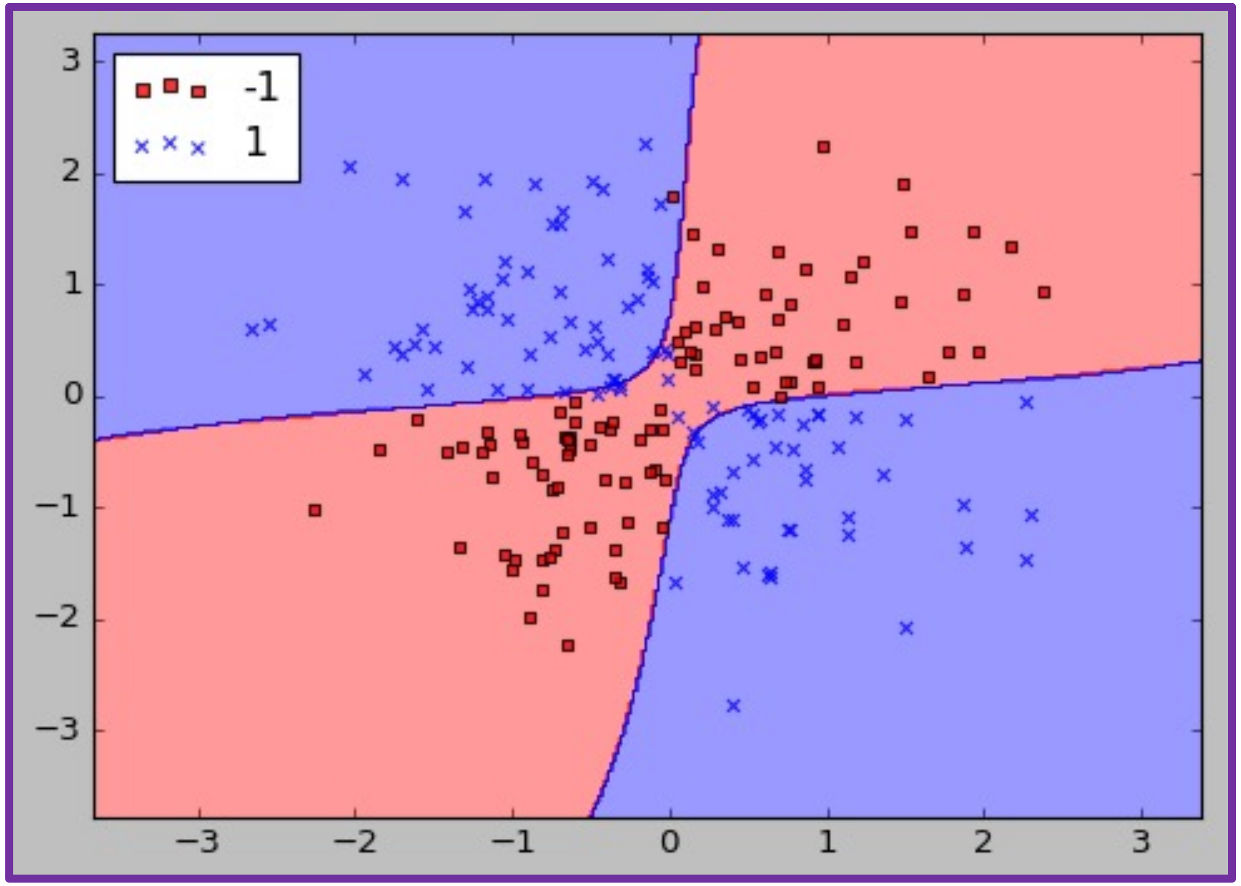
A brief introduction to Kernels

Ariel Shamir

Zohar Yakhini

IDC HERZLIYA



A systematic approach to classification by mapping into higher dimension

- Try to map into the full rational variety of increasing degrees.
- Apply the Perceptron in the mapping (ambient) space.

- Overfitting?
- The Perceptron uses inner products. What would the time complexity of the operation be?

Here come the kernels …

Kernels

- A function $K: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is called a **kernel** if there exists a mapping function $\varphi: \mathbb{R}^n \to \mathbb{R}^N$ so that the following always holds:

$$\forall x, y \qquad K(x, y) = \varphi(x) \cdot \varphi(y)$$

- A mapping function $\varphi: \mathbb{R}^n \to \mathbb{R}^N$ is said to **afford a kernel** if such K exists.
- Kernels transform the learning into direct operations in the (lower dimension) input space (will see how).
- Kernels help us avoid the explicit search for an ambient space and for mapping functions, φ, into higher dimensions. We explore kernels instead.
- In fact, kernels can also support learning in infinite dimensional mapping spaces (general Hilbert spaces)

$$
\begin{tikzcd}
 & \mathbb{R}^N \times \mathbb{R}^N \arrow[dr, "\text{Inner product}"] & \\
\mathbb{R}^n \times \mathbb{R}^n \arrow[ur, "\varphi"] \arrow[rr, "K"'] & & \mathbb{R}
\end{tikzcd}
$$

Kernel Example

For $x = (x_1, x_2) \in R^2$ let $\varphi(\vec{x}) = (x_1^2, \sqrt{2}x_1 x_2, x_2^2)$

Given $x = (x_1, x_2)$ $y = (y_1, y_2)$ we then get

$$\varphi(x)\varphi(y) = (x_1^2, \sqrt{2}x_1 x_2, x_2^2)\begin{pmatrix} y_1^2 \\ \sqrt{2}y_1 y_2 \\ y_2^2 \end{pmatrix}$$

$$= x_1^2 y_1^2 + 2x_1 x_2 y_1 y_2 + x_2^2 y_2^2$$

$$= \left((x_1, x_2)\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \right)^2 = (x \cdot y)^2$$

Defining $K(x, y) = (x \cdot y)^2$ we therefore have a kernel for the mapping φ

More Kernel Examples

$$\text{Homogenous Polynomial kernel}: \quad k(x,y) = (x \cdot y)^d$$

$$\text{Inhomogenous Polynomial kernel}: \quad k(x,y) = (x \cdot y + 1)^d$$

$$\text{Radial Basis function kernel}: k(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$$

Sums and products of kernels are kernels as well!

A kernel for the full rational varieties mapping?

$$\varphi(\vec{x}) = \ (1, x_1, x_2, x_3, x_4, x_1 \cdot x_2,\ x_1 \cdot x_3, x_1 \cdot x_4, x_2 \cdot x_3, \dots, x_1^2, x_2^2, x_3^2, x_4^2)$$

- What is N? (the ambient dimension)
- Note that in seeking a kernel we might as well find one for a version of φ which involves coefficients.

Mercer-Gram Thm

A function $K: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a kernel
iff
For every finite set of vectors in $\mathbb{R}^n$: $\{x_1, x_2, \dots, x_\ell\}$, the matrix G defined by $G(i,j) = K\left(x_i, x_j\right)$ is positive semi definite.

Proof for the easy direction:

For any vector $\mathbf{v}$ we have

$$
\begin{aligned}
\mathbf{v}'\mathbf{G}\mathbf{v} &= \sum_{i,j=1}^{\ell} v_i v_j \mathbf{G}_{ij} = \sum_{i,j=1}^{\ell} v_i v_j \left\langle \phi\left(\mathbf{x}_i\right), \phi\left(\mathbf{x}_j\right)\right\rangle \\
&= \left\langle \sum_{i=1}^{\ell} v_i \phi\left(\mathbf{x}_i\right), \sum_{j=1}^{\ell} v_j \phi\left(\mathbf{x}_j\right) \right\rangle \\
&= \left\| \sum_{i=1}^{\ell} v_i \phi\left(\mathbf{x}_i\right) \right\|^2 \geq 0,
\end{aligned}
$$

Next week

- The Dual Perceptron
- The Kernel Perceptron
  (efficiently implementing the rational varieties approach)
- Large margin classifiers
- Lagrange optimization
- SVMs, their geometric representation and related topics
- Slack variables and examples