EXERCISE SHEET - KERNEL METHODS

STATISTICAL LEARNING COURSE, ECOLE NORMALE SUPÉRIEURE

Raphaël Berthier
raphael.berthier@inria.fr

Let $\mathcal{X}$ be a set. We recall that that a symmetric function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite kernel if the two following equivalent conditions hold:

(a) for all $\alpha_1, \dots, \alpha_n \in \mathbb{R}$, for all $x_1, \dots, x_n \in \mathcal{X}$

$$\sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) \geq 0 \,.$$

(b) there exists a Hilbert space $\mathcal{H}$ and a feature map $\phi : \mathcal{X} \to \mathcal{H}$ such that for all $x, x' \in \mathcal{X}$,

$$k(x, x') = \left\langle \phi(x), \phi(x') \right\rangle \,.$$

1. Examples of positive definite kernels

(1) **Basic operations.** Let k_1, k_2 be two positive definite kernels on $\mathcal{X}$.
  (a) Show that $k_1 + k_2$ is a positive definite kernel on $\mathcal{X}$.
  (b) Show that the pointwise product $k_1 k_2$ is a positive definite kernel on $\mathcal{X}$.

(2) **Minimum.** Show that the function $k(x, y) = \min(x, y)$ is a positive definite kernel on $\mathbb{R}_+$.

(3) **Chi-2.** Show that the function $k(x, y) = \frac{xy}{x+y}$ is a positive definite kernel on $\mathbb{R}_+^*$

(4) **On finite sets.** Let X be a finite set, and $\mathcal{X} = \mathcal{P}(X)$ the collection of all subsets of X. Show that the function $k(A, B) = \frac{|A \cap B|}{|A \cup B|}$ is a positive definite kernel on $\mathcal{X}$.

(5) **Greatest common divisor.** Show that the function $k(n, m) = \mathrm{GCD}(n, m)$, where $\mathrm{GCD}(n, m)$ is the greatest common divisor of n and m, is a positive definite kernel on $\mathbb{N}$.

2. Distance in the feature space

(1) (a) Let k be a positive definite kernel on $\mathcal{X}$. Let $\mathcal{H}$ be a Hilbert space and $\phi : \mathcal{X} \to \mathcal{H}$ be a feature map such that $k(x, y) = \langle \phi(x), \phi(y) \rangle$. Express the distance $\|\phi(x) - \phi(y)\|^2$ as a function of k only.
  (b) What is the distance in the case of the Chi-2 kernel?

(2) In this question, we want to apply the kernel trick for a simple method for classification: the distance to the mean. More precisely, we are given points $x_1, \dots, x_n \in \mathcal{X}$ and their classes $y_1, \dots, y_n \in \{-1, 1\}$. We define the centers of each class

$$\mu_+ = \frac{1}{n_+} \sum_{i, y_i = 1} \phi(x_i) \qquad\qquad \mu_- = \frac{1}{n_-} \sum_{i, y_i = -1} \phi(x_i)$$

where n_+ (resp. n_-) is the number of points labeled $+1$ (resp. -1). We classify a new point $\Phi(x)$ in the class $+1$ if and only if it is closer to μ_+ than μ_- in the kernel space.

(a) Express the distance $\|\phi(x) - \mu_+\|^2$ using the kernel k and the data only.
(b) Express the classication rule explained above using the kernel k and the data only.
(c) Assume now and in the next question that $\|\mu_+\| = \|\mu_-\|$, and that the two classes are balanced, i.e., $n_+ = n_-$. Simplify the rule of the previous question.
(d) We now take $\mathcal{X} = \mathbb{R}^d$ and the Gaussian kernel $k(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$. Show that the classification rule boils down to a local averaging rule.