Kernel Functions

If we let $k(\overline{x}, \overline{y}) = \Phi(\overline{x}) \bullet \Phi(\overline{y})$ be a kernel function, then we can write our support vector machine in terms of kernels,

$$\hat{f}(\overline{x}) = \text{sign}\left(\sum_{i=1}^{l} \alpha_i^* y_i k(\overline{x}_i, \overline{x}) - \sum_{i=1}^{l} \alpha_i^* y_i k(\overline{x}_i, \overline{x}_{sv+}) + 1\right)$$

We can write our training algorithm in terms of kernel functions as well,

$$\overline{\alpha}^* = \underset{\overline{\alpha}}{\text{argmax}}\left(\sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j k(\overline{x}_i, \overline{x}_j)\right),$$

subject to the constraints,

$$\sum_{i=1}^{l} \alpha_i y_i = 0,$$

$$\alpha_i \geq 0, \quad i = 1, \ldots, l.$$

Selecting the right kernel for a particular non-linear classification problem is called *feature search*.

Kernel Functions

Kernel Name	Kernel Function	Free Parameters
Linear Kernel	$k(\overline{x}, \overline{y}) = \overline{x} \bullet \overline{y}$	none
Homogeneous Polynomial Kernel	$k(\overline{x}, \overline{y}) = (\overline{x} \bullet \overline{y})^d$	$d \geq 2$
Non-Homogeneous Polynomial Kernel	$k(\overline{x}, \overline{y}) = (\overline{x} \bullet \overline{y} + c)^d$	$d \geq 2$, $c > 0$
Gaussian Kernel	$k(\overline{x}, \overline{y}) = e^{-\frac{\lvert\overline{x}-\overline{y}\rvert^2}{2\sigma^2}}$	$\sigma > 0$

Non-linear Classifiers

Let's review classification with non-linear SVMs:

1. We have a non-linear data set.
2. Pick a kernel other than the linear kernel, this means that the input space will be transformed into a higher dimensional feature space.
3. Solve our dual maximum margin problem in the feature space (we are solving now a linear classification problem).
4. The resulting model is a linear model in feature space and a *non-linear model* in input space.

A Closer Look at Kernels

We have shown that for $\Phi(x_1, x_2) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$ the kernel

$$k(\overline{x}, \overline{y}) = \Phi(\overline{x}) \bullet \Phi(\overline{y}) = (\overline{x} \bullet \overline{y})^2.$$

That is, we picked our mapping from input space into feature space in such a way that the kernel in feature space can be evaluated in input space.

This begs the question: What about the other kernels? What do the associated feature spaces and mappings look like?

It turns out that for each kernel function we can construct a canonical feature space and mapping. This implies that features spaces and mappings for kernels are not unique!

Properties of Kernels

Definition: [Positive Definite Kernel] A function $k : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$ such that

$$\sum_{i=1}^{l} \sum_{j=1}^{l} \theta_i \theta_j k(\overline{x}_i, \overline{x}_j) \geq 0$$

holds is called a ***positive definite kernel***. Here, $\theta_i, \theta_j \in \mathbb{R}$ and $\overline{x}_1, \ldots, \overline{x}_l$ is a set of points in $\mathbb{R}^n$.

Properties of Kernels

New notation: Let $k : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ be a kernel, then $k(\cdot, \overline{x})$ is a partially evaluated kernel with $\overline{x} \in \mathbb{R}^n$ and represents a function $\mathbb{R}^n \rightarrow \mathbb{R}$.

Theorem: [Reproducing Kernel Property] Let $k : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a positive definite kernel, then the following property holds,

$$k(\overline{x}, \overline{y}) = k(\overline{x}, \cdot) \bullet k(\cdot, \overline{y}),$$

with $\overline{x}, \overline{y} \in \mathbb{R}^n$.

Feature Spaces are not Unique

We illustrate that feature spaces are not unique using our homogeneous polynomial kernel to the power of two, that is, $k(\overline{x}, \overline{y}) = (\overline{x} \bullet \overline{y})^2$ with $\overline{x}, \overline{y} \in \mathbb{R}^2$. Let $\Phi : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ such that

$$\Phi(\overline{x}) = \Phi(x_1, x_2) = (x_1^2, x_2^2, \sqrt{2}x_1^2 x_2^2)$$

and $\Psi : \mathbb{R}^2 \rightarrow \{\mathbb{R}^2 \rightarrow \mathbb{R}\}$ with

$$\Psi(\overline{x}) = k(\cdot, \overline{x}) = ((\cdot) \bullet \overline{x})^2,$$

be two mappings from our input space to two different feature spaces, then

$$\begin{aligned}
\Phi(\overline{x}) \bullet \Phi(\overline{y}) &= (\overline{x}_1^2, \overline{x}_2^2, \sqrt{2}\overline{x}_1^2\overline{x}_2^2) \bullet (\overline{y}_1^2, \overline{y}_2^2, \sqrt{2}\overline{y}_1^2\overline{y}_2^2) \\
&= (\overline{x} \bullet \overline{y})^2 \\
&= k(\overline{x}, \overline{y}) \\
&= k(\cdot, \overline{x}) \bullet k(\cdot, \overline{y}) \\
&= ((\cdot) \bullet \overline{x})^2 \bullet ((\cdot) \bullet \overline{y})^2 \\
&= \Psi(\overline{x}) \bullet \Psi(\overline{y}).
\end{aligned}$$

The section on kernels in the book shows that the construction $\Psi(\overline{x}) \bullet \Psi(\overline{y})$ is indeed well defined and represents a dot product in an appropriate feature space.