# Machine learning theory

## Kernel methods

Hamid Beigy

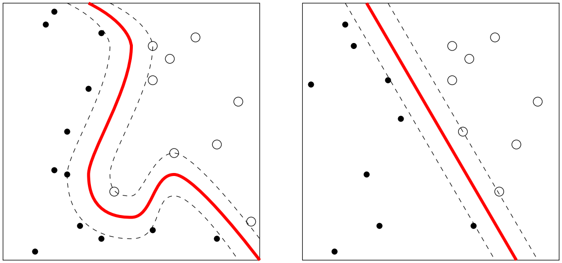
Sharif university of technology

May 6, 2023

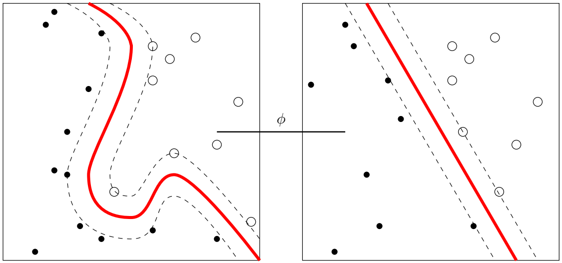
## Table of contents

# Motivation

1. Most of learning algorithms are linear and are not able to classify non-linearly-separable data.
2. How do you separate these two classes?



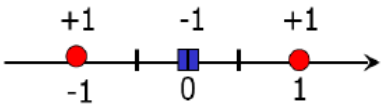
3. Linear separation impossible in most problems.
4. Non-linear mapping from input space to high-dimensional feature space: $\phi : \mathcal{X} \mapsto \mathbb{H}$.



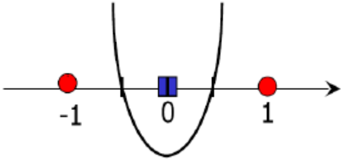

5. Generalization ability: independent of $dim(\mathbb{H})$, depends only on $\rho$ and $m$.

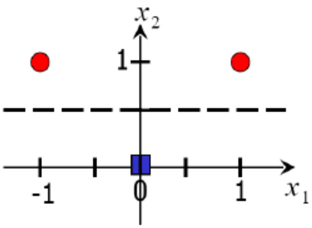
# Kernel methods

1. Most datasets are not linearly separable, for example



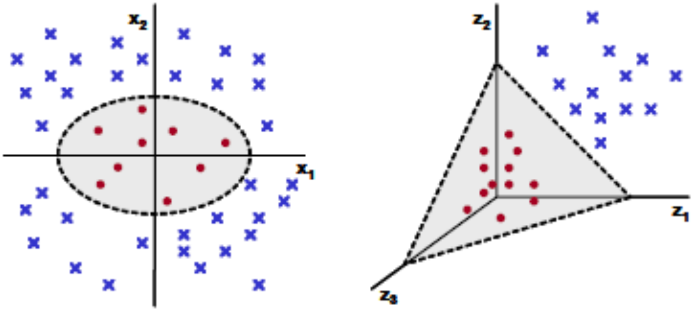

2. Instances that are not linearly separable in $\mathbb{R}$, may be linearly separable in $\mathbb{R}^2$ by using mapping $\phi(x) = (x, x^2)$.




3. In this case, we have two solutions
   - Increase dimensionality of data set by introducing mapping $\phi$.
   - Use a more complex model for classifier.

1. To classify the non-linearly separable dataset, we use mapping $\phi$.
2. For example, let $\mathbf{x} = (x_1, x_2)^T$, $\mathbf{z} = (z_1, z_2.z_3)^T$, and $\phi : \mathbb{R}^2 \to \mathbb{R}^3$.
3. If we use mapping $\mathbf{z} = \phi(x) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)^T$, the dataset will be linearly separable in $\mathbb{R}^3$.




4. Mapping dataset to higher dimensions has two major problems.
   - In high dimensions, there is risk of over-fitting.
   - In high dimensions, we have more computational cost.
5. The generalization capability in higher dimension is ensured by using large margin classifiers.
6. The mapping is an implicit mapping not explicit.

1. Kernel methods avoid explicitly transforming each point $\mathbf{x}$ in the input space into the mapped point $\phi(\mathbf{x})$ in the feature space.
2. Instead, the inputs are represented via their $m \times m$ pairwise similarity values.
3. The similarity function, called a **kernel**, is chosen so that it represents a dot product in some high-dimensional feature space.
4. The kernel can be computed without directly constructing $\phi$.
5. The pairwise similarity values between points in $S$ represented via the $m \times m$ kernel matrix, defined as

$$\mathbf{K} = \begin{pmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & k(\mathbf{x}_1, \mathbf{x}_2) & \cdots & k(\mathbf{x}_1, \mathbf{x}_m) \\ k(\mathbf{x}_2, \mathbf{x}_1) & k(\mathbf{x}_2, \mathbf{x}_2) & \cdots & k(\mathbf{x}_2, \mathbf{x}_m) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{x}_m, \mathbf{x}_1) & k(\mathbf{x}_m, \mathbf{x}_2) & \cdots & k(\mathbf{x}_m, \mathbf{x}_m) \end{pmatrix}$$

6. Function $K(\mathbf{x}_i, \mathbf{x}_j)$ is called kernel function and defined as

**Definition (Kernel)**

Function $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a kernel if

6.1 $\exists \phi : \mathcal{X} \mapsto \mathbb{R}^N$ such that $K(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$.

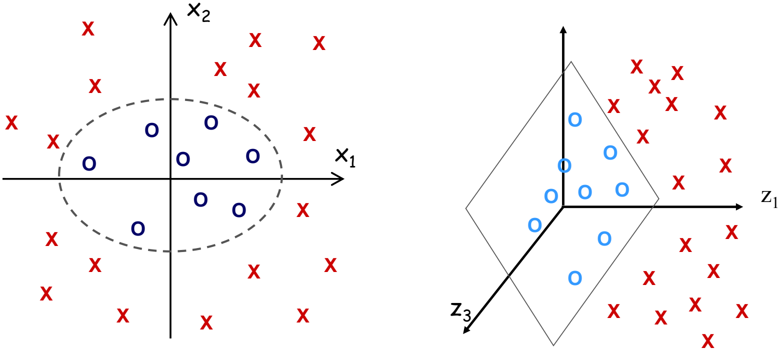
6.2 Range of $\phi$ is called the feature space.

6.3 $N$ can be very large.

1. Let $\phi : \mathbb{R}^2 \mapsto \mathbb{R}^3$ be defined as $\phi(\mathbf{x}) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$.
2. Then $\langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle$ equals to

$$\begin{aligned}\langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle &= \left\langle (x_1^2, x_2^2, \sqrt{2}x_1x_2), (z_1^2, z_2^2, \sqrt{2}z_1z_2) \right\rangle \\ &= (x_1z_1 + x_2z_2)^2 \\ &= (\langle \mathbf{x}, \mathbf{z} \rangle)^2 \\ &= K(\mathbf{x}, \mathbf{z}).\end{aligned}$$

3. The above mapping can be described




Input space feature space

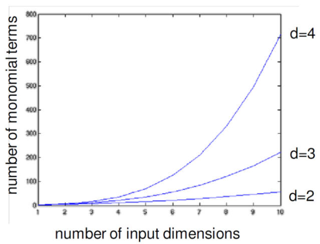
1. Let $\phi_1 : \mathbb{R}^2 \mapsto \mathbb{R}^3$ be defined as $\phi(\mathbf{x}) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$.
2. Then $\langle \phi_1(\mathbf{x}), \phi_1(\mathbf{z}) \rangle$ equals to

$$\begin{aligned}\langle \phi_1(\mathbf{x}), \phi_1(\mathbf{z}) \rangle &= \left\langle (x_1^2, x_2^2, \sqrt{2}x_1x_2), (z_1^2, z_2^2, \sqrt{2}z_1z_2) \right\rangle \\ &= (x_1z_1 + x_2z_2)^2 \\ &= (\langle \mathbf{x}, \mathbf{z} \rangle)^2 = K(\mathbf{x}, \mathbf{z}).\end{aligned}$$

3. Let $\phi_2 : \mathbb{R}^2 \mapsto \mathbb{R}^4$ be defined as $\phi(x) = (x_1^2, x_2^2, x_1x_2, x_2x_1)$.
4. Then $\langle \phi_2(\mathbf{x}), \phi_2(\mathbf{z}) \rangle$ equals to

$$\begin{aligned}\langle \phi_2(\mathbf{x}), \phi_2(\mathbf{z}) \rangle &= \left\langle (x_1^2, x_2^2, x_1x_2, x_2x_1), (z_1^2, z_2^2, z_1z_2, z_2z_1) \right\rangle \\ &= (\langle \mathbf{x}, \mathbf{z} \rangle)^2 = K(\mathbf{x}, \mathbf{z}).\end{aligned}$$

5. Feature space can grow really large and really quickly.
6. Let $K$ be a kernel $\mathbf{K}(\mathbf{x}, \mathbf{z}) = (\langle \mathbf{x}, \mathbf{z} \rangle)^d = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle$
7. The dimension of feature space equals to $\binom{d+n-1}{d}$.
8. Let $n = 100$, $d = 6$, there are 1.6 billion terms.



1. The kernel methods have the following benefits.
   **Efficiency:** $K$ is often more efficient to compute than $\phi$ and the dot product.
   **Flexibility:** $K$ can be chosen arbitrarily so long as the existence of $\phi$ is guaranteed (Mercer's condition).

**Theorem (Mercer's condition)**

*For all functions $c$ that are square integrable (i.e., $\int c(x)^2 dx < \infty$), other than the zero function, the following property holds:*

$$\int \int c(x) K(x, z) c(z) dx dz \geq 0.$$

2. This Theorem states that $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a kernel if matrix $\mathbf{K}$ is positive semi-definite (PSD).
3. Suppose $\mathbf{x}, \mathbf{z} \in \mathbb{R}^n$ and consider the following kernel

$$K(\mathbf{x}, \mathbf{z}) = (\langle \mathbf{x}, \mathbf{z} \rangle)^2$$

4. It is a valid kernel because

$$\begin{aligned} K(\mathbf{x}, \mathbf{z}) &= \left( \sum_{i=1}^{n} x_i z_i \right) \left( \sum_{j=1}^{n} x_j z_j \right) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i x_j)(z_i z_j) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle \end{aligned}$$

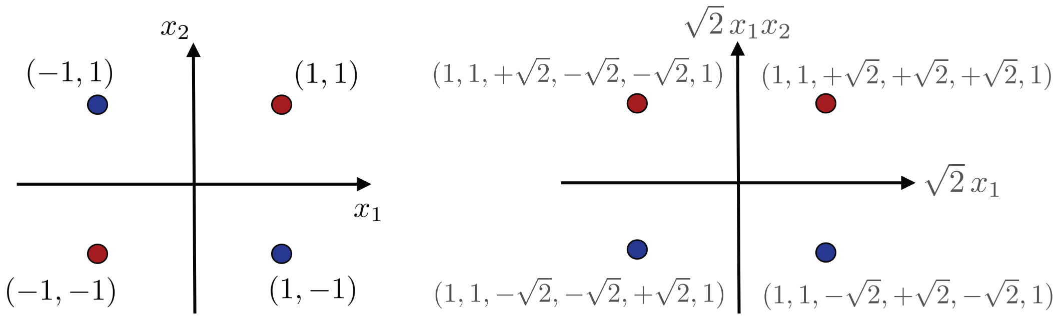
   where the mapping $\phi$ for $n = 2$ is

$$\phi(\mathbf{x}) = (x_1 x_1, x_1 x_2, x_2 x_1, x_2 x_2)^T$$

1. Consider the polynomial kernel $K(x, z) = (\langle \mathbf{x}, \mathbf{z} \rangle + c)^d$ for all $\mathbf{x}, \mathbf{z} \in \mathbb{R}^n$.
2. For $n = 2$ and $d = 2$,

$$\begin{aligned} K(\mathbf{x}, \mathbf{z}) &= (x_1 z_1 + x_2 y_2 + c)^2 \\ &= \left\langle \left[x_1^2, x_2^2, \sqrt{2} x_1 x_2, \sqrt{2} c x_1, \sqrt{2} c x_2, c\right], \left[z_1^2, z_2^2, \sqrt{2} z_1 z_2, \sqrt{2} c z_1, \sqrt{2} c z_2, c\right] \right\rangle \end{aligned}$$

3. Using second-degree polynomial kernel with $c = 1$:




4. The left data is not linearly separable but the right one is.

## Some valid kernels

1. Some valid kernel functions
   - **Polynomial kernels** consider the kernel defined by

     $$K(\mathbf{x}, \mathbf{z}) = (\langle \mathbf{x}, \mathbf{z} \rangle + c)^d$$

     $d$ is the degree of the polynomial and specified by the user and $c$ is a constant.
   - **Radial basis function kernels** consider the kernel defined by

     $$K(\mathbf{x}, \mathbf{z}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{z}\|^2}{2\sigma^2}\right)$$

     The width $\sigma$ is specified by the user. This kernel corresponds to an infinite dimensional mapping $\phi$ .
   - **Sigmoid kernel** consider the kernel defined by

     $$K(\mathbf{x}, \mathbf{z}) = \tanh\left(\beta_0 \langle \mathbf{x}, \mathbf{z} \rangle + \beta_1\right)$$

     This kernel only meets Mercer's condition for certain values of $\beta_0$ and $\beta_1$.
2. **Homework:** Please prove VC-dimension of the above kernels.

We give the crucial property of PDS kernels, which is to induce an inner product in a Hilbert space.

**Lemma (Cauchy-Schwarz inequality for PDS kernels)**

*Let* $\mathbf{K}$ *be a PDS kernel matrix. Then, for any* $\mathbf{x}, \mathbf{z} \in \mathcal{X}$,

$$K(\mathbf{x}, \mathbf{z})^2 \leq K(\mathbf{x}, \mathbf{x})K(\mathbf{z}, \mathbf{z})$$

**Proof.**

1. Consider the kernel matrx $\mathbf{K} = \begin{pmatrix} K(x,x) & K(x,x') \\ K(x',x) & K(x'x') \end{pmatrix}$.
2. By definition, if $K$ is PDS, then $\mathbf{K}$ is SPSD for all $x, x' \in \mathcal{X}$.
3. Then, the product of the eigenvalues of $\mathbf{K}$, $\det(\mathbf{K})$, must be non-negative.
4. Using $K(x,x') = K(x',x)$, we have $\det(\mathbf{K}) = K(x,x)K(x'x') - K(x,x')^2 \geq 0$.

□

**Theorem (Reproducing kernel Hilbert space (RKHS))**

*Let* $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ *be a PDS kernel. Then, there exists a Hilbert space* $\mathbb{H}$ *and a mapping* $\phi : \mathcal{X} \mapsto \mathbb{H}$ *such that for all* $\mathbf{x}, \mathbf{y} \in \mathcal{X}$

$$K(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle .$$

This Theorem implies that PDS kernels can be used to implicitly define a feature space.

For any kernel $\mathbf{K}$, we can associate a normalized kernel $\mathbf{K}_n$ defined by

$$K_n(\mathbf{x}, \mathbf{z}) = \begin{cases} 0 & \text{if } ((K(\mathbf{x}, \mathbf{x}) = 0) \vee (K(\mathbf{z}, \mathbf{z}) = 0)) \\ \dfrac{K(\mathbf{x}, \mathbf{z})}{\sqrt{K(\mathbf{x}, \mathbf{x})K(\mathbf{z}, \mathbf{z})}} & \text{otherwise} \end{cases}$$

**Lemma (Normalized PDS kernels)**

*Let $K$ be a PDS kernel. Then, the normalized kernel $K_n$ associated to $K$ is PDS.*

**Proof.**

1. Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \subseteq \mathcal{X}$ and let $\mathbf{c}$ be an arbitrary vector in $\mathbb{R}^n$.
2. We will show that $\sum_{i,j=1}^{m} c_i c_j K_n(\mathbf{x}_i, \mathbf{x}_j) \geq 0$.
3. By Lemma Cauchy-Schwarz inequality for PDS kernels, if $K(\mathbf{x}_i, \mathbf{x}_i) = 0$, then $K(\mathbf{x}_i, \mathbf{x}_j) = 0$ and thus $K_n(\mathbf{x}_i, \mathbf{x}_i) = 0$ for all $j \in \{1, 2, \ldots, m\}$.
4. We can assume that $K(\mathbf{x}_i, \mathbf{x}_i) > 0$ for all $i \in \{1, 2, \ldots, m\}$.
5. Then, the sum can be rewritten as follows:

$$\sum_{i,j=1}^{m} c_i c_j K_n(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i,j=1}^{m} \frac{c_i c_j K(\mathbf{x}_i, \mathbf{x}_j)}{\sqrt{K(\mathbf{x}_i, \mathbf{x}_i)K(\mathbf{x}_j, \mathbf{x}_j)}} = \sum_{i,j=1}^{m} \frac{c_i c_j \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle}{\|\phi(\mathbf{x}_i)\|_{\mathbb{H}} \cdot \|\phi(\mathbf{x}_j)\|_{\mathbb{H}}} = \left\| \sum_{i=1}^{m} \frac{c_i \phi(\mathbf{x}_i)}{\|\phi(\mathbf{x}_i)\|_{\mathbb{H}}} \right\|_{\mathbb{H}}^2 \geq 0.$$

□

The following theorem provides closure guarantees for all of these operations.

**Theorem (Closure properties of PDS kernels)**

*PDS kernels are closed under*

1. *sum*
2. *product*
3. *tensor product*
4. *pointwise limit*
5. *composition with a power series* $\sum_{k=1}^{\infty} a_k x^k$ *with* $a_k \geq 0$ *for all* $k \in \mathbb{N}$.

**Proof.**

We only proof the closeness under sum. Consider two valid kernel matrices $\mathbf{K}_1$ and $\mathbf{K}_2$.

1. For any $\mathbf{c} \in \mathbb{R}^m$, we have $\mathbf{c}^T\mathbf{K}_1\mathbf{c} \geq 0$ and $\mathbf{c}^T\mathbf{K}_2\mathbf{c} \geq 0$.
2. This implies that $\mathbf{c}^T\mathbf{K}_1\mathbf{c} + \mathbf{c}^T\mathbf{K}_2\mathbf{c} \geq 0$.
3. Hence, we have $\mathbf{c}^T(\mathbf{K}_1 + \mathbf{K}_2)\mathbf{c} \geq 0$.
4. Let $\mathbf{K} = \mathbf{K}_1 + \mathbf{K}_2$, which is a valid kernel.

□

**Homework:** Please prove other closure properties of PDS kernels.

# Basic kernel operations in feature space

1. **Norm of a point:** we can compute the norm of a point $\phi(\mathbf{x})$ in feature space as

$$\|\phi(\mathbf{x})\|^2 = \langle \phi(\mathbf{x}), \phi(x) \rangle = K(\mathbf{x}, \mathbf{x}),$$

which implies that $\|\phi(\mathbf{x})\| = \sqrt{K(\mathbf{x}, \mathbf{x})}$.

2. **Distance between Points:** the distance between two points $\phi(\mathbf{x}_i)$ and $\phi(\mathbf{x}_j)$ can be computed as

$$\begin{aligned} \|\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)\|^2 &= \|\phi(\mathbf{x}_i)\|^2 + \|\phi(\mathbf{x}_j)\|^2 - 2\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle \\ &= K(\mathbf{x}_i, \mathbf{x}_i) + K(\mathbf{x}_j, \mathbf{x}_j) - 2K(\mathbf{x}_i, \mathbf{x}_j), \end{aligned}$$

which implies that $\|\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)\| = \sqrt{K(\mathbf{x}_i, \mathbf{x}_i) + K(\mathbf{x}_j, \mathbf{x}_j) - 2K(\mathbf{x}_i, \mathbf{x}_j)}$.

3. **Mean in feature space:** the mean of the points in feature space is given as

$$\mu_\phi = \frac{1}{m} \sum_{i=1}^{m} \phi(\mathbf{x}_i).$$

Since we haven't access to $\phi(\mathbf{x})$, we cannot explicitly compute the mean point in feature space but we can compute the squared norm of the mean as follows.

$$\begin{aligned} \|\mu_\phi\|^2 &= \langle \mu_\phi, \mu_\phi \rangle \\ &= \left\langle \frac{1}{m} \sum_{i=1}^{m} \phi(\mathbf{x}_i), \frac{1}{m} \sum_{i=1}^{m} \phi(\mathbf{x}_i) \right\rangle \\ &= \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} K(\mathbf{x}_i, \mathbf{x}_j). \end{aligned}$$

1. **Total variance in feature space:** the squared distance of a point $\phi(x_i)$ to the mean $\mu_\phi$ in feature space:

$$\begin{aligned}\|\phi(\mathbf{x}) - \mu_\phi\|^2 &= \|\phi(\mathbf{x}_i)\|^2 - 2\langle \phi(\mathbf{x}_i), \mu_\phi\rangle + \|\mu_\phi\|^2 \\ &= K(\mathbf{x}_i, \mathbf{x}_i) - \frac{2}{m}\sum_{j=1}^{m} K(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{m^2}\sum_{a=1}^{m}\sum_{b=1}^{m} K(\mathbf{x}_a, \mathbf{x}_b).\end{aligned}$$

The total variance in feature space is obtained by taking the average squared deviation of points from the mean in feature space

$$\begin{aligned}\sigma_\phi^2 &= \frac{1}{m}\sum_{i=1}^{m} \|\phi(\mathbf{x}_i) - \mu_\phi\|^2 \\ &= \frac{1}{m}\sum_{i=1}^{m}\left(K(\mathbf{x}_i, \mathbf{x}_i) - \frac{2}{m}\sum_{j=1}^{m} K(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{m^2}\sum_{a=1}^{m}\sum_{b=1}^{m} K(\mathbf{x}_a, \mathbf{x}_b)\right) \\ &= \frac{1}{m}\sum_{i=1}^{m} K(\mathbf{x}_i, \mathbf{x}_i) - \frac{2}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} K(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{m^2}\sum_{a=1}^{m}\sum_{b=1}^{m} K(\mathbf{x}_a, \mathbf{x}_b) \\ &= \frac{1}{m}\sum_{i=1}^{m} K(\mathbf{x}_i, \mathbf{x}_i) - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} K(\mathbf{x}_i, \mathbf{x}_j) \\ &= \frac{1}{m}\mathrm{Tr}\left[\mathbf{K}\right] - \|\mu_\phi\|^2.\end{aligned}$$

1. **Centering in feature space:**
   - We can center each point in feature space by subtracting the mean from it

$$\hat{\phi}(\mathbf{x}_i) = \phi(\mathbf{x}_i) - \mu_\phi.$$

   - We have not $\phi(\mathbf{x}_i)$ and $\mu_\phi$, hence, we cannot explicitly center the points.
   - However, we can still compute the centered kernel matrix $\hat{\mathbf{K}}$, that is, the kernel matrix over centered points.

$$\begin{aligned}
\hat{K}(x_i, x_j) &= \left\langle \hat{\phi}(\mathbf{x}_i), \hat{\phi}(\mathbf{x}_j) \right\rangle \\
&= \left\langle \phi(\mathbf{x}_i) - \mu_\phi, \phi(\mathbf{x}_j) - \mu_\phi \right\rangle \\
&= \left\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \right\rangle - \left\langle \phi(\mathbf{x}_i), \mu_\phi \right\rangle - \left\langle \phi(\mathbf{x}_j), \mu_\phi \right\rangle + \left\langle \mu_\phi, \mu_\phi \right\rangle \\
&= K(\mathbf{x}_i, \mathbf{x}_j) - \frac{1}{m}\sum_{k=1}^{m} \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_k) \rangle - \frac{1}{m}\sum_{k=1}^{m} \langle \phi(\mathbf{x}_j), \phi(\mathbf{x}_k) \rangle + \left\|\mu_\phi\right\|^2 \\
&= K(\mathbf{x}_i, \mathbf{x}_j) - \frac{1}{m}\sum_{k=1}^{m} K(\mathbf{x}_i, \mathbf{x}_k) - \frac{1}{m}\sum_{k=1}^{m} K(\mathbf{x}_j, \mathbf{x}_k) + \left\|\mu_\phi\right\|^2
\end{aligned}$$

   - In other words, we can compute the centered kernel matrix using only the kernel function.

1. **Normalizing in feature space:**
   - A common form of normalization is to ensure that points in feature space have unit length by replacing $\phi(\mathbf{x})$ with the corresponding unit vector $\phi_n(\mathbf{x}) = \frac{\phi(\mathbf{x})}{\|\phi(\mathbf{x})\|}$ .
   - The dot product in feature space then corresponds to the cosine of the angle between the two mapped points, because

$$\langle \phi_n(\mathbf{x}_i), \phi_n(\mathbf{x}_j) \rangle = \frac{\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle}{\|\phi(\mathbf{x}_i)\| . \|\phi(\mathbf{x}_j)\|} = \cos\theta.$$

   - If the mapped points are both centered and normalized, then a dot product corresponds to the correlation between the two points in feature space.
   - The normalized kernel function, $K_n$, can be computed using only the kernel function $K$, as

$$K_n(\mathbf{x}_i, \mathbf{x}_j) = \frac{\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle}{\|\phi(\mathbf{x}_i)\| . \|\phi(\mathbf{x}_j)\|} = \frac{K(\mathbf{x}_i, \mathbf{x}_j)}{\sqrt{K(\mathbf{x}_i, \mathbf{x}_i).K(\mathbf{x}_j, \mathbf{x}_j)}}$$

# Kernel-based algorithms

1. The optimization problem for SVM is defined as

$$\textit{Minimize}\frac{1}{2}\left\|\mathbf{w}\right\|^2 \qquad \text{subject to } y_k\left(\langle \mathbf{w}, \mathbf{x}_k\rangle + b\right) \geq 1 \text{ for all } k = 1, 2, \ldots, m$$

2. In order to solve this constrained optimization problem, we use the Lagrangian function

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\left\|\mathbf{w}\right\|^2 - \sum_{k=1}^{m} \alpha_k \left[y_k\left(\langle \mathbf{w}, \mathbf{x}_k\rangle + b\right) - 1\right]$$

   where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)^T$.

3. Eliminating $\mathbf{w}$ and $b$ from $L(\mathbf{w}, b, a)$ using these conditions then gives the dual representation of the problem in which we maximize

$$\psi(\alpha) = \sum_{k=1}^{m} \alpha_k - \frac{1}{2}\sum_{k=1}^{m}\sum_{j=1}^{m} \alpha_k \alpha_j y_k y_j \langle \mathbf{x}_k, \mathbf{x}_j\rangle$$

4. We need to maximize $\psi(\alpha)$ subject to constraints $\sum_{k=1}^{m} \alpha_k y_k = 0$ and $\alpha_k \geq 0$ $\forall k$.
5. For optimal $\alpha_k$'s, we have $\alpha_k\left[1 - y_k\left(\langle \mathbf{w}, \mathbf{x}_k\rangle + b\right)\right] = 0$.
6. To classify a data $\mathbf{x}$ using the trained model, we evaluate the following function

$$h(\mathbf{x}) = \operatorname{sgn}\left(\sum_{k=1}^{m} \alpha_k y_k \langle \mathbf{x}_k, \mathbf{x}\rangle\right)$$

7. This solution depends on the dot-product between two pints $\mathbf{x}_k$ and $\mathbf{x}$.

1. By using kernel $K$, the dual representation of the problem in which we maximize

$$\psi(\alpha) = \sum_{k=1}^{m} \alpha_k - \frac{1}{2} \sum_{k=1}^{m} \sum_{j=1}^{m} \alpha_k \alpha_j y_k y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

2. To classify a data $x$ using the trained model, we evaluate the following function

$$h(\mathbf{x}) = \text{sgn}\left( \sum_{k=1}^{m} \alpha_k y_k K(\mathbf{x}_k, \mathbf{x}) \right)$$

3. This solution depends on the dot-product between two pints $\mathbf{x}_k$ and $\mathbf{x}$.

**Theorem (Rademacher complexity of kernel-based hypotheses)**

*Let* $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ *be a PDS kernel and let* $\phi : \mathcal{X} \mapsto \mathbb{H}$ *be a feature mapping associated to* $K$*. Let also* $S \subseteq \{\mathbf{x} \mid \mathbf{K}(\mathbf{x}, \mathbf{x}) \leq r^2\}$ *be a sample of size* $m$ *and let* $H = \{\mathbf{x} \mapsto \langle \mathbf{w}, \phi(\mathbf{x})\rangle \mid \|\mathbf{w}\|_{\mathbb{H}} \leq \Lambda\}$ *for some* $\Lambda \geq 0$*. Then*

$$\hat{\mathcal{R}}_S(H) \leq \frac{\Lambda\sqrt{\operatorname{Tr}[\mathbf{K}]}}{m} \leq \sqrt{\frac{r^2\Lambda^2}{m}}.$$

**Proof.**

$$\begin{aligned}
\hat{\mathcal{R}}_S(H) &= \frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{\|\mathbf{w}\|\leq\Lambda}\sum_{i=1}^{m}\sigma_i\langle\mathbf{w},\phi(\mathbf{x}_i)\rangle\right] = \frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{\|\mathbf{w}\|\leq\Lambda}\left\langle\mathbf{w},\sum_{i=1}^{m}\sigma_i\phi(\mathbf{x}_i)\right\rangle\right] \\
&\leq \frac{\Lambda}{m}\mathop{\mathbb{E}}_{\sigma}\left[\left\|\sum_{i=1}^{m}\sigma_i\phi(\mathbf{x}_i)\right\|_{\mathbb{H}}\right] \leq \frac{\Lambda}{m}\sqrt{\mathop{\mathbb{E}}_{\sigma}\left[\left\|\sum_{i=1}^{m}\sigma_i\phi(\mathbf{x}_i)\right\|_{\mathbb{H}}^2\right]} = \frac{\Lambda}{m}\sqrt{\mathop{\mathbb{E}}_{\sigma}\left[\sum_{i,j=1}^{m}\sigma_i\sigma_j\langle\phi(\mathbf{x}_i),\phi(\mathbf{x}_j)\rangle\right]} \\
&\leq \frac{\Lambda}{m}\sqrt{\mathop{\mathbb{E}}_{\sigma}\left[\sum_{i=1}^{m}\|\phi(\mathbf{x}_i)\|^2\right]} = \frac{\Lambda}{m}\sqrt{\mathop{\mathbb{E}}_{\sigma}\left[\sum_{i=1}^{m}\mathbf{K}(\mathbf{x}_i,\mathbf{x}_i)\right]} \\
&\leq \frac{\Lambda\sqrt{\operatorname{Tr}[\mathbf{K}]}}{m} = \sqrt{\frac{r^2\Lambda^2}{m}}
\end{aligned}$$

□

**Theorem (Margin bounds for kernel-based hypotheses)**

*Let* $\mathbf{K} : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ *be a PDS kernel with* $r^2 = \sup_{\mathbf{x} \in \mathcal{X}} \mathbf{K}(\mathbf{x}, \mathbf{x})$*. Let* $\phi : \mathcal{X} \mapsto \mathbb{H}$ *be a feature mapping associated to* $\mathbf{K}$ *and let* $H = \{x \mapsto \langle \mathbf{w}, \phi(\mathbf{x}) \rangle \mid \|\mathbf{w}\|_{\mathbb{H}} \leq \Lambda\}$ *for some* $\Lambda \geq 0$*. Fix* $\rho > 0$*. Then for any* $\delta > 0$*, each of the following statements holds with probability at least* $(1 - \delta)$ *for any* $h \in H$*:*

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}_{S,\rho}(h) + 2\sqrt{\frac{r^2\Lambda^2/\rho^2}{m}} + \sqrt{\frac{\log(1/\delta)}{2m}}$$

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}_{S,\rho}(h) + 2\sqrt{\frac{\mathrm{Tr}\,[\mathbf{K}]\Lambda^2/\rho^2}{m}} + 3\sqrt{\frac{\log(2/\delta)}{2m}}$$

1. Chapter 16 of Shai Shalev-Shwartz and Shai Ben-David Book[1]
2. Chapter 6 of Mehryar Mohri and Afshin Rostamizadeh and Ameet Talwalkar Book[2].

[1] Shai Shalev-Shwartz and Shai Ben-David (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

[2] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar (2018). *Foundations of Machine Learning*. Second Edition. MIT Press.

# References

Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar (2018). *Foundations of Machine Learning*. Second Edition. MIT Press.

Shalev-Shwartz, Shai and Shai Ben-David (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

**Questions?**