Kernel Functions

Benedetto Manca

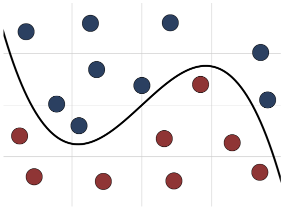
What are kernels?

- The kernels (or kernel functions) are a tool that allow us to extend linear classifiers (like SVM) to non-linear classifiers
- The idea is to define a function that defines implicitly an inner product on a high dimensional vector space
- By replacing the inner product with the one defined via the kernel we are able to determine a non linear separation surface using the same algorithm that yield a linear separation surface

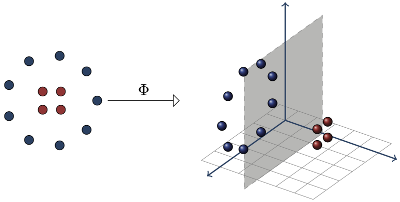
What to do when the input data are not linearly separable?

- If we are given two sets of points that are not linearly separable we can separate them in two ways:

Define a non-linear separator



Map the input point in a space where they become linearly separable



Definition of Kernel

- Let $\mathcal{X} \subseteq \mathbb{R}^n$ be the input sample and $\Phi : \mathcal{X} \to \mathbb{R}^N$ a map onto a higher dimensional space called *feature space*

Definition

A kernel function on $\mathcal{X}$ is a map $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that

$$\forall x, x' \in \mathcal{X} \qquad \mathcal{K}(x, x') = \langle \Phi(x), \Phi(x') \rangle$$

- Usually the computation of the map Φ is expensive compare to computing the kernel function $\mathcal{K}$
- We want to be able to extend the linear separation using only $\mathcal{K}$

Definizione di Kernel

- Not all the function $\mathcal{K}$ are kernel functions, but it is possible to impose some conditions on $\mathcal{K}$ that guarantees the existence of the feature map Φ

Theorem (Mercer's Condition)

Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a compact set and $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric and continuous function. Then $\mathcal{K}$ admits a uniformly convergent expansion of the form

$$K(x, x') = \sum_{i=0}^{\infty} a_i \phi(x)\phi(x'), \quad a_i > 0$$

if and only if, for every $c \in L_2(\mathcal{X})$ the following holds

$$\int_{\mathcal{X}\times\mathcal{X}} c(x)c(x')\mathcal{K}(x, x')dxdx' \geq 0$$

- The Mercer's condition also guarantees that the corresponding optimization problem is convex (unique solution)
- Under the same hypothesis, the Mercer's Condition is equivalent to $\mathcal{K}$ being positive-semidefinite

Positive semidefinite Kernels

Definition

A kernel $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive semidefinite if for every $\{x_1, \dots, x_m\} \subseteq \mathcal{X}$ the matrix $\mathbf{K} := (K(x_i, x_j))_{i,j} \in \mathbb{R}^{m \times m}$ is positive semidefinite.

- $\mathbf{K}$ is positive semidefinite if satisfies one of the following equivalent conditions:
  - The eigenvalues of $\mathbf{K}$ are non-negatives;
  - For every $x \in \mathbb{R}^m$ holds

$$x^\top \mathbf{K} x = \sum_{i,j} x_i x_j K(x_i, x_j) \geq 0$$

- The matric $\mathbf{K}$ is called kernel matrix (or Gram matrix) of the kernel $\mathcal{K}$ associated to the sample $\{x_1, \dots, x_m\}$

Examples

Definition

Given a constant $c > 0$, the polynomial kernel of degree $d \in \mathbb{N}$ on $\mathbb{R}^n$ is given by

$$\mathcal{K}(x, x') = (x \cdot x' + c)^d$$

Definition

Given a constant $\sigma > 0$, the Gaussian (or RBF) kernel on $\mathbb{R}^n$ is given by

$$\mathcal{K}(x, x') = \exp\left(-\frac{\|x' - x\|_2^2}{2\sigma^2}\right)$$

Definition (This is not PSD)

Given two constants $a, b, > 0$ the Sigmoidal kernel on $\mathbb{R}^n$ is given by

$$\mathcal{K}(x, x') = \tanh(a(x \cdot x') + b)$$

Examples

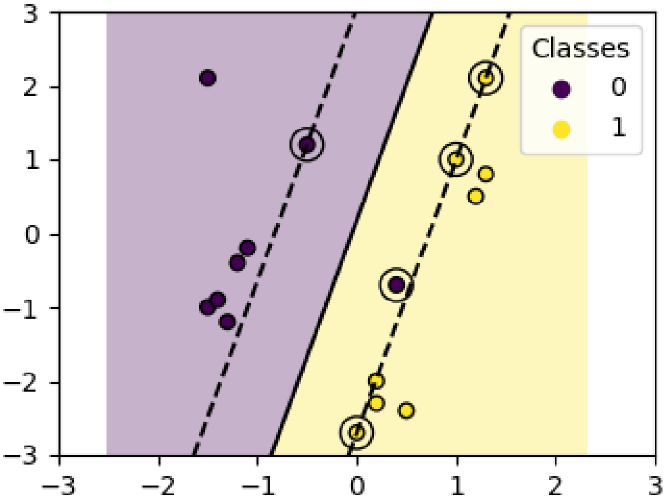


Figure: Linear SVM

Examples

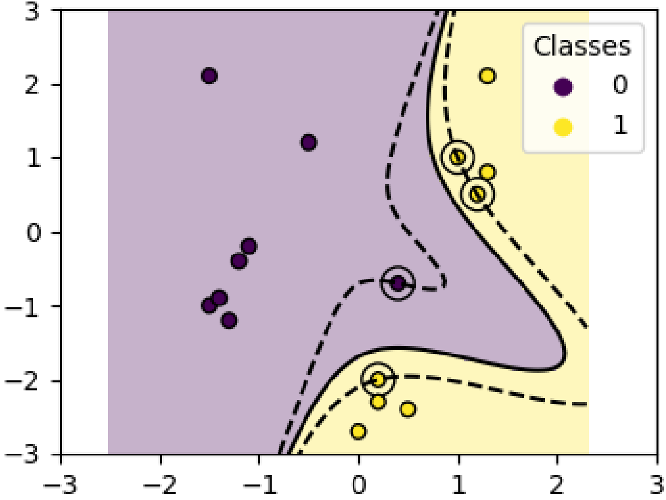


Figure: Polynomial SVM

Examples

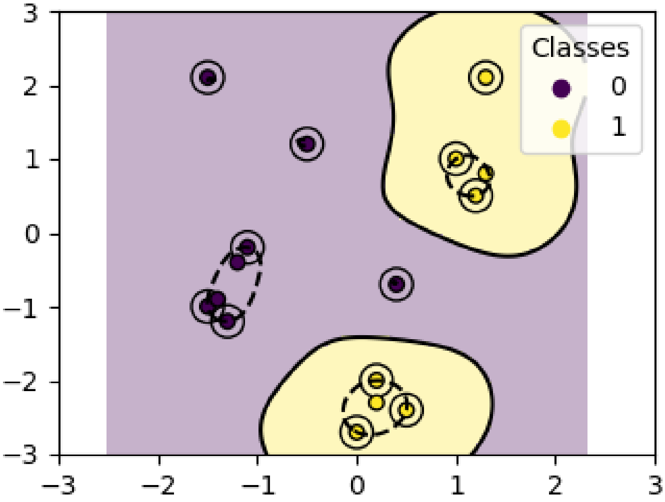


Figure: Gaussian SVM

Examples

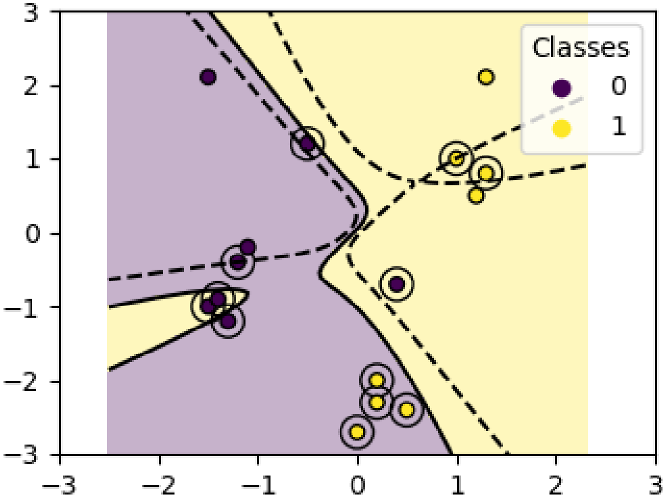


Figure: Sigmoidal SVM

Using a kernel on SVM

- We have seen that the dual optimization problem of SVM is

$$\begin{cases} \max\limits_{\alpha} & \sum_i \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) \\ s.t. & \sum_i \alpha_i y_i = 0 \\ & 0 \le \alpha_i \le C, \quad i \le m \end{cases}$$

- Given a positive semidefinite kernel $\mathcal{K}$, we can apply it to SVM by replacing $x_i \cdot x_j$ with $\mathcal{K}(x_i, x_j)$

$$\begin{cases} \max\limits_{\alpha} & \sum_i \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \mathcal{K}(x_i, x_j) \\ t.c. & \sum_i \alpha_i y_i = 0 \\ & 0 \le \alpha_i \le C, \quad i \le m \end{cases}$$

- The corresponding classifying function becomes:

$$f(x) = \mathsf{sgn}\left(\sum_i \alpha_i y_i \mathcal{K}(x, x_i) + b\right)$$

Homework

- Implement the Gaussian and Sigmoidal kernel function in the Helper file and adjust the SVM classifier so that it is possible to choose between its linear version, the polynomial, Gaussian and Sigmoidal kernel.