Tutorial 11 : Kernels

1 Theory

Non-Linear Classifier

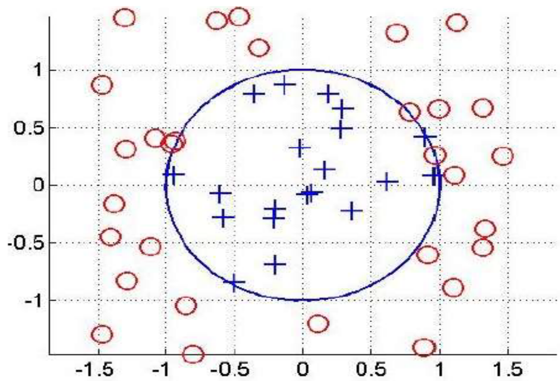
A non-linear classifier is of the form

$$sign\left(w^T\phi(x) + b\right)$$

where $\phi(x) \triangleq [\phi_1(x)\ \phi_2(x)\ \cdots \phi_M(x)]$ with non-linear functions $\phi_i(x) : \mathbb{R}^d \to \mathbb{R}$.

Example

Consider the following problem



There is no linear classifier which classify the examples with no error. However, define the transformation $\phi(x) = [x_1^2, x_2^2]$, the following classifier

$$sign(-x_1^2 - x_2^2 + R^2)$$

can classify the examples above with no error.

Kernel Functions

The use of non-linear transformation $\phi(x)$ allows to extend our frameworks to non-linear classifier and solve problems for which the examples are not linearly separable in the space. However, the dimension of $\phi(x)$ might be large and even infinite, thus, leading to a huge computational load. As in SVM, When the weights can be written as

$$w = \sum_{i=1}^{n} \alpha_i y_i \phi(x_i),$$

there is no need to compute $\phi(x)$ explicitly since the classifier requires to compute only inner products

$$\hat{f}_w(x) = sign(\sum_{i=1}^{n} \alpha_i y_i \phi(x_i)^T \phi(x) + b).$$

We define a kernel function on a set $X \subseteq \mathbb{R}^d$ as a function $K : X \times X \to \mathbb{R}$ which satisfies

1. Symmetric $K(x, z) = K(z, x)$.
2. For every finite set of points $\{x_1, x_2, ..., x_n\}$ the matrix $K_{il} = K(x_i, x_l)$ is positive semi-definite (PSD).

Then, under some reasonable technical conditions, there exists a basis $\phi(x)$ such that the kernel function is a dot product of the form $K(x_i, x_l) = \phi(x_i)^T \phi(x_l)$. In this case, the classifier is

$$\hat{f}_w(x) = sign(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b).$$

Examples of kernel functions:

- Gaussian kernel - $K(x, z) = \exp\big(-||x - z||_2^2/c\big)$ where $c > 0$.
- Polynomial kernel - $K(x, z) = \big(1 + x^T z)^p\big)$ where $p \geq 1$.

2 Practice

Question 1

(a) For $x \in \mathbb{R}^2$ we define the following feature vector

$$\phi(x) = \begin{bmatrix} x_1^2 \\ \sqrt{2}x_1 x_2 \\ x_2^2 \end{bmatrix} \in \mathbb{R}^3.$$

Prove for the kernel $K(x, z) = (x^T z)^2$ it holds that $K(x, z) = \langle \phi(x), \phi(z) \rangle$.

In the general case we have input of dimension d, that is $x \in \mathbb{R}^d$ and the features are $\phi_m(x) = c_m x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_d^{\alpha_d}$ such that $\alpha_i \in \mathbb{N}$ and $\sum_{i=1}^{d} \alpha_i = p$, that is, every feature is a product of p coordinates of x with possible repetitions. For example, in item (a) we saw inputs and features for $d = 2$ and $p = 2$.

(b) It is given that for the kernel $K(x, z) = \langle x, z \rangle$ it holds that $K(x, z) = \langle \phi(x), \phi(z) \rangle$. What is the dimension of $\phi(x)$ for general d, p?

Solution

(a) We compute the following inner product:

$$\begin{aligned}
\phi(x)^T \phi(z) &= \begin{bmatrix} x_1^2 & \sqrt{2}x_1 x_2 & x_2^2 \end{bmatrix} \begin{bmatrix} z_1^2 \\ \sqrt{2}z_1 z_2 \\ z_2^2 \end{bmatrix} \\
&= x_1^2 z_1^2 + 2x_1 x_2 z_1 z_2 + x_2^2 z_2^2 \\
&= (x_1 z_1 + x_2 z_2)^2 \\
&= (x^T z)^2 = K(x, z).
\end{aligned}$$

(b) In each coordinate of the features vector the sum of exponents is $p \to \sum_{i=1}^{d} \alpha_i$. Hence, the number of coordinates is determined by the number of possible partitions of p exponents to d components. This problem can be described by the following illustration:

$$\Big| \ \alpha_1 \ \Big| \ \alpha_2 \ \Big| \ \cdots \ \Big| \ \alpha_i \ \Big| \ \cdots \ \Big| \ \alpha_d \ \Big|$$

where we have $d+1$ dividers for d components. the left most and right most dividers does not affect the partition, hence, we have $d-1$ dividers and p exponents to arrange. Thus, the number of possible partitions is $\frac{((p+d-1)!}{p!(d-1)!} = \binom{p+d-1}{p}$. This implies the dimension of $\phi(x)$ is exponential in p and using the kernel function can reduce the computation significantly.

Question 2

Consider the following two examples

$$x_1 = (+1, +1), \ y_1 = +1,$$
$$x_2 = (-1, -1), \ y_2 = -1.$$

Compute the separating plane for the Gaussian kernel

$$K(x, z) = \exp\big(- ||x - z||_2^2 \big).$$

Solution

Recall that the classification is given by

$$sign\big(w^T \phi(x)\big) = sign\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) \right).$$

Therefore, the separating plane is defined by the following equation:

$$\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) = 0.$$

To find the coefficients $\{\alpha_i\}$ we solve the dual problem:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{l=1}^{n} \alpha_i \alpha_l y_i y_l K(x_i, x_l) \\ s.t. \quad & \alpha_i \geq 0, \ i = 1, 2, ..., n, \\ & \sum_{i=1}^{n} \alpha_i y_i = 0. \end{aligned}$$

In our case, we have

$$\begin{aligned} \max_{\alpha} \quad & \alpha_1 + \alpha_2 - \frac{1}{2} \big(\alpha_1^2 K(x_1, x_1) + \alpha_2^2 K(x_2, x_2) - 2\alpha_1 \alpha_2 K(x_1, x_2) \big) \\ s.t. \quad & \alpha_i \geq 0, \ i = 1, 2, \\ & \alpha_1 - \alpha_2 = 0. \end{aligned}$$

Notice that $K(x_1, x_1) = K(x_2, x_2) = 1$ and $K(x_1, x_2) = e^{-||x_1 - x_2||_2^2} = e^{-8} \leq 1$. In additions, the second constraint implies that $\alpha_1 = \alpha_2$, hence we get

$$\begin{aligned} \max_{\alpha_1} \quad & 2\alpha_1 - \alpha_1^2 (1 - e^{-8}) \\ s.t. \quad & \alpha_1 \geq 0. \end{aligned}$$

The objective function is a simple quadratic function, hence, we compute the derivative and set it to zero:

$$\alpha_1 (1 - e^{-8}) = 1 \ \Rightarrow \ \alpha_1 = \frac{1}{1 - e^{-8}} > 0.$$

Thus, the separating plane is given by

$$\begin{aligned}
&\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) = 0, \\
&\Rightarrow \alpha_1 \left(K(x_1, x) - K(x_2, x)\right) = 0, \\
&\Rightarrow K(x_1, x) = K(x_2, x), \\
&\Rightarrow e^{-||x-x_1||_2^2} = e^{-||x-x_2||_2^2}, \\
&\Rightarrow ||x - x_1||_2^2 = ||x - x_2||_2^2,
\end{aligned}$$

which is a line in $\mathbb{R}^2$ as expected.