Kernel Method

Jerry Peng

1 Kernel Method

The kernel method is a technique used in machine learning and statistics to transform data into a different space, making it easier to work with. The idea is to map the original data into a higher-dimensional space where it becomes linearly separable or easier to model. This is particularly useful for algorithms like Support Vector Machines (SVMs) that work well with linearly separable data but struggle with data that is not easily separable in its original form. For example, in Fig 1, we can see that in the input space, data points are not easy to be classified, and if we use a very complicate model to perform this task, we may get overfitting problem. Hence, through some kernel method ϕ, we can map the data points from the input space (low-dimension) to the feature space (high-dimension) and make those data points linear separable.

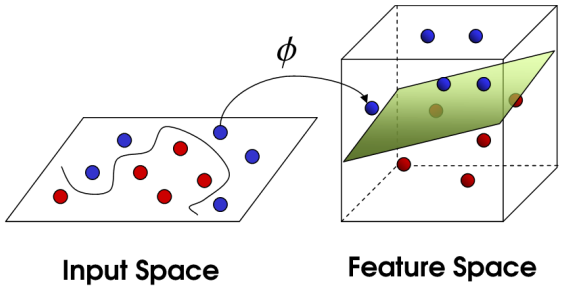

Figure 1: Kernel Method

Mapping Function The mapping function $\phi(x)$ is used to transform data points from the original feature space to a higher-dimensional space. This is done to make the data linearly separable or easier to work with. It takes a data point x in the original feature space and maps it to a new point $\phi(x)$ in the higher-dimensional space.

$$\phi : \mathbb{R}^n \to \mathbb{R}^m \tag{1}$$

where n is the dimensionality of the original space and m is the dimensionality of the transformed space, and $n < m$. The mapping is often explicit, meaning you could technically compute $\phi(x)$ for each x, although this is usually computationally expensive.

Kernel Function The kernel function $K(x, y)$ is used to compute the dot product between two transformed vectors in the higher-dimensional space without explicitly performing the transformation. This makes the computation more efficient. It takes two data points x and y in the original feature space and returns the dot product of their images in the higher-dimensional space.

$$K(x, y) = \phi(x) \cdot \phi(y) \tag{2}$$

The kernel function allows for an implicit transformation of the data, meaning you don't have to compute $\phi(x)$ and $\phi(y)$ explicitly to get their dot product.

The kernel method is widely used in various machine learning tasks like classification, regression, and clustering. It's a powerful tool for capturing complex relationships in data without having to explicitly compute the transformation.

1.1 Example

Let's consider a simple example using a polynomial kernel to illustrate the difference between the mapping function and the kernel function. Suppose we have two data points in a 2D space: $X = (x_1, x_2)$ and $Y = (y_1, y_2)$

Mapping Function For a polynomial of degree 2, the mapping function could be:

$$\phi(x) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$

Therefore, if we have $x_1 = 1$ and $x_2 = 2$, then $\phi(x) = (1, 2\sqrt{2}, 4)$

Kernel Function The polynomial kernel function of degree 2 is defined as:

$$K(x, y) = \phi(x) \cdot \phi(y) = x_1^2x_2^2 + 2x_1x_2y_1y_2 + y_1^2y_2^2$$

After using Kernel Function, we can deploy SVM classifier on this dataset.

1.2 Some Kernel Functions

Linear Kernel The linear kernel is the simplest type of kernel, which essentially doesn't transform the input data. It's equivalent to the standard dot product in the original feature space.

$$K_{linear}(x, y) = x \cdot y \tag{3}$$

Polynomial Kernel The polynomial kernel allows you to fit polynomial decision boundaries. The degree of the polynomial determines the complexity of the decision boundary.

$$K_{polynomial}(x, y) = (x \cdot y + c)^d \tag{4}$$

where c is a constant term (usually is 1), and d is the degree of the polynomial

Radial Basis Function (RBF) or Gaussian Kernel The RBF kernel is a popular choice for Support Vector Machines and other algorithms. It can map an input space into an infinite-dimensional space, making it a powerful choice for handling non-linear data.

$$K_{RBF}(x, y) = e^{-\gamma\|x-y\|^2} \tag{5}$$

where γ is a parameter that determines the spread of the function.

Sigmoid Kernel The sigmoid kernel is inspired by the sigmoid activation function commonly used in neural networks. It's not positive definite for some parameter values, so it's less commonly used for SVMs.

$$K_{sigmoid} = \tanh(\alpha x \cdot y + c) \tag{6}$$

where α is the scaling factor and c is the constant term

Laplacian Kernel Similar to the RBF kernel but uses the Manhattan distance (L_1 norm) instead of the Euclidean distance (L_2 norm)

$$K_{laplacian}(x, y) = e^{-\gamma\|x-y\|_1} \tag{7}$$

Cosine Similarity Kernel Measures the cosine of the angle between two vectors, often used in text analysis.

$$K_{cosine}(x, y) = \frac{x \cdot y}{\|x\|_2 \times \|y\|_2} \tag{8}$$

References