Lecture-05: PDS Kernels

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 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 $N = 2$ 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 $x,x' \in \mathbb{R}^2$. For general N and d, can you find the dimension of $\mathbb{H}$ for the $\Phi : \mathcal{X} \to \mathbb{H}$ corresponding to the Kernel function?

Example 1.4. 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 $h(x_1,x_2) = x_1x_2$ to bring these points to a higher-dimensional space, we find that these points are indeed separable along the x_1x_2 dimension.

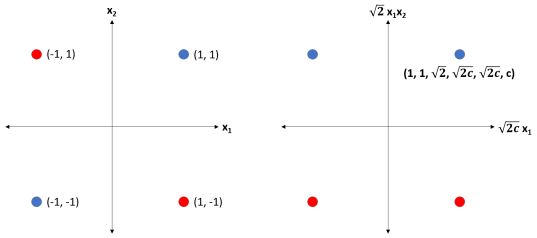

Figure 1: Left: Four points from two classes plotted on the x_1, x_2 axes. These points are not separable by any hyperplane. Right: The same four points are plotted on the $\sqrt{2}x_1x_2$ and $\sqrt{2c}x_1$ axes. These points are now separable.

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.

2 PDS Kernels

Theorem 2.1 (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

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

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 $\mathcal{X}$.

Definition 2.2 (Gram matrix). For a sample $x \in \mathcal{X}^m$, the **kernel matrix** or the **Gram matrix** associated to the kernel K and the sample x is denoted by $\mathbf{K} \in \mathbb{R}^{m\times m}$ and given by

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

Definition 2.3 (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
(i) symmetric, i.e. $\mathbf{K}_{ij} = \mathbf{K}_{ji}$,
(ii) positive semi-definite: for any column vector $c \in \mathbb{R}^m$, we have $c^T\mathbf{K}c \geqslant 0$.

Example 2.4 (Inner product). Consider kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by the inner product $K(x,y) \triangleq \langle x,y\rangle$ for all $x,y \in \mathcal{X}$. For any unlabeled training sample $x \in \mathcal{X}^m$, we denote the corresponding gram matrix by $\mathbf{K}$. We observe that $\mathbf{K}$ is symmetric, since the inner product is symmetric. That is,

$$\mathbf{K}_{ij} = \langle x_i,x_j\rangle = \langle x_j,x_i\rangle = \mathbf{K}_{ji}.$$

Further, we observe that $\mathbf{K}$ is positive semi definite, since for any $c \in \mathbb{R}^m$, we have

$$\langle c,\mathbf{K}c\rangle = \sum_{i,j=1}^{m} c_j\langle x_i,x_j\rangle c_j = \left\langle \sum_{i=1}^{m} c_i x_i, \sum_{j=1}^{m} c_j x_j \right\rangle = \left\| \sum_{i=1}^{m} c_i x_i \right\|_2^2 \geqslant 0$$

Since the gram matrix $\mathbf{K}$ is SPSD for any sample x, it follows that the kernel K is PDS.

Definition 2.5 (Normalized kernels). To any kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we can associate a **normalized kernel** $K' : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined for all $x,y \in \mathcal{X}$ by

$$K'(x,y) = \begin{cases} \frac{K(x,y)}{\sqrt{K(x,x)K(y,y)}}, & K(x,x)K(y,y) \neq 0, \\ 0, & K(x,x)K(y,y) = 0. \end{cases}$$

Remark 5. For any $x \in \mathcal{X}$ such that $K(x,x) \neq 0$, we have $K'(x,x) = 1$. For any PDS kernel, we have $|K'| \leqslant 1$.

Example 2.6 (Gaussian kernel). For $\sigma > 0$, let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be defined as $K(x,y) = \exp\left(\frac{\langle x,y\rangle}{\sigma^2}\right)$. The normalized kernel associated with this kernel is the **Gaussian kernel** $K' : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with parameter $\sigma > 0$, defined for all $x, y \in \mathcal{X}$ as

$$K'(x,y) = \exp\left(\frac{1}{2\sigma^2}\left(2\langle x,y\rangle - \|x\|^2 - \|y\|^2\right)\right) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right).$$

Lemma 2.7 (Normalized PDS kernels). *Let K be a PDS kernel. Then, the normalized kernel K' associated to K is PDS.*

Proof. Consider an m-sized unlabeled training sample $x \in \mathcal{X}^m$. We will show that the gram matrix $\mathbf{K}'$ generated by the sample x and kernel K' is SPSD. Symmetry of K' follows from the symmetry of K, and hence the gram matrix $\mathbf{K}'$ is symmetric. To see the positive semi-definiteness of the gram matrix $\mathbf{K}'$, we note that its (i,j)-th entry $\mathbf{K}'_{ij} = K'(x_i,x_j) = \frac{\langle \Phi(x_i), \Phi(x_j)\rangle_{\mathbb{H}}}{\|\Phi(x_i)\|_{\mathbb{H}} \left\|\Phi(x_j)\right\|_{\mathbb{H}}}$. Hence, for any vector $c \in \mathbb{R}^m$, we have $c^T \mathbf{K}' c = \left\|\sum_{i=1}^m c_i \frac{\Phi(x_i)}{\|\Phi(x_i)\|_{\mathbb{H}}}\right\|_{\mathbb{H}}^2 \geqslant 0$. □

3 Closure Properties

Definition 3.1 (Tensor product). The **tensor product** of two kernels K_1, K_2 is denoted by $K_1 \otimes K_2 : \mathcal{X}^4 \to \mathbb{R}$ and defined as $(K_1 \otimes K_2)(x_1,x_2,y_1,y_2) = K_1(x_1,y_1)K_2(x_2,y_2)$ for all $x_1,y_1,x_2,y_2 \in \mathcal{X}$.

Theorem 3.2 (Closure properties of PDS kernels). *PDS kernels are closed under sum, product, tensor product, point-wise limit, and composition with a power series $\sum_{n=0}^{\infty} a_n x^n$ with $a_n \geqslant 0$ for all $n \in \mathbb{N}$.*

Proof. Let $(K_n : n \in \mathbb{N})$ be a sequence of PDS kernels on $\mathbb{R}^{\mathcal{X}\times\mathcal{X}}$, and let $\mathbf{K}_n$ be the gram matrix generated by a sample $x \in \mathcal{X}^m$ for the kernel K_n for each $n \in \mathbb{N}$.

(i) It suffices to show that $\mathbf{K}_1 + \mathbf{K}_2$ is SPSD. Since $\mathbf{K}_1, \mathbf{K}_2$ are SPSD, it follows that $\mathbf{K}_1 + \mathbf{K}_2$ is symmetric. From the linearity of inner products and positive semi definiteness of $\mathbf{K}_1, \mathbf{K}_2$, we have $\langle c, (\mathbf{K}_1 + \mathbf{K}_2)c\rangle = \langle c, \mathbf{K}_1 c\rangle + \langle c, \mathbf{K}_2 c\rangle \geqslant 0$ for any $c \in \mathbb{R}^m$.

(ii) It suffices to show that the matrix $\mathbf{K}_{ij} = [(\mathbf{K}_1)_{ij}(\mathbf{K}_2)_{ij}]$ is SPSD. Symmetry follows from the symmetry of SPSD matrices $\mathbf{K}_1$ and $\mathbf{K}_2$.

Since $\mathbf{K}_1$ is SPSD, we have $\mathbf{K}_1 = \mathbf{M}\mathbf{M}^T$ by singular value decomposition or Cholesky decomposition. Therefore, $(\mathbf{K}_1)_{ij}(\mathbf{K}_2)_{ij} = \sum_{k=1}^m \mathbf{M}_{ik}\mathbf{M}_{jk}(\mathbf{K}_2)_{ij}$ and hence for any $c \in \mathbb{R}^m$, we can write

$$\sum_{i,j=1}^m c_i c_j \Big(\sum_{k=1}^m \mathbf{M}_{ik}\mathbf{M}_{jk}\Big)(\mathbf{K}_2)_{ij} = \sum_{k=1}^m \sum_{i,j=1}^m (c_i\mathbf{M}_{ik})(\mathbf{K}_2)_{ij}(c_j\mathbf{M}_{jk}).$$

Defining $z_k = (c_i\mathbf{M}_{ik} : i \in [m])$, we see that $c^T\mathbf{K}c = \sum_{k=1}^m z_k^T \mathbf{K}_2 z_k \geqslant 0$.

(iii) The tensor product of two kernels K_1, K_2 can be thought of as the product of two PDS kernels

$$(x_1,x_2,y_1,y_2) \mapsto K_1(x_1,y_1), \qquad (x_1,x_2,y_1,y_2) \mapsto K_2(x_2,y_2).$$

(iv) Let K be the point-wise limit of the sequence of PDS kernels $(K_n : n \in \mathbb{N})$. Let $\mathbf{K}$ be the gram matrix generated by the map K and the sample $x \in \mathcal{X}^m$. Symmetry of $\mathbf{K}$ follows from the symmetry of each $\mathbf{K}_n$. From the continuity of inner products, we have $\langle c, \mathbf{K}c\rangle = \lim_n \langle c, \mathbf{K}_n c\rangle \geqslant 0$ for any $c \in \mathbb{R}^m$.

(v) Let's assume that K is a PDS kernel with $|K(x,y)| < \rho$ for all $x, y \in \mathcal{X}$, and let $f : x \mapsto \sum_{n=0}^{\infty} a_n x^n$, be a power series with $a_n \geqslant 0$ and radius of convergence ρ. Then, for any $n \in \mathbb{N}$, both K^n and thus $a_n K^n$ are PDS by closure under product. For any $N \in \mathbb{N}$, the sum $\sum_{n=0}^N a_n K^n$ is PDS by closure under sum of PDS kernels $(a_n K^n : n \geqslant 0)$ and $f \circ K$ is PDS by closure under the limit of $\sum_{n=0}^N a_n K^n$ as $N \to \infty$.

□

Example 3.3 (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 3.4 (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.

Example 3.5 (Gaussian kernels). For any PDS kernel K, the kernel $\exp(K)$ is also PDS since it can be written as a power series with an infinite radius of convergence. We can check that a kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by $K(x,y) = \langle x,y\rangle$ for all $x,y \in \mathcal{X}$ is PDS kernel, and hence $K' = \exp(K)$ defined by $K'(x,y) = \exp\left(\frac{\langle x,y\rangle}{\sigma^2}\right)$ for all $x,y \in \mathcal{X}$ is PDS kernel. Therefore, the Gaussian kernel is PDS since it is normalized kernel of K'.