02471 Machine Learning for Signal Processing

Solution

Exercise 12: Kernel methods

12.1 Obtaining linear separability using Kernels

Exercise 12.1.1

The points are generated using two uniform distributions. The (x, y) coordinates for the points are generated using the real and imaginary part of the complex exponential function

$$f(\mathrm{r}, \theta) = \mathrm{r}\exp(i\theta), \quad \theta \sim \mathcal{U}[0; 2\pi]$$

For class 1 we have $\mathrm{r} \sim \mathcal{U}[0; 1]$ and for class 2 we have $\mathrm{r} \sim \mathcal{U}[1.5; 2.5]$, where $\mathcal{U}[a, b]$ denotes the uniform distribution on the interval $[a, b]$.

Exercise 12.1.3

$$\begin{aligned}
\boldsymbol{\phi}^T(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{y}) &= \begin{bmatrix} x_1^2 & \sqrt{2}x_1x_2 & x_2^2 \end{bmatrix} \begin{bmatrix} y_1^2 \\ \sqrt{2}y_1y_2 \\ y_2^2 \end{bmatrix} \\
&= x_1^2y_1^2 + 2x_1x_2y_1y_2 + x_2^2y_2^2 \\
&= (x_1y_1 + x_2y_2)^2 \\
&= (\boldsymbol{x}^T\boldsymbol{y})^2
\end{aligned}$$

12.2 Derivation of the kernel ridge regression

Exercise 12.2.1

From def 8.15 we use $\langle a\boldsymbol{x}, \boldsymbol{y}\rangle = a\langle \boldsymbol{x}, \boldsymbol{y}\rangle$, and additionally, since we are in $\mathbb{R}$ we have $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \langle \boldsymbol{y}, \boldsymbol{x}\rangle^* = \langle \boldsymbol{y}, \boldsymbol{x}\rangle$.

Now, by inserting f in the inner product, we get

$$\langle \sum_{n=1}^{N} \theta_n \kappa(\cdot, \boldsymbol{x}_n), \sum_{m=1}^{N} \theta_m \kappa(\cdot, \boldsymbol{x}_m)\rangle = \sum_{n=1}^{N} \theta_n \sum_{m=1}^{N} \theta_m \langle \kappa(\cdot, \boldsymbol{x}_n), \kappa(\cdot, \boldsymbol{x}_m)\rangle$$

We now use the property $\langle \kappa(\cdot, \boldsymbol{y}), \kappa(\cdot, \boldsymbol{x})\rangle = \kappa(\boldsymbol{x}, \boldsymbol{y}) = \kappa(\boldsymbol{y}, \boldsymbol{x})$ (eq. (11.9) in the book) to get

$$\sum_{n=1}^{N} \theta_n \sum_{m=1}^{N} \theta_m \langle \kappa(\cdot, \boldsymbol{x}_n), \kappa(\cdot, \boldsymbol{x}_m)\rangle = \sum_{n=1}^{N} \theta_n \sum_{m=1}^{N} \theta_m \kappa(\boldsymbol{x}_n, \boldsymbol{x}_m)$$

Using $\mathcal{K} = \mathcal{K}^T$ and the definition of the kernel matrix (Eq 11.10–11.12 in the book), we obtain

$$\sum_{n=1}^{N} \theta_n \sum_{m=1}^{N} \theta_m \kappa(\boldsymbol{x}_n, \boldsymbol{x}_m) = \boldsymbol{\theta}^T \mathcal{K}^T \boldsymbol{\theta}$$

Exercise 12.2.2

If we define

$$\kappa(\cdot) := [\kappa(\cdot, \boldsymbol{x}_1), \cdots, \kappa(\cdot, \boldsymbol{x}_n)]^T \quad \Rightarrow$$
$$\sum_{n=1}^{N} \theta_n \kappa(\cdot, \boldsymbol{x}_n) = \boldsymbol{\theta}^T \kappa(\cdot)$$

This allows the following rewrite

$$\sum_{n=1}^{N} \left(y_n - \boldsymbol{\theta}^T \kappa(\cdot)\right)^2 = \|\boldsymbol{y} - \mathcal{K}\boldsymbol{\theta}\|^2 = (\boldsymbol{y} - \mathcal{K}\boldsymbol{\theta})^T(\boldsymbol{y} - \mathcal{K}\boldsymbol{\theta}) = \boldsymbol{y}\boldsymbol{y}^T - 2\boldsymbol{y}^T\mathcal{K}\boldsymbol{\theta} + \boldsymbol{\theta}^T\mathcal{K}^T\mathcal{K}\boldsymbol{\theta}$$

Now using the following properties from Appendix A:

$$\frac{\partial \boldsymbol{x}^T A \boldsymbol{x}}{\partial \boldsymbol{x}} = (A + A^T)\boldsymbol{x}$$
$$\frac{\partial A\boldsymbol{x}}{\partial \boldsymbol{x}} = A^T$$

We obtain

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -2(\boldsymbol{y}^T\mathcal{K})^T + 2\mathcal{K}^T\mathcal{K}\boldsymbol{\theta} + 2C\mathcal{K}^T\boldsymbol{\theta} = 0$$

Which leads to

$$\mathcal{K}^T\boldsymbol{y} = (\mathcal{K}^T\mathcal{K} + C\mathcal{K}^T)\hat{\boldsymbol{\theta}}$$

If K^{-1} exists we then obtain

$$(\mathcal{K}^T)^{-1}\mathcal{K}\boldsymbol{y} = (\mathcal{K}^T)^{-1}\mathcal{K}^T(\mathcal{K} + CI)\hat{\boldsymbol{\theta}} \quad \Rightarrow$$
$$\boldsymbol{y} = (\mathcal{K} + CI)\hat{\boldsymbol{\theta}} \quad \Rightarrow$$
$$\hat{\boldsymbol{\theta}} = (\mathcal{K} + CI)^{-1}\boldsymbol{y}$$