Lecture-04: PDS Kernels and RKHS

1 Kernel Methods

Kernel methods are extensions of SVMs to define non-linear decision boundaries, and can also be used for other algorithms that depend solely on inner products between sample points.

Kernel functions map the data to higher dimensional space. Under symmetry and positive definiteness of these kernel functions, we can define inner product in this high dimensional space. A linear separation in this high dimensional space is non-linear separation in the original space.

Example 1.1 (Document classification). Let $\mathcal{X}$ be the set of words in a document, which has a typical size of $|\mathcal{X}| = 10^5$ words. Classifying the document into different types based on single words (elements from the set $\mathcal{X}$) will be difficult because many types of documents will share the same words. A better way to classify documents is to look for patterns in groups of adjacent words. For example, consider $\mathcal{X}^3$, which is the set of trigrams (triplets of words). Classifying documents in the space of trigrams will yield better results despite the increased size of the space $|\mathcal{X}^3| = 10^{15}$.

Remark 1. The complexity of linear separation algorithm like SVM doesn't depend on the dimension of the space, rather on the margin ρ. However, the higher dimension inner product may become costly.

Definition 1.2 (Kernels). For the input space $\mathcal{X}$, we let the non-linear map $\Phi : \mathcal{X} \to \mathbb{H}$ be a **feature mapping** that takes our feature vectors to a higher dimensional space Hilbert $\mathbb{H}$ called a **feature space**. A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a **kernel** over $\mathcal{X}$. For this mapping Φ, we define a kernel K by the inner product in the space $\mathbb{H}$, such that

$$K(x,x') = \left\langle \Phi(x), \Phi(x') \right\rangle_{\mathbb{H}}, \text{ for all } x,x' \in \mathcal{X}.$$

Remark 2. The inner product $\langle \cdot, \cdot \rangle$ is similarity measure between two feature vectors in the feature space $\mathbb{H}$. The kernel K is a similarity measure between elements of the input space $\mathcal{X}$.

Example 1.3 (Polynomial kernel). For $c > 0$ and degree $d \in \mathbb{N}$, we define a kernel

$$K(x,x') \triangleq (\langle x,x' \rangle + c)^d, \text{ for all } x,x' \in \mathcal{X} \subseteq \mathbb{R}^N.$$

For this mapping $\Phi : \mathcal{X} \to \mathbb{H}$, can you find the dimension of $\mathbb{H}$? For $N = 2$ and $d = 2$, we see that $\Phi : \mathcal{X} \to \mathbb{H}$ given by $\Phi(x) = \begin{bmatrix} x_1^2 & x_2^2 & \sqrt{2}x_1x_2 & \sqrt{2c}x_1 & \sqrt{2c}x_2 & c \end{bmatrix}$ suffices to give us $K(x,x') = \langle \Phi(x), \Phi(x') \rangle_{\mathbb{H}}$ for all $x,x' \in \mathbb{R}^2$.

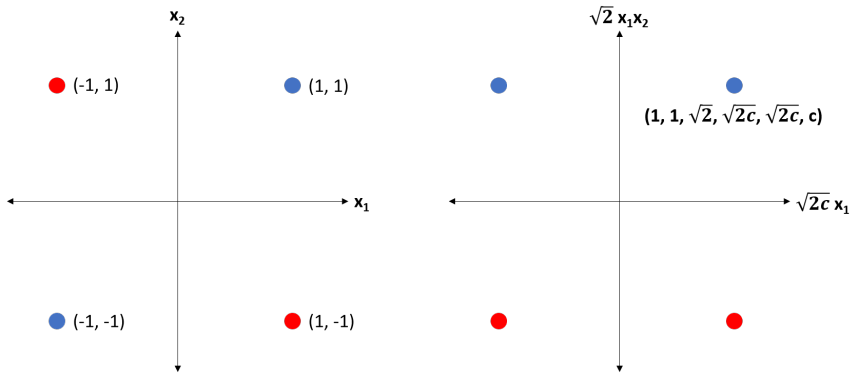

Figure 1: Left: Four points from two classes plotted on the x_1, x_2 axes. These points are not separable by any hyperplane. Right: The same four points are plotted on the $\sqrt{2}x_1x_2$ and $\sqrt{2c}x_1$ axes. These points are now separable.

Consider the following classification problem shown in Figure 1, where the red and the blue points must be separated by a hyperplane. This is not possible in the space $\mathbb{R}^2$ since there is no hyperplane that can separate

the blue and red points. However, when we use the function $h(x_1, x_2) = x_1 x_2$ to bring these points to a higher-dimensional space, we find that these points are indeed separable along the $x_1 x_2$ dimension.

Remark 3. Why do we work with kernels?

- **Efficiency**: Inner product in higher dimensional space is equal to the computation of kernel function in the input space. Computation in the input space $\mathcal{X}$ is more efficient than computation in the feature space $\mathbb{H}$ because $\dim(\mathbb{H}) >> \dim(\mathcal{X})$ and $\langle x, y \rangle = O(\dim(\mathcal{X}))$.
- **Flexibility**: There is no need to explicitly define the map Φ but its existence is guaranteed if K satisfies Mercer's condition.

Theorem 1.4 (Mercer's condition). *Let $\mathcal{X} \subseteq \mathbb{R}^N$ be a compact set and let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous and symmetric function. Then, the kernel K admits a uniformly convergent expansion of the form*

$$K(x, x') = \sum_{n=0}^{\infty} a_n \phi_n(x) \phi_n(x'),$$

with $a_n > 0$ iff for any square integrable function $c \in L_2(x)$, the following condition holds

$$\iint_{\mathcal{X} \times \mathcal{X}} c(x) c(x') K(x, x') dx dx' \geqslant 0.$$

This is the positive semi-definiteness condition on the kernel K.

This condition is important to guarantee the convexity of the optimization problem for algorithms such as SVMs and thus convergence guarantees. A condition that is equivalent to Mercer?s condition under the assumptions of the theorem is that the kernel K be **positive definite symmetric** (PDS). This property is in fact more general since in particular it does not require any assumption about $\mathcal{X}$.

1.1 PDS Kernels

Definition 1.5 (Gram matrix). For a sample $S = (x_1, \ldots, x_m)$, the **kernel matrix** or the **Gram matrix** associated to the kernel K and the sample S is denoted by $\mathbf{K} \in \mathbb{R}^{m \times m}$ and given by

$$\begin{bmatrix} K(x_1, x_1) & \ldots & K(x_1, x_m) \\ \vdots & \ddots & \vdots \\ K(x_m, x_1) & \ldots & K(x_m, x_m) \end{bmatrix}.$$

Definition 1.6 (PDS kernels). A kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be **positive definite symmetric (PDS)** if for any $x \in \mathcal{X}^m$, the Gram matrix $\mathbf{K} = [K(x_i, x_j)]_{ij} \in \mathbb{R}^{m \times m}$ is *symmetric positive semi-definite (SPSD)*.

Remark 4. The matrix $\mathbf{K}$ is SPSD if it is

- symmetric, i.e. $\mathbf{K}_{ij} = \mathbf{K}_{ji}$,
- positive semi-definite: for any column vector $c \in \mathbb{R}^m$, we have $c^T K c \geqslant 0$.

Example 1.7 (Gaussian kernel). For any $\sigma > 0$, a *Gaussian kernel* is defined as $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that

$$K(x, x') \triangleq \exp\left(\frac{-||x - x'||^2}{2\sigma^2} \right), \text{ for all } x, x' \in \mathcal{X}.$$

This is a PDS kernel derived by normalization of the following kernel

$$K'(x, x') = \exp\left(\frac{\langle x, x' \rangle}{\sigma^2} \right) = \sum_{n=0}^{\infty} \frac{1}{n!} \left(\frac{\langle x, x' \rangle}{\sigma^2} \right)^n, \text{ for all } x, x' \in \mathcal{X}.$$

Example 1.8 (Sigmoid kernel). For any $a, b \geqslant 0$, a *Sigmoid kernel* is defined as $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that

$$K(x, x') \triangleq \tanh(a \langle x, x' \rangle + b).$$

This kernel is used in sigmoid perceptrons in neural networks due to its similarity to the sign function.

2 Reproducing Kernel Hilbert Space (RKHS)

Lemma 2.1 (Cauchy-Schwarz inequality for PDS kernel). *Let K be a PDS kernel. Then*

$$K^2(x,x') \leqslant K(x,x)K(x',x') \text{ for all } x,x' \in \mathcal{X}.$$

Proof. We can write the following Gram matrix for samples x,x' and PDS kernel K as

$$\mathbf{K} = \begin{bmatrix} K(x,x) & K(x,x') \\ K(x',x) & K(x',x') \end{bmatrix}.$$

Since K is a PDS Kernel, the Gram matrix $\mathbf{K}$ is symmetric and positive semi-definite. In particular, $K(x,x') = K(x',x)$ and the $\det(\mathbf{K}) \geqslant 0$. Hence, the result follows. □

Theorem 2.2 (RKHS). *Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a PDS kernel. Then, there exists a Hilbert space $\mathbb{H}$ and a mapping $\Phi : \mathcal{X} \to \mathbb{H}$ such that for all $x,x' \in \mathcal{X}$,*

$$K(x,x') = \left\langle \Phi(x), \Phi(x') \right\rangle_{\mathbb{H}}.$$

Furthermore, $\mathbb{H}$ has the following reproducing property, for all $h \in \mathbb{H}$ and $x \in \mathcal{X}$,

$$h(x) = \langle (h(\cdot), K(x,\cdot)) \rangle_{\mathbb{H}}.$$

The Hilbert space $\mathbb{H}$ is called the RKHS associated with the kernel K.

Remark 5. We make the following observations from the Theorem statement.

1. The Hilbert space $\mathbb{H} \subseteq \mathbb{R}^{\mathcal{X}}$.
2. For any $x \in \mathcal{X}$, we have $K(x,\cdot) \in \mathbb{H}$.

Proof. For any $x \in \mathcal{X}$, define $\Phi_x : \mathcal{X} \to \mathbb{R}$ such that $\Phi_x(x') = K(x,x')$. Let us take $\mathbb{H}_0$, the span of kernel evaluations at finitely many elements of $\mathcal{X}$. That is,

$$\mathbb{H}_0 \triangleq \left\{ \sum_{i \in I} a_i \Phi_{x_i} : I \text{ finite }, a_i \in \mathbb{R}, x_i \in \mathcal{X}, \text{ for each } i \in I \right\}.$$

Then, we define a map $\langle \cdot,\cdot \rangle : \mathbb{H}_0 \times \mathbb{H}_0 \to \mathbb{R}$ such that fo $f = \sum_{i \in I} a_i \Phi_{x_i}$ and $g = \sum_{j \in J} b_j \Phi_{x_j}$, we have

$$\langle f,g \rangle_{\mathbb{H}_0} \triangleq \sum_{i \in I} \sum_{j \in J} a_i b_j K(x_i,x_j) = \sum_{j \in J} b_j f(x_j) = \sum_{i \in I} a_i g(x_i).$$

We can verify that the $\langle \cdot,\cdot \rangle : \mathbb{H}_0 \times \mathbb{H}_0 \to \mathbb{R}$ has the follow properties.

1. **Symmetry**: By definition, $\langle \cdot,\cdot \rangle$ is symmetric.
2. **Bilinearity**: $\langle \cdot,\cdot \rangle$ is bilinear. Can you show that $\langle \alpha f + \beta h, g \rangle = \alpha \langle f,g \rangle + \beta \langle f,g \rangle$?
3. **Positive semi-definiteness**: For any $f \in \mathbb{H}_0$, we have $f = \sum_{i \in I} a_i \Phi_{x_i}$ and since the Gram matrix $\mathbf{K}$ is symmetric and positive semidefinite for kernel K and samples $S = (x_i : i \in I)$, we have

$$\langle f,f \rangle = \sum_{i \in I} \sum_{j \in I} a_i a_j K(x_i,x_j) = a^T \mathbf{K} a \geqslant 0.$$

4. **Reproducing property:** Let $f \in \mathbb{H}_0$ and $f = \sum_{i \in I} a_i \Phi_{x_i}$. Then,

$$\langle f, \Phi_x \rangle = \sum_{i \in I} a_i K(x_i,x) = \sum_{i \in I} a_i \Phi_{x_i}(x) = f(x).$$

5. **Definiteness**: We will show that for any $f \in \mathbb{H}_0$ and $x \in \mathcal{X}$, we have bounded $f(x)$. From the reproducing property, it suffices to show that $\langle f, \Phi_x \rangle^2 \leqslant \langle f,f \rangle \langle \Phi_x, \Phi_x \rangle$ for any $x \in \mathcal{X}$. Can you show that $\langle \cdot,\cdot \rangle$ is a PDS kernel? Then the result will follow from Lemma 2.1.
6. From properties $1,2,3,5$, it follows that $\mathbb{H}_0$ is a pre-Hilbert space which can be made complete to form the Hilbert space $\mathbb{H} = \overline{\mathbb{H}_0}$, where $\mathbb{H}_0$ is dense in $\mathbb{H}$. This Hilbert space $\mathbb{H}$ is the RKHS associated with the kernel K.

□