 2x_p^2, 2x_1, \ldots, 2x_p, 1, \sqrt{8}x_1x_2, \sqrt{8}x_1x_3, \ldots, \sqrt{8}x_{p-1}x_p).$$

The dimension of feature space is $p + p + 1 + \frac{p(p-1)}{2} = \frac{(p+1)(p+2)}{2}$.

e) (3P) The Kernel calssifer replaces the inner product in dual problem:

$$\begin{aligned} \max \quad & \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,j} y_i y_j \lambda_i \lambda_j K(\mathbf{x}_i, \mathbf{x}_j) \\ \text{s.t.} \quad & 0 \le \lambda_i \\ & \sum_{i=1}^{n} \lambda_i y_i = 0. \end{aligned}$$

For the proposed K we have:

$$\begin{aligned} \max \quad & \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,j} y_i y_j \lambda_i \lambda_j \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2) \\ \text{s.t.} \quad & 0 \le \lambda_i \\ & \sum_{i=1}^{n} \lambda_i y_i = 0. \end{aligned}$$

f) (3P) From this optimization problem, the vector $\phi(\mathbf{a})$ in the feature space is obtained as:

$$\phi(\mathbf{a}) = \sum_{i=1}^{n} \lambda_i y_i \phi(\mathbf{x}_i).$$

and for two vectors with $y_k = 1$ and $y_l = -1$ and $0 < \lambda$.

$$\begin{aligned}
b^\star &= \frac{-1}{2}\phi(\mathbf{a}^\star)^T(\phi(\mathbf{x}_k) + \phi(\mathbf{x}_l)) \\
&= \frac{-1}{2}(\sum_{i=1}^{n} \lambda_i y_i \phi(\mathbf{x}_i))^T(\phi(\mathbf{x}_k) + \phi(\mathbf{x}_l)) \\
&= \frac{-1}{2}\sum_{i=1}^{n} \lambda_i y_i (\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_k) + \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_l)) \\
&= \frac{-1}{2}\sum_{i=1}^{n} \lambda_i y_i (K(\mathbf{x}_i, \mathbf{x}_k) + K(\mathbf{x}_i, \mathbf{x}_l)).
\end{aligned}$$

The kernel classifier is given as:

$$\phi(\mathbf{a})^T\phi(\mathbf{x}) + b^\star \geq 1 \implies \mathbf{x} \in \mathcal{C}_1$$

$$\phi(\mathbf{a})^T\phi(\mathbf{x}) + b^\star \leq -1 \implies \mathbf{x} \in \mathcal{C}_2$$

where

$$\phi(\mathbf{a})^T\phi(\mathbf{x}) = \sum_{i=1}^{n} \lambda_i y_i K(\mathbf{x}_i, \mathbf{x})$$

Solution of Problem 3

Note that the regression problem should be written as

$$y_i = \begin{bmatrix} 1 & x_i \end{bmatrix} \begin{bmatrix} \vartheta_0 \\ \vartheta_1 \end{bmatrix}$$

and for all n samples of (x_i, y_i), we have the following definition :

$$\mathbf{y} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \vartheta_0 \\ \vartheta_1 \end{bmatrix}$$

See that firstly:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{n\sum x_i^2 - (\sum x_i)^2}\begin{bmatrix}\sum x_i^2 & -\sum x_i \\ -\sum x_i & n\end{bmatrix} = \frac{1}{n}\frac{1}{\overline{x^2} - \overline{x}^2}\begin{bmatrix}\overline{x^2} & -\overline{x} \\ -\overline{x} & 1\end{bmatrix}$$

On the other hand we have:

$$\mathbf{X}^T\mathbf{y} = \begin{bmatrix} n\overline{y} \\ \sum x_i y_i \end{bmatrix}.$$

So finally the solution is given by:

$$\begin{aligned}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} &= \frac{1}{n}\frac{1}{\overline{x^2} - \overline{x}^2}\begin{bmatrix}\overline{x^2} & -\overline{x} \\ -\overline{x} & 1\end{bmatrix}\begin{bmatrix} n\overline{y} \\ \sum x_i y_i \end{bmatrix} = \frac{1}{n}\frac{1}{\overline{x^2} - \overline{x}^2}\begin{bmatrix} n\overline{y}\overline{x^2} - \overline{x}(\sum x_i y_i) \\ -n\overline{y}.\overline{x} + \sum x_i y_i \end{bmatrix} \\ &= \frac{1}{\overline{x^2} - \overline{x}^2}\begin{bmatrix} \overline{y}\overline{x^2} - \overline{x}\rho_{xy} \\ -\overline{y}.\overline{x} + \rho_{xy} \end{bmatrix} = \frac{1}{\sigma_x^2}\begin{bmatrix} \overline{y}\overline{x^2} - \overline{x}\rho_{xy} \\ \sigma_{xy} \end{bmatrix}\end{aligned}$$

Therefore $\vartheta_1 = \frac{\sigma_{xy}}{\sigma_x^2}$ and

$$\vartheta_0 = \frac{1}{\sigma_x^2}(\overline{y}\overline{x^2} - \overline{x}\rho_{xy}) = \frac{1}{\sigma_x^2}(\overline{y}\overline{x^2} - \overline{x}(\overline{y}.\overline{x} + \sigma_{xy})) = \frac{1}{\sigma_x^2}(\overline{y}(\overline{x^2} - \overline{x}^2)) - \overline{x}\frac{\sigma_{xy}}{\sigma_x^2}$$

hence $\vartheta_0 = \overline{y} - \vartheta_1\overline{x}$.