MATH 6397 - Mathematics of Data Science

Instructor: Demetrio Labate

April 11, 2023

Course Outline

1. Mathematics of machine learning

    1.1 Support Vector Machines

References:

□ *The Mathematics of Signal Processing*, by Damelin and Miller

□ *Foundations of Data Science*, by Blum, Hopcroft and Kannan

□ *Foundations of Machine Learning*, by Mohri, Rostamizadeh and Talwalkar

□ *Deep Learning with PyTorch*, by Stevens, Antiga and Viehmann

Support Vector Machines

Support Vector Machines

Support Vector Machines (SVMs) are one of the most theoretically well motivated and most effective classification algorithms in modern machine learning [Boser, Guyon, Vapnik, 1992]

Given a set of labeled training data, each marked as belonging to either one of two categories, an SVM algorithm computes an **optimal hyperplane** that separates the two categories.

The optimal criterion for the hyperplane consists in determining the hyperplane achieving the **widest possible gap** between the two categories.

In addition to performing linear classification, SVMs can also perform a non-linear classification using what is called the **kernel trick** - a method that implicitly maps input data into an appropriate feature space where feature vectors are linearly separable.

Linear classification

Let $X \subset \mathbb{R}^N$, $Y = \{-1, +1\}$ and $f : X \to Y$ be a target function.

Given a hypothesis set H of functions mapping X to Y, the **binary classification task** is formulated as follows.

The learner receives a training sample S of size m drawn i.i.d. from X according to some unknown distribution $\mathcal{D}$

$$S = ((x_1, y_1), \dots, (x_m, y_m)) \in X \times Y$$

with $y_i = f(x_i)$ for all $i \in [m]$.

The problem consists of determining a hypothesis $h \in H$, a binary classifier, with small generalization error:

$$\mathcal{R}_{\mathcal{D}}(h) = \underset{x \sim \mathcal{D}}{P}(h(x) \neq f(x))$$

Linear classification

Different hypothesis sets H can be selected for this binary classification task.

In view of the discussion of the model selection problem presented above, hypothesis sets with smaller complexity (e.g., smaller VC-dimension or Rademacher complexity) provide better learning guarantees, everything else being equal.

A natural hypothesis set with relatively small complexity is that of linear classifiers, or hyperplanes, which can be defined as follows:

$$H = \{x \to \mathrm{sign}(w \cdot x + b) : w \in \mathbb{R}^N, b \in \mathbb{R}\}$$

The learning problem is then referred to as a **linear classification problem.**

Linear classification

The general equation of an hyperplane in $\mathbb{R}^N$ is

$$HP_{w,b} = \{x \in \mathbb{R}^N : w \cdot x + b = 0\},$$

where $w \in \mathbb{R}^N$ and $b \in \mathbb{R}$.
The vector $\frac{w}{\|w\|}$ can be identified with the unit normal vector to the hyperplane and b with the offset or distance of the hyperplane from the origin.

Accordingly, we can define a *decision function* $\text{sign}(w \cdot x + b)$ that takes values in the set $\{-1, +1\}$ depending on x falling on either side of the hyperplane $HP_{w,b}$.

Linear classification

We start by assuming that the training sample S can be linearly separated, that is, we assume the existence of a hyperplane that perfectly separates the training sample into two populations of positively and negatively labeled points.

Definition. *Let* $S = \{(x_i, y_i) \subset X \times Y : i = 1, \ldots, m\}$*. The set* S *is said to be* **linearly separable** *if there exist* $w \in \mathbb{R}^N \setminus \{0\}$ *and* $b \in \mathbb{R}$ *such that*

$$y_i\,(w \cdot x_i + b) > \delta \quad \forall i = 1, \ldots N,$$

for some $\delta > 0$*. In this case,* $HP_{w,b}$ *is said to be a* **separating hyperplane***.*

If S is separable, there are infinitely many separating hyperplanes.

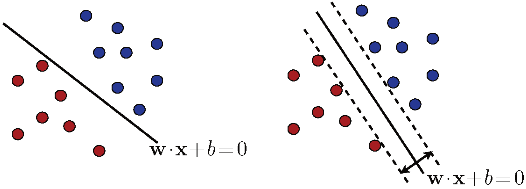
Which hyperplane should a learning algorithm select?

Linear SVM

Among all possible separating hyperplanes, the SVM approach seeks to find the one with the **maximum margin** of separation between any (training) point and the hyperplane.

Definition. *The* **optimal separating hyperplane** *for a set* $S = \{(x_i, y_i) \subset \mathbb{R}^N \times \{-1, +1\} : i = 1, \ldots, m\}$ *is the solution of:*

$$\max_{w \in \mathbb{R}^N, b \in \mathbb{R}} \{\min \|x_i - x\| : x \in \mathbb{R}^N, w \cdot x + b = 0, i = 1, \ldots, m\}$$


$\mathbf{w} \cdot \mathbf{x} + b = 0$
$\mathbf{w} \cdot \mathbf{x} + b = 0$

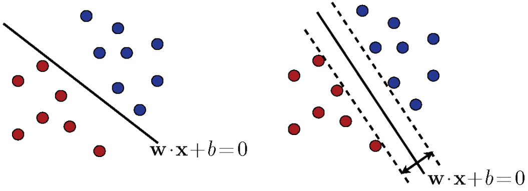
Linear SVM

The **geometric margin** ρ of a linear classifier for a sample $S = (x_1, \ldots, x_m)$ is the minimum distance $d(x_i, HP_{w,b})$ over the points in the sample,

$$\rho = \min_{i \in [m]} d(x_i, HP_{w,b}),$$

that is, the distance of the hyperplane to the closest sample point(s).

The SVM solution is the separating hyperplane with the maximum geometric margin and is thus known as the maximum-margin hyperplane.


$\mathbf{w} \cdot \mathbf{x} + b = 0$
$\mathbf{w} \cdot \mathbf{x} + b = 0$

Linear SVM

It is easy to see that any hyperplane $HP_{w,b}$ can be rescaled by multiplying w and b by the same non-zero constant λ so that

$$HP_{w,b} = HP_{\lambda w, \lambda b}$$

We can remove this unnecessary degree of freedom by rescaling w and b so that the point(s) closest to the hyperplane satisfy $|w \cdot x_i + b| = 1$.

Definition. *The hyperplane $HP_{w,b}$ is said to be in* **canonical form** *with respect to $X = \{x_i \in H : i = 1, \ldots, m\}$ if*

$$\min_{i=1,\ldots m} |w \cdot x_i + b| = 1$$

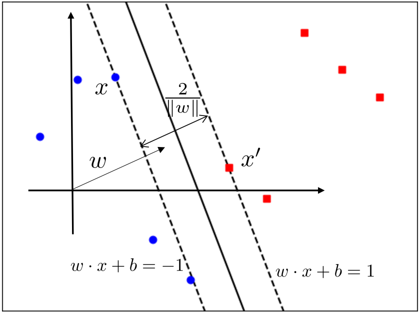
If $HP_{w,b}$ is a canonical hyperplane, then a vector in $x_i \in X$ is said to be a **support vector** *if it belongs to either one of the hyperplanes HP_{-1} or HP_1, where $HP_k := \{x \in H : w \cdot x + b = k\}$.*

Linear SVM

If $x \in HP_{-1}$ and $x' \in HP_1$ (i.e., x, x' are support vectors) then

$$\begin{aligned} 2 &= |(w \cdot x + b) - (w \cdot x' + b)| \\ &= |w \cdot x - w \cdot x'| \\ &= |w \cdot (x - x')| \implies |\tfrac{w}{\|w\|} \cdot (x - x')| = \tfrac{2}{\|w\|}. \end{aligned}$$

Hence, the distance between HP_{-1} and HP_1 is $\frac{2}{\|w\|}$ and the distance between $HP_{w,b}$ and a support vector is $\frac{1}{\|w\|}$.



Linear SVM

Alternatively, using the distance formula from a point to a plane, one can derive that

$$d(x, H_{w,b}) = \frac{|w \cdot x + b|}{\|w\|}$$

Thus, for a support vector, this distance is $\frac{1}{\|w\|}$.

Linear SVM - Primal optimization problem

We can set up an optimization problem that defines the SVM solution. As above, we assume that S is separable.

The maximum margin of a separating hyperplane is given by

$$\rho = \max_{w,b: y_i(w\cdot x_i+b)\geq 0} \min_{i\in[m]} \frac{|w\cdot x_i + b|}{\|w\|} = \max_{w,b} \min_{i\in[m]} \frac{y_i(w\cdot x_i + b)}{\|w\|}$$

As observed above, we can rescale the hyperplanes, so that it is sufficient to consider the pairs (w, b) such that $\min_{i\in[m]} y_i(w\cdot x_i + b) = 1$. Thus we have

$$\begin{aligned}\rho &= \max_{(w,b):\, \min_{i\in[m]} y_i(w\cdot x_i+b)=1} \frac{1}{\|w\|} \\ &= \max_{(w,b):\, y_i(w\cdot x_i+b)\geq 1,\, i\in[m]} \frac{1}{\|w\|}\end{aligned}$$

Linear SVM - Primal optimization problem

Maximizing $\frac{1}{\|w\|}$ is equivalent to minimizing $\frac{1}{2}\|w\|^2$.
Thus, the pair (w, b) returned by SVM in the separable case is the solution of the following convex optimization problem:

$$\min_{(w,b)\in\mathbb{R}^N\times\mathbb{R}} \tau(w) = \frac{1}{2}\|w\|^2$$

$$\text{subject to:} \quad y_i(w \cdot x_i + b) \geq 1, \quad \forall i = 1, \ldots, m.$$

Remark. The objective function τ is strictly convex. In fact τ is infinitely differentiable and the eigenvalues of the Hessian $\nabla^2\tau(w) = I$ are strictly positive. Also, the constraints are all defined by affine functions.
It follows that the optimization problem admits a unique solution.

Moreover, since the objective function is quadratic and the constraints are affine, this optimization problem is a specific instance of **quadratic programming**, a family of problems extensively studied in optimization that have efficient numerical solutions.

Linear SVM - KKT conditions

We can reformulate the optimization problem using the method of Lagrange multipliers.

We introduce non-negative constants $\alpha_1, \ldots, \alpha_m$ and denote $\alpha = (\alpha_1, \ldots, \alpha_m)^t$.
For $w \in \mathbb{R}^N, b \in \mathbb{R}$, we define the Lagrangian

$$\mathcal{L}(w, b, \alpha) := \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i (y_i (w \cdot x_i + b) - 1).$$

Then the primal optimization problem can be solved as

$$\min_{w \in \mathbb{R}^N, b \in \mathbb{R}} \mathcal{L}(w, b, \alpha)$$

$$\text{subject to } \alpha_i \geq 0, \text{ for all } i = 1, \ldots m.$$

Linear SVM - KKT conditions

The Karush-Kuhn-Tucker (KKT) conditions for the solution of the optimization problem are obtained by setting the gradient of the Lagrangian with respect to the primal variables w and b to zero and by writing the complementarity conditions:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial b} &= -\sum_{i=1}^{m} \alpha_i y_i = 0 & \implies \quad & \sum_{i}^{m} \alpha_i y_i = 0, \\
\frac{\partial \mathcal{L}}{\partial w} &= w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0 & \implies \quad & w = \sum_{i=1}^{m} \alpha_i y_i x_i \\
\alpha_i(y_i(w \cdot x_i + b) - 1) &= 0, \forall i & \implies \quad & \alpha_i = 0 \vee y_i(w \cdot x_i + b) = 1
\end{aligned}$$

Remark. Eq. $w = \sum_{i=1}^{m} \alpha_i y_i x_i$ shows that the solution w is a linear combination of the **support vectors**, that is, those training vectors x_i for which $\alpha_i > 0$.
By the complementarity condition, if $\alpha_i \neq 0$, then $y_i(w \cdot x_i + b) = 1$. Thus the support vectors lie on the marginal hyperplanes $w \cdot x + b = \pm 1$.

Linear SVM - KKT conditions

Support vectors fully define the maximum-margin hyperplane or SVM solution.

This explains the name of the SVM algorithm.

Vectors not lying on the marginal hyperplanes do not affect the definition of these hyperplanes and, in their absence, the solution to the SVM problem remains unchanged.

Note that while the solution w of the SVM problem is unique, the support vectors are not.

In dimension N, $N + 1$ points are sufficient to define a hyperplane. When more than $N + 1$ points lie on a marginal hyperplane, different choices are possible for the $N + 1$ support vectors.

Linear SVM - Dual optimization problem

We derive the dual form of the constrained optimization problem stated above by plugging into the Lagrangian the definition of w in terms of the dual variables.

This yields

$$\mathcal{L} = \frac{1}{2}\|\sum_{i=1}^{m} \alpha_i y_i x_i\|^2 - \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i,j=1}^{m} \alpha_i \alpha_j b + \sum_{i=1}^{m} \alpha_i .$$

which simplifies to

$$\mathcal{L} = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

Linear SVM - Dual optimization problem

By applying the constraint, we obtain the dual optimization problem for SVMs (valid in the separable case)

$$\max_{\alpha_1,\ldots,\alpha_m} \mathcal{L}(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

$$\text{subject to } \sum_{i=1}^{m} \alpha_i y_i = 0 \text{ and } \alpha_i \geq 0, \text{ for all } i = 1, \ldots, m.$$

Remark. The objective function $\mathcal{L}(\alpha)$ is convex. In fact $\mathcal{L}$ is infinitely differentiable and the Hessian $\nabla^2 \mathcal{L}(\alpha) = -A$, where $A = (y_i x_i \cdot y_j x_j)_{i,j}$ is positive semidefinite. Also, the constraints are affine and convex.

Since $\mathcal{L}(\alpha)$ is a quadratic function of α, this dual optimization problem is also a quadratic programming problem, as in the case of the primal optimization and once again both general-purpose and specialized quadratic programming solvers can be used to obtain the solution.

Linear SVM - Dual optimization problem

Since the constraints are affine, strong duality holds.
Thus, the primal and dual problems are equivalent: the solution α of the dual problem can be used directly to determine the hypothesis returned by SVMs:

$$h(x) = \mathrm{sign}(w \cdot x + b) = \mathrm{sign}\left(\sum_{i=1}^{m} \alpha_i y_i (x_i \cdot x) + b\right)$$

Since support vectors lie on the marginal hyperplanes, *for any support vector* x_i, $w \cdot x_i + b = y_i$, and thus b can be obtained via

$$b = y_i - \sum_{j=1}^{m} \alpha_j y_j (x_j \cdot x_i)$$

Linear SVM - Dual optimization problem

Multiplying both sides by $\alpha_i y_i$ and taking the sum leads to

$$\sum_{i=1}^{m} \alpha_i y_i b = \sum_{i=1}^{m} \alpha_i y_i^2 - \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_j \cdot x_i)$$

Since $y_i^2 = 1$, $w = \sum_{i=1}^{m} \alpha_i y_i x_i$ and $\sum_{i=1}^{m} \alpha_i y_i = 0$, we have

$$0 = \sum_{i=1}^{m} \alpha_i - \|w\|^2$$

This implies that the margin can be expressed as

$$\rho^2 = \frac{1}{\|w\|^2} = \frac{1}{\sum_{i=1}^{m} \alpha_i} = \frac{1}{\|\alpha\|_1}$$

Linear SVM - Dual optimization problem

Remark. The dual optimization problem

$$\max_{\alpha_1,\ldots,\alpha_m} \mathcal{L}(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

$$\text{subject to } \sum_{i=1}^{m} \alpha_i y_i = 0 \text{ and } \alpha_i \geq 0, \text{ for all } i = 1, \ldots, m.$$

and the related expressions

$$h(x) = \operatorname{sign}\left(\sum_{i=1}^{m} \alpha_i y_i (x_i \cdot x) + b \right)$$

$$b = y_i - \sum_{j=1}^{m} \alpha_j y_j (x_j \cdot x_i)$$

reveal a very important property of SVMs: *the hypothesis solution depends only on inner products between vectors and not directly on the vectors themselves.*

Linear SVM - Learning guarantee

We can derive a learning guarantee for SVMs based on the fraction of support vectors in the training set.

Definition. *Let h_S denote the hypothesis returned by a learning algorithm $\mathcal{A}$, when trained on a fixed sample S. Then, the* **leave-one-out error** *of $\mathcal{A}$ on a sample S of size m is defined by*

$$\hat{\mathcal{R}}_{LOO}(\mathcal{A}) = \frac{1}{m} \sum_{i=1}^{m} 1_{h_{S-\{x_i\}}(x_i) \neq y_i}$$

That is, for each $i \in [m]$, $\mathcal{A}$ is trained on all the points in S except for x_i, i.e., $S - \{x_i\}$ and its error is then computed using x_i. The leave-one-out error is the average of these errors.

Linear SVM - Learning guarantee

Proposition. *The average leave-one-out error for samples of size $m \geq 1$ is an unbiased estimate of the average generalization error for samples of size $m-1$:*

$$\underset{S\sim\mathcal{D}^m}{E}[\hat{\mathcal{R}}_{LOO}(\mathcal{A})] = \underset{S'\sim\mathcal{D}^{m-1}}{E}[\mathcal{R}(h_{S'})]$$

where $\mathcal{D}$ is the distribution according to which points are drawn.

Proof. Since the points of S are drawn in an i.i.d. fashion,

$$\begin{aligned}\underset{S\sim\mathcal{D}^m}{E}[\hat{\mathcal{R}}_{LOO}(\mathcal{A})] &= \frac{1}{m}\sum_{i=1}^{m}\underset{S\sim\mathcal{D}^m}{E}[1_{h_{S-\{x_i\}}(x_i)\neq y_i}]\\ &= \underset{S\sim\mathcal{D}^m}{E}[1_{h_{S-\{x_1\}}(x_1)\neq y_1}]\\ &= \underset{S'\sim\mathcal{D}^{m-1},x_1\sim\mathcal{D}}{E}[1_{h_{S'}(x_1)\neq y_1}]\\ &= \underset{S'\sim\mathcal{D}^{m-1}}{E}[\underset{x_1\sim\mathcal{D}}{E}[1_{h_{S'}(x_1)\neq y_1}]]\\ &= \underset{S'\sim\mathcal{D}^{m-1}}{E}[\mathcal{R}(h_{S'})] \qquad \square\end{aligned}$$

Linear SVM - Learning guarantee

Theorem. *Let h_S be the hypothesis returned by SVMs for a sample S and $N_{SV}(S)$ be the number of support vectors that defines h_S. Then*

$$\underset{S\sim\mathcal{D}^m}{E}[\mathcal{R}(h_S)] \leq \underset{S\sim\mathcal{D}^{m+1}}{E}[\tfrac{N_{SV}(S)}{m+1}]$$

Proof. Let S be a linearly separable sample of $m+1$.
If x is not a support vector for h_S, removing it does not change the SVM solution.
Thus, $h_{S-\{x\}} = h_S$ and $h_{S-\{x\}}$ correctly classifies x.
By contraposition, if $h_{S-\{x\}}$ misclassifies x, x must be a support vector, which implies

$$\hat{\mathcal{R}}_{LOO}(\mathcal{A}) \leq \tfrac{N_{SV}(S)}{m+1}$$

The proof is completed by taking the expectation of both sides and applying the Proposition above. □

Linear SVM - Learning guarantee

Remark. The theorem shows that the average error of the SVM algorithm is upper bounded by the average fraction of support vectors.

One may hope that for many distributions seen in practice, a relatively small number of training points will lie on the marginal hyperplanes. The solution will then be **sparse** in the sense that a small fraction of the dual variables α_i will be non-zero.

However, this bound is relatively weak since it applies only to the average generalization error of the algorithm over all samples of size m.
It provides no information about the variance of the generalization error.

A stronger result about the margin theory will be proved below.

Linear SVM - Non-separable case

In most practical settings, the training data is not linearly separable, which implies that for any hyperplane $w \cdot x + b = 0$, there exists $x_i \in S$ such that

$$y_i(w \cdot x_i + b) \ngeq 1$$

The constraints imposed in the linearly separable case cannot all hold simultaneously.
However, we can impose a relaxed version of these constraints, that is, for each $i \in [m]$, there exist $\xi_i \geq 0$ such that

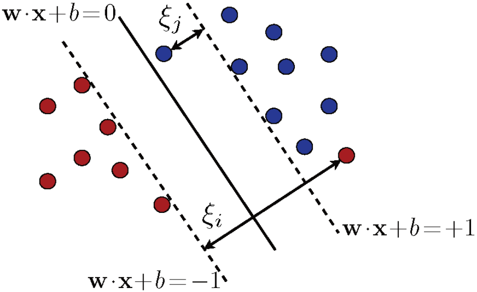
$$y_i(w \cdot x_i + b) \geq 1 - \xi_i$$

The variables ξ_i are known as **slack variables** and they measure the distance by which the vector x_i violate the desired inequality $y_i(w \cdot x_i + b) \geq 1$

Linear SVM - Non-separable case

For a hyperplane $w \cdot x + b = 0$, a vector x_i with $\xi_i > 0$ can be viewed as an outlier.

Each xi must be positioned on the correct side of the appropriate marginal hyperplane to not be considered an outlier. As a consequence, a vector x_i with $y_i(w \cdot x_i + b) < 1$ is correctly classified by the hyperplane $w \cdot x + b = 0$ but is nonetheless considered to be an outlier, that is $\xi_i > 0$.



$\mathbf{w}\cdot\mathbf{x}+b=0$
ξ_j
ξ_i
$\mathbf{w}\cdot\mathbf{x}+b=+1$
$\mathbf{w}\cdot\mathbf{x}+b=-1$

Linear SVM - Non-separable case

If we omit the outliers, the training data is correctly separated by $w \cdot x + b = 0$ with a margin $\rho = \frac{1}{\|w\|^2}$ that we refer to as the **soft margin**, as opposed to the **hard margin** in the separable case.

How should we select the hyperplane in the non-separable case?

There are two conflicting objectives:

1. on one hand, we wish to limit the total amount of slack due to outliers, which can be measured by $\sum_{i=1}^{m} \xi_i$ or, more generally, by $\sum_{i=1}^{m} \xi_i^p$, for some $p \geq 1$;
2. on the other hand, we seek a hyperplane with a large margin, though a larger margin can lead to more outliers and thus larger amounts of slack.

Linear SVM - Non-separable case

The following general optimization problem defining SVMs in the non-separable case where the parameter $C \geq 0$ determines the trade-off between margin-maximization (or minimization of $\|w\|^2$) and the minimization of the slack penalty.

$$\min_{w,b,\xi} \tau(w) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i^p$$

$$\text{subject to:} \quad y_i(w \cdot x_i + b) \geq 1 - \xi_i \wedge \xi_i \geq 0, \quad \forall i = 1, \ldots, m.$$

where $\xi = (\xi_1, \ldots, \xi_m)^t$ and the parameter C is determined via n-fold cross-validation.

Remark. As in the separable case, this is a **convex optimization** problem since the constraints are affine and thus convex and since the objective function is convex for any $p \geq 1$.

Linear SVM - Non-separable case

There are many possible choices for p in the slack penalty $\sum_{i=1}^{m} \xi_i^p$ leading to more or less aggressive penalizations of the slack terms.

The loss functions associated with $p = 1$ and $p = 2$ are called the **hinge loss** and the **quadratic hinge loss**, respectively.

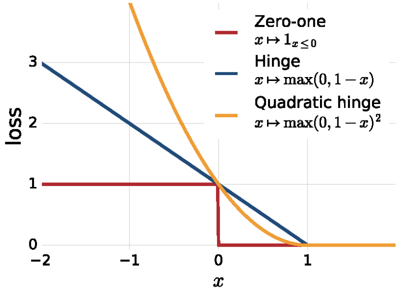


Figure: The hinge loss and the quadratic hinge loss provide convex upper bounds on the binary zero-one loss.

Linear SVM - KKT conditions

As in the separable case, the objective function and the afine constraints are convex and differentiable. We can reformulate the optimization problem using Lagrange multipliers and apply the KKT conditions.

Let $\alpha = (\alpha_1, \ldots, \alpha_m)^t$, $\beta = (\beta_1, \cdots, \beta_m)^t$, with $\alpha_i, \beta_i \geq 0$ for all $i \in [m]$. For $w \in \mathbb{R}^N, b \in \mathbb{R}$, $\xi = (\xi_1, \ldots, \xi_m)^t \in \mathbb{R}^m_+$ we define the Lagrangian

$$\mathcal{L}(w, b, \xi, \alpha, \beta) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^m \xi_i - \sum_{i=1}^m \alpha_i(y_i(w \cdot x_i + b) - 1 + \xi_i) - \sum_{i=1}^m \beta_i \xi_i.$$

Then the primal optimization problem can be solved as

$$\min_{w \in \mathbb{R}^N, b \in \mathbb{R}, \xi \in \in \mathbb{R}^m_+} \mathcal{L}(w, b, \alpha, \beta)$$

$$\text{subject to } \alpha_i, \beta_i \geq 0, \text{ for all } i = 1, \ldots m.$$

Linear SVM - KKT conditions

The KKT conditions for the solution of the optimization problem are obtained by setting the gradient of the Lagrangian with respect to the primal variables w, b and ξ to zero and by writing the complementarity conditions:

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial b} &= -\sum_{i=1}^{m} \alpha_i y_i = 0 && \implies \sum_{i}^{m} \alpha_i y_i = 0, \\
\frac{\partial \mathcal{L}}{\partial w} &= w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0 && \implies w = \sum_{i=1}^{m} \alpha_i y_i x_i \\
\frac{\partial \mathcal{L}}{\partial \xi_i} &= C - \alpha_i - \beta_i = 0 && \implies \alpha_i + \beta_i = C \\
\alpha_i(y_i(w \cdot x_i + b) &- 1 + \xi_i) = 0, \forall i && \implies \alpha_i = 0 \vee y_i(w \cdot x_i + b) = 1 - \xi_i \\
\beta_i \xi_i &= 0, \forall i && \implies \beta_i \vee \xi_i = 0
\end{aligned}
$$

Linear SVM - KKT conditions

Remark. As in the separable case, eq. $w = \sum_{i=1}^{m} \alpha_i y_i x_i$ shows that the solution w is a linear combination of the **support vectors**, that is, those training vectors x_i for which $\alpha_i > 0$.

By the complementarity condition, there are two types of support vectors.
If $\alpha_i = 0$, then $y_i(w \cdot x_i + b) = 1$. Thus the support vectors lie on the marginal hyperplanes $w \cdot x + b = \pm 1$.
If $\alpha_i \neq 0$, then $y_i(w \cdot x_i + b) = 1 - \xi_i$ and x_i is an outlier. In this case, eq. $\beta_i \xi_i = 0$ implies $\beta_i = 0$ and eq. $\alpha_i + \beta_i = C$ then requires $\alpha_i = C$.

Thus, support vectors x_i are either outliers, in which case $\alpha_i = C$, or vectors lying on the marginal hyperplanes.

As in the separable case, while the weight vector w solution is unique, the support vectors are not.

Linear SVM - Dual optimization problem

We derive the dual form of the constrained optimization problem stated above by plugging into the Lagrangian the definition of w in terms of the dual variables.

This yields

$$\mathcal{L} = \frac{1}{2}\|\sum_{i=1}^{m} \alpha_i y_i x_i\|^2 - \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i,j=1}^{m} \alpha_i \alpha_j b + \sum_{i=1}^{m} \alpha_i .$$

which - exactly as in the separable case - simplifies to

$$\mathcal{L} = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

However, here, in addition to $\alpha_i \geq 0$, we must impose the constraint on the Lagrange variables $\beta_i \geq 0$.
In view of $\alpha_i + \beta_i = C$, this is equivalent to $\alpha_i \leq C$.

Linear SVM - Dual optimization problem

Hence, by applying the constraint, we obtain the following dual optimization problem for SVMs

$$\max_{\alpha_1,\ldots,\alpha_m} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

$$\text{subject to } \sum_{i=1}^{m} \alpha_i y_i = 0 \text{ and } 0 \leq \alpha_i \leq C, \text{ for all } i = 1, \ldots, m.$$

which differs from the separable case for the constraints $\alpha_i \leq C$.

Remark. As in the separable case, the objective function and the constraints are convex. This dual optimization problem is also a quadratic programming problem which is equivalent to the primal problem.

Linear SVM - Dual optimization problem

As in the linear case, the solution of the dual problem can be used directly to determine the hypothesis returned by SVMs and the expression of b, yielding

$$h(x) = \operatorname{sign}(w \cdot x + b) = \operatorname{sign}\left(\sum_{i=1}^{m} \alpha_i y_i (x_i \cdot x) + b\right)$$

and

$$b = y_i - \sum_{j=1}^{m} \alpha_j y_j (x_j \cdot x_i),$$

which are valid for $0 \leq \alpha_i \leq C$.

As in the separable case, the dual optimization problem and the expressions above show that the hypothesis solution depends only on inner products between vectors and not directly on the vectors themselves.

Linear SVM - Margin theory

Recall that the VC-dimension of the family H of hyperplanes or linear hypotheses in R^N is $N+1$.

Hence, the application of the of the VC-dimension bound to this hypothesis gives that, for any $\delta > 0$, with probability at least $1-\delta$, the following holds for all $h \in H$

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}_S(h) + \sqrt{\frac{2(N+1)\log \frac{m}{(N+1)}}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$$

When the dimension of the feature space N is large compared to the sample size m, this bound is uninformative.

However, we will derive learning guarantees presented that are independent of the dimension N and thus hold regardless of its value.

Linear SVM - Margin theory

Definition. *The* **confidence margin** *of a real-valued function h at a point x labeled with y is the quantity $y\,h(x)$.*

Thus, when $y\,h(x) > 0$, h classifies x correctly but we interpret the magnitude $|h(x)|$ as the **confidence** of the prediction made by h.

The notion of *confidence margin* is distinct from that of *geometric margin* and does not require a linear separability assumption.

The two notions are related in the separable case: For $h(x) = w \cdot x + b$ with geometric margin ρ_{geom}, the confidence margin at any point x of the training sample with label y satisfies

$$|y\,h(x)| \geq \rho_{geom} \|w\|$$

Linear SVM - Margin theory

For any parameter $\rho > 0$, we define a ρ-**margin loss function** that penalizes h with the cost of 1 when it misclassifies a point x but also penalizes h (linearly) when it correctly classifies x with confidence less than or equal to ρ.

Definition. *For any $\rho > 0$, the ρ-**margin loss function** $L_\rho : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ is defined for all $y, y' \in \mathbb{R}$ by $L_\rho(y, y') = \Phi_\rho(y\,y')$ with*

$$\Phi_\rho(x) = \min(1, \max(0, 1 - \tfrac{x}{\rho})) = \begin{cases} 1 & \text{if } x \leq 0 \\ 1 - \frac{x}{\rho} & \text{if } 0 < x \leq \rho \\ 0 & \text{if } \rho \leq x. \end{cases}$$

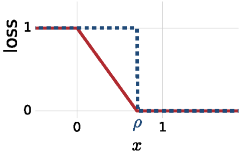


Figure: Margin loss function (in red) defined with $\rho = 0.7$

Linear SVM - Margin theory

We define the empirical margin loss as the margin loss over the training sample.

Definition. *Given a sample $S = (x_1, \ldots, x_m)$ and a hypothesis h, the* **empirical margin loss** *is defined by*

$$\hat{\mathcal{R}}_{S,\rho}(h) = \frac{1}{m}\sum_{i=1}^{m} \Phi_\rho(y_i h(x_i))$$

Since $\Phi_\rho(y_i h(x_i)) \leq 1_{y_i h(x_i) \leq \rho}$ then the empirical margin loss satisfies

$$\hat{\mathcal{R}}_{S,\rho}(h) \leq \frac{1}{m}\sum_{i=1}^{m} 1_{y_i h(x_i) \leq \rho}$$

Interpretation: the empirical margin loss can be replaced by this upper bound, which represents the *fraction of the points in the training sample S that have been misclassified or classified with confidence less than ρ.*

Linear SVM - Margin theory

In other words, the upper bound of the empirical margin loss is the *fraction of the points in the training data with margin less than* ρ.

This corresponds to the loss function indicated by the blue dotted line in the figure

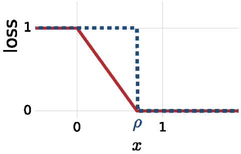


Figure: Margin loss function (in red) defined with $\rho = 0.7$

The advantage of using a loss function based on Φ_ρ as opposed to the zero-one loss or the loss defined by the blue dotted line in the figure is that Φ_ρ is Lispchitz continuous.

Linear SVM - Margin theory

The lemma below bounds the empirical Rademacher complexity of a hypothesis set H after composition with a Lipschitz function in terms of the empirical Rademacher complexity of H.

Lemma (Talagrand's lemma). *Let $\Phi_1, \ldots, \Phi_m$ be λ-Lipschitz functions from $\mathbb{R}$ to $\mathbb{R}$ and let $\sigma_1, \ldots, \sigma_m$ be Rademacher random variables. Then , for any hypothesis set H of real value function,*

$$\frac{1}{m} \underset{\sigma}{E} \left[\sup_{h \in H} \sum_{i=1}^{m} \sigma_i (\Phi_i \circ h)(x_i) \right] \leq \frac{\lambda}{m} \underset{\sigma}{E} \left[\sup_{h \in H} \sum_{i=1}^{m} \sigma_i h(x_i) \right] = \lambda \mathfrak{R}_S(H)$$

In particular, if $\Phi_1 = \Phi$ for all $i \in [m]$, then the following holds:

$$\mathfrak{R}_S(\Phi \circ H) \leq \lambda \mathfrak{R}_S(H).$$

This lemma will be needed for the proof of the margin-based generalization bound.

Linear SVM - Margin theory

We can now prove the following general margin-based generalization bound.

Theorem (Margin bound for binary classification). *Let H be a set of real-valued functions and fix $\rho > 0$. then, for any $\delta > 0$, with probability at least $1 - \delta$, each of the following holds for all $h \in H$:*

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}_{S,\rho}(h) + \tfrac{2}{\rho}\,\mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$$

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}_{S,\rho}(h) + \tfrac{2}{\rho}\,\mathfrak{R}_m(\mathcal{H}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}$$

Linear SVM - Margin theory

Remarks. The generalization bounds of the theorem suggest a trade-off:

A larger value of ρ decreases the complexity term (second term), but tends to increase the empirical margin-loss $\hat{\mathcal{R}}_{S,\rho}(h)$ (first term) by requiring from a hypothesis h a higher confidence margin.

Thus, if for a relatively large value of ρ the empirical margin loss of h remains relatively small, then h benefits from a very favorable guarantee on its generalization error.

For the theorem to hold, the margin parameter ρ must be selected beforehand. However, we will show next that the bounds of the theorem can be generalized to hold uniformly for all $\rho \in (0, 1]$ at the cost of a modest additional term.

Linear SVM - Margin theory

Proof. Let $\tilde{H} = \{z = (x, y) \mapsto yh(x : h \in H)\}$ and consider the functions of the form $g = \Phi_\rho \circ f$, where $f \in \tilde{H}$.
Using the generalization bound for the Rademacher complexity presented above, for any $\delta > 0$, with probability at least $1 - \delta$, for any such g

$$E[g(z)] \leq \frac{1}{m} \sum_{i=1}^{m} g(z_i) + 2\,\mathfrak{R}_m(\tilde{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$$

and, thus, for all $h \in H$.

$$E[\Phi_\rho(yh(x))] \leq \hat{\mathcal{R}}_{S,\rho}(h) + 2\,\mathfrak{R}_m(\Phi_\rho \circ \tilde{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

Since $1_{u \leq 0} \leq \Phi_\rho(u)$ for all $u \in \mathbb{R}$, then

$$\mathcal{R}(h) = E[1_{yh(x) \leq 0}] \leq E[\Phi_\rho(yh(x))]$$

Linear SVM - Margin theory

From the last two inequalities, we derive

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}_{S,\rho}(h) + 2\,\mathfrak{R}_m(\Phi_\rho \circ \tilde{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$$

Since Φ_ρ is $1/\rho$-Lipschitz, by Talagrand's lemma it follows that $\mathfrak{R}_m(\Phi_\rho \circ \tilde{H}) \leq \frac{1}{\rho}\mathfrak{R}_m(\tilde{H})$. In addition, we have

$$\mathfrak{R}_m(\tilde{H}) = \frac{1}{m} \underset{S,\sigma}{E} \left[\sup_{h \in H} \sum_{i=1}^{m} \sigma_i y_i h(x_i) \right] = \frac{1}{m} \underset{S,\sigma}{E} \left[\sup_{h \in H} \sum_{i=1}^{m} \sigma_i h(x_i) \right] = \mathfrak{R}_m(H).$$

Using these observations in the inequality above, we obtain the first bound of the theorem.

The second bound is proved similarly using the other generalization bound for the Rademacher complexity. □

Linear SVM - Margin theory

Here is the version of the Margin Bound theorem valid uniformly for $\rho \in (0, r]$ for some $r > 0$.

Theorem. *Let H be a set of real-valued functions and fix $r > 0$. then, for any $\delta > 0$, with probability at least $1 - \delta$, each of the following holds for all $h \in H$ and $\rho \in (0, r]$:*

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}_{S,\rho}(h) + \tfrac{4}{\rho}\,\mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log \log_2 \frac{2r}{\rho}}{m}} + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}$$

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}_{S,\rho}(h) + \tfrac{4}{\rho}\,\mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log \log_2 \frac{2r}{\rho}}{m}} + 3\sqrt{\frac{\log \frac{4}{\delta}}{2m}}$$

Linear SVM - Margin theory

Proof. Let (ρ_k) and (ϵ_k), with $\epsilon_k \in (0,1]$, be two sequences. By the Margin Bound Theorem above, for any fixed $k \geq 1$,

$$P\left[\sup_{h \in H, k \geq 1} \mathcal{R}(h) - \hat{\mathcal{R}}_{S,\rho_k}(h) > \tfrac{2}{\rho_k}\,\mathfrak{R}_m(\mathcal{H}) + \epsilon_k\right] \leq \exp\left(-2m\epsilon_k^2\right)$$

Choosing $\epsilon_k = \epsilon + \sqrt{\frac{\log k}{m}}$, we have that

$$\begin{aligned}
P\left[\sup_{h \in H, k \geq 1} \mathcal{R}(h) - \hat{\mathcal{R}}_{S,\rho_k}(h) - \tfrac{2}{\rho_k}\,\mathfrak{R}_m(\mathcal{H}) - \epsilon_k > 0\right] &\leq \sum_{k=1}^{\infty} \exp\left(-2m\epsilon_k^2\right) \\
&\leq \sum_{k=1}^{\infty} \exp\left(-2m(\epsilon + \sqrt{\tfrac{\log k}{m}})^2\right) \\
&= \sum_{k=1}^{\infty} \exp\left(-2m\epsilon^2\right) \exp\left(-2\log k\right) \\
&= \left(\sum_{k=1}^{\infty} \tfrac{1}{k^2}\right) \exp\left(-2m\epsilon^2\right) \\
&\leq 2\exp\left(-2m\epsilon^2\right)
\end{aligned}$$

Linear SVM - Margin theory

We choose $\rho_k = r/2^k$. For any $\rho \in (0, r]$, there exists $k \geq 1$ such that $\rho \in (\rho_{k-1}, \rho_k]$ with $\rho_0 = r$.
For that k, $\rho \leq \rho_{k-1} = 2\rho_k$, hence $1/\rho_k \leq 2/\rho$ and

$$\sqrt{\log k} = \sqrt{\log \log_2(r/\rho_k)} \leq \sqrt{\log \log_2(2r/\rho)}$$

Further, for any $h \in H$, $\hat{\mathcal{R}}_{S,\rho_k}(h) \leq \hat{\mathcal{R}}_{S,\rho}(h)$. Thus

$$P[\sup_{h\in H, k\geq 1} \mathcal{R}(h) - \hat{\mathcal{R}}_{S,\rho}(h) - \tfrac{4}{\rho}\,\mathfrak{R}_m(\mathcal{H}) - \sqrt{\tfrac{\log \log_2(2r/\rho)}{m}} - \epsilon > 0] \leq \exp\left(-2m\epsilon_k^2\right)$$

which proves the first statement.
The second statement can be proven in a similar way. □

Linear SVM - Margin theory

We can finally state the following general margin bound for linear hypotheses with bounded weight vectors.

Corollary. Let $H = \{x \mapsto w \cdot x : \|w\| \leq \Lambda\}$, assume that $X \subset \{x : \|x\| \leq r\}$ and fix $\rho > 0$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the choice of a sample S of size m, the following holds for any $h \in H$:

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

Proof. The proof follows from the Margin Bound theorems above and the observation that, under the assumption of the theorem, the empirical Rademacher complexity can be bounded as

$$\mathfrak{R}_m(\mathcal{H}) \leq \sqrt{\frac{r^2\Lambda^2}{m}}.$$

Linear SVM - Margin theory

As with the Margin Bound theorem, the bound of the corollary can be generalized to hold uniformly for all $\rho \in (0, 1]$ at the cost of a additional term. Namely,

Corollary. Let $H = \{x \mapsto w \cdot x : \|w\| \leq \Lambda\}$, assume that $X \subset \{x : \|x\| \leq r\}$ and let $\rho \in (0, 1]$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the choice of a sample S of size m, the following holds for any $h \in H$:

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}_{S,\rho}(h) + 2\sqrt{\frac{r^2\Lambda^2/\rho^2}{m}} + \sqrt{\frac{\log \log_2 \frac{1}{\rho}}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$$

Linear SVM - Margin theory

Remarks. The generalization bound for linear hypotheses in the Corollary *does not depend directly on the dimension of the feature space, but only on the margin.*

It suggests that a small generalization error can be achieved when $\rho/(r\Lambda)$ is large (small second term) while the empirical margin loss is relatively small (first term). The latter occurs when few points are either classified incorrectly or correctly, but with margin less than ρ.

When S is linearly separable, for a linear hypothesis with geometric margin ρ_{geom} and the choice $\rho = \rho_{geom}$, the empirical margin loss term is zero.
Thus, if ρ_{geom} is relatively large, this provides a strong guarantee for the generalization error of the corresponding linear hypothesis.

Linear SVM - Margin theory

Remarks. Is there a contradiction with the VC-dimension lower bounds stating that for any learning algorithm $\mathcal{A}$ there exists a bad distribution for which the error of the hypothesis returned by the algorithm is $\Omega(\sqrt{d/m})$ with a non-zero probability?
No. The bound of the corollary does not rule out such bad cases. However, for such bad distributions, the empirical margin loss would be large even for a relatively small margin ρ, and thus the bound of the corollary would be loose in that case.

The learning guarantee of the corollary hinges upon the hope of a good margin value ρ.
If there exists a relatively large margin value $\rho > 0$ for which the empirical margin loss is small, then a small generalization error is guaranteed by the corollary.
This favorable margin situation depends on the distribution: while the learning bound is distribution-independent, the existence of a good margin is in fact distribution-dependent.

A favorable margin appear relatively often in applications.

Linear SVM - Margin theory

Choose $\Lambda = 1$ in the Corollary. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the choice of a sample S of size m, the following holds for any $h \in \{x \mapsto w \cdot x : \|w\| \leq 1\}$, $\rho \in (0, r]$:

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}_{S,\rho}(h) + 4\sqrt{\frac{r^2/\rho^2}{m}} + \sqrt{\frac{\log \log_2 \frac{2r}{\rho}}{m}} + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}$$

The margin loss function is upper bounded by the hinge loss

$$\Phi_\rho(u) = \min\left(1, \max(0, 1 - \tfrac{u}{\rho})\right) \leq \max(0, 1 - \tfrac{u}{\rho})$$

Thus, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for any $h \in \{x \mapsto w \cdot x : \|w\| \leq 1\}$:

$$\mathcal{R}(h) \leq \frac{1}{m}\sum_{i=1}^{m} \max(0, 1 - y_i(w \cdot x_i)) + 4\sqrt{\frac{r^2/\rho^2}{m}} + \sqrt{\frac{\log \log_2 \frac{2r}{\rho}}{m}} + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}$$

This inequality can be used to derive an algorithm that selects w and $\rho > 0$ to minimize the right-hand side.

Linear SVM - Margin theory

Since only the first term of the right-hand side depends on w, for any $\rho > 0$, the bound suggests selecting w as the solution of the following optimization problem:

$$\min_{\|w\| \leq \frac{1}{\rho}} \frac{1}{m} \sum_{i=1}^{m} \max(0, 1 - y_i(w \cdot x_i))$$

Introducing a Lagrange variable $\lambda \geq 0$, the optimization problem can be written equivalently as

$$\min_{w} \lambda \|w\|^2 + \frac{1}{m} \sum_{i=1}^{m} \max(0, 1 - y_i(w \cdot x_i))$$

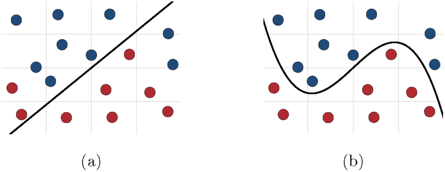
The resulting algorithm precisely coincides with SVMs.

Non-Linear SVM

Kernel methods are use to extend SVMs to define non-linear decision boundaries.

The main idea is based on **kernel functions** which are used to implicitly define an inner product in a high-dimensional space $\mathcal{H}$.

Replacing the original inner product in the input space X with such kernels extends SVMs to a linear separation in $\mathcal{H}$, or, equivalently, to a non-linear separation in X.



(a)

(b)

Figure: The classification task consists of discriminating between blue and red points. (a) No hyperplane can separate the two populations but (b) a non-linear mapping can be used instead.

Non-Linear SVM

We recall that the dual optimization problem for SVMs

$$\max_{\alpha_1,\ldots,\alpha_m} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

$$\text{subject to } \sum_{i=1}^{m} \alpha_i y_i = 0 \text{ and } 0 \le \alpha_i \le C, \text{ for all } i = 1, \ldots, m.$$

which leads to writing the hypothesis h returned by SVMs as

$$h(x) = \operatorname{sign}(w \cdot x + b) = \operatorname{sign}\left(\sum_{i=1}^{m} \alpha_i y_i (x_i \cdot x) + b \right)$$

with

$$b = y_i - \sum_{j=1}^{m} \alpha_j y_j (x_j \cdot x_i),$$

Non-Linear SVM

Let $\Phi : X \mapsto \mathcal{H}$ be a **feature map** that maps the input data to some Hilbert space $\mathcal{H}$ called **feature space**.

The feature map Φ is typically nonlinear and $\mathcal{H}$ may be infinite dimensional.

By mapping the input data $x_1, \ldots, x_m \in X$ to $\mathcal{H}$, we expect that the features $\Phi(x_1), \ldots, \Phi(x_m)$ will be linearly separable in $\mathcal{H}$.

Next, assume that there is a kernel function $K : X \times X \mapsto \mathbb{R}$ on the input space satisfying

$$K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle_{\mathcal{H}}.$$

Non-Linear SVM

Under these assumptions, we can now reformulate the SVM optimization problem by replacing the inner products $x \cdot x'$ with kernels $K(x, x')$:

$$\max_{\alpha_1,\ldots,\alpha_m} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

$$\text{subject to } \sum_{i=1}^{m} \alpha_i y_i = 0 \text{ and } 0 \leq \alpha_i \leq C, \text{ for all } i = 1, \ldots, m.$$

so that we write the hypothesis h returned by SVMs as

$$h(x) = \operatorname{sign}(w \cdot x + b) = \operatorname{sign}\left(\sum_{i=1}^{m} \alpha_i y_i K(x_i, x) + b\right)$$

with

$$b = y_i - \sum_{j=1}^{m} \alpha_j y_j K(x_j, x_i).$$

Kernel Methods

Definition Let $X \neq \emptyset$ be a set. A function $k: X \times X \to \mathbb{R}$ is called a **kernel** on X iff there is a Hilbert space $\mathcal{H}$ and a feature map $\Phi: X \to \mathcal{H}$ such that for any $x, x' \in X$

$$k(x, x') = \langle \Phi(x'), \Phi(x) \rangle_{\mathcal{H}}$$

holds.

Remark. Given a kernel k, neither Φ nor $\mathcal{H}$ are uniquely determined.

Kernel Methods

Example 1. Let $X = \mathbb{R}$ and $k(x, x') = x' x$. Obviously, k is a kernel on X with $\Phi_1(x) = x$ being the identity map and $\mathcal{H}_1 = \mathbb{R}$.

Next Consider $\Phi_2 : X \to \mathbb{R}^2 = \mathcal{H}_2$ given by

$$\Phi_2(x) = \frac{1}{\sqrt{2}}(x,\, x).$$

We have

$$\langle \Phi_2(x'), \Phi_2(x) \rangle_{\mathbb{R}^2} = \frac{x'x}{\sqrt{2}} + \frac{x'x}{\sqrt{2}} = x'x = k(x, x'),$$

and hence k is a kernel on X also for Φ_2 and $\mathcal{H}_2$.

Kernel Methods

Example 2. Let $X \neq \emptyset$ and $\{f_n\}_{n=1}^{\infty}$ be a set of functions $f_n : X \to \mathbb{R}$ with the property that $f_n(x) \in \ell^2$ for any $x \in X$. Then

$$k(x, x') = \sum_{i=1}^{\infty} f_n(x)\, \overline{f_n(x')}$$

is a kernel on X with $\Phi(x) = \overline{f_n(x)}$, $\Phi : X \to \ell^2$, i.e., the sum

$$\langle \Phi(x'), \Phi(x) \rangle_{\ell^2} = \sum_{i=1}^{\infty} f_n(x)\, \overline{f_n(x')} = k(x, x')$$

is well defined since $f_n(x) \in \ell^2$ for any $x \in X$ by Hölder's inequality.

Kernel Methods

Properties of kernels

1. If k_1, k_2 are kernels then $k_1 + k_2$ is a kernel.
2. If $\alpha \geq 0$ and k is a kernel, then αk is a kernel.

   Remark: The space of kernels forms a cone but not a vector space as shown by the argument below.

   Let k_1, k_2 be kernels on X such that, for some $x \in X$,

   $$k_1(x,x) - k_2(x,x) < 0.$$

   If $k_1 - k_2$ is kernel, then there exist a map $\Phi : X \to H$ such that

   $$0 \leq \langle \Phi(x), \Phi(x) \rangle = k_1(x,x) - k_2(x,x) < 0,$$

   giving a contradiction. So $k_1 - k_2$ is not a kernel.

Kernel Methods

3. Let k be a kernel on X and A be a map, $A : \bar{\underline{\mathrm{Y}}} \to X$, where Y is another set.
   Then, $\overline{k}(x, x') = k(A(x), A(x'))$, for $x, x' \in X$ defines a kernel on Y.
   This include the special case where A is a restriction map. Hence, if $Y \subset X$, then $k_{|Y \times Y}$ is a kernel.
4. If k_1 is a kernel on X_1 and k_2 is a kernel on X_2, then $k_1.k_2$ is a kernel on the tensor space $X_1 \times X_2$.
   In particular, if $X_1 = X_2 = X$, then

   $$k(x, x') = k_1(x, x')k_2(x, x'),\ x, x' \in X$$

   defines a kernel on X.

Kernel Methods

Example 3 (Polynomial kernels).
By the properties of kernels, for any $n > 0$, the map $k_n(x, x') = (xx')^n$, where $x, x' \in X$ is a kernel.
Hence, if $p : X \to \mathbb{R}$ is of the form,

$$p(t) = a_n t^n + \ldots + a_1 t + a_0$$

with non-negative coefficients a_i, then $k(x, x') = p(xx')$, with $x, x' \in X$ is a kernel.

In general, the function: $k(z, z_1) = (\langle z, z' \rangle + c)^m$ with $z, z' \in \mathbb{C}^d, c \geq 0$, is a polynomial kernel on $\mathbb{C}^d$.

Kernel Methods

Example 4 (Exponential kernels).
Using the Taylor expansion, one can express the exponential function in terms of polynomials.

Hence, for $d \in \mathbb{N}$, $x, x' \in \mathbb{R}^d$, $k(x, x') = exp(\langle x, x' \rangle)$ is a kernel on $\mathbb{R}^d$.

Similarly, let $d \in \mathbb{N}$, $\gamma > 0$, $z = (z_1, ..., z_d)$, $z' = (z'_1, ..., z'_d) \in \mathbb{C}^d$. Then

$$k_{\gamma,\mathbb{C}^d}^{(z,z')} = exp(-\gamma^{-2} \sum_{j=1}^{d} (z_j, -\bar{z}'_j)^2)$$

is a kernel on $\mathbb{C}^d$.
Its restriction $k_\gamma = exp(-\frac{||x-x'||_2^2}{\gamma^2})$, for $x, x' \in \mathbb{R}^d$, is a kernel on $\mathbb{R}^d$.

Kernel Methods

Definition. A function $k : X \times X \to \mathbb{R}$ is **positive definite** if for all $n \in \mathbb{N}$, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, and all $x_1, \ldots, x_n \in X$, we have

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0$$

Furthermore, it is **strictly positive definite** if for mutually distinct $x_1, \ldots, x_n \in X$, equality only occurs when $\alpha_1 = \cdots = \alpha_n = 0$.
k is **symmetric** if $k(x, x') = k(x', x)$, for all $x, x' \in X$.

Given a function $k : X \times X \to \mathbb{R}$, the matrix $K = (k(x_i, x_j))_{i,j}$ is the **Gram matrix** of k with respect to the vectors $x_1, \ldots, x_n$ in X.

We have that

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0 \iff K \text{ is positive definite.}$$

Kernel Methods

Theorem *A function $k : X \times X \to \mathbb{R}$ is a kernel if and only if it is symmetric and positive definite*

Proof. $(\Longrightarrow)$
If k is a kernel, then

$$k(x, x') = \langle \Phi(x'), \Phi(x) \rangle = \langle \Phi(x), \Phi(x') \rangle = k(x', x)$$

is symmetric.
Also, for any $n \in \mathbb{N}$, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, $x_1, \ldots, x_n \in X$, we observe that

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) = \langle \sum_{i=1}^{n} \alpha_i \Phi(x_i), \sum_{j=1}^{n} \alpha_j \Phi(x_j) \rangle = || \sum_{i=1}^{n} \alpha_i \Phi(x_i) ||^2 \geq 0$$

Hence, k is positive definite.

Kernel Methods

($\Longleftarrow$)
Assume $k : X \times X \to \mathbb{R}$ is symmetric and positive definite.
Define

$$\mathcal{H}_{pre} = \left\{ \sum_{i=1}^{n} \alpha_i \, k(\cdot\,,\; x_i) : \; n \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_i \in X \right\}.$$

For any $f = \sum_{i=1}^{n} \alpha_i k(\cdot\,,\; x_i),\; g = \sum_{j=1}^{n} \beta_j k(\cdot\,,\; x_j') \in \mathcal{H}_{pre}$, set

$$\langle f, g \rangle := \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \beta_j k(x_j', x_i).$$

We want to show that this operation defines an inner product on $\mathcal{H}_{pre}$, i.e., that $\langle \cdot, \cdot \rangle$ is bilinear, symmetric and positive definite.

First we observe that, for any $x_j' \in X$, we have
$f(x_j') = \sum_{i=1}^{n} \alpha_i k(x_j', x_i)$, hence $\langle f, g \rangle = \sum_{j=1}^{m} \beta_j f(x_j')$. Similarly, we can write $\langle f, g \rangle = \sum_{i=1}^{n} \alpha_i g(x_i)$.
This shows that $\langle f, g \rangle$ is independent of the representation of f and g.

Kernel Methods

By the assumption on k, it follows that $\langle f, g\rangle$ is symmetric, bilinear and positive, that is

$$\langle f, f\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0$$

for any $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, $x_1, \ldots, x_n \in X$, $f \in \mathcal{H}_{pre}$.

Note that these properties imply the Cauchy-Schwartz inequality,

$$|\langle f, g\rangle|^2 \leq \langle f, f\rangle\langle g, g\rangle$$

for all $f, g, \in \mathcal{H}_{pre}$.

It also follows that if $f = 0$, then $\langle f, f\rangle = 0$.

It remains to show that $\langle f, f\rangle = 0$ implies $f = 0$.

Kernel Methods

Since $\langle f, g\rangle = \sum_{i=1}^{n} \alpha_i g(x_i)$, it follows that

$$\sum_{i=1}^{n} \alpha_i k(x, x_i) = \langle f, k(x, x_i)\rangle = f(x)$$

Using this observation and Cauchy-Schwartz inequality, for any $x \in X$ we have

$$|f(x)|^2 = |\sum_{i=1}^{n} \alpha_i k(x, x_i)|^2 = |\langle f, k(\cdot, x)\rangle|^2 \leq \langle k(\cdot, x), k(\cdot, x)\rangle\langle f, f\rangle.$$

Thus, $\langle f, f\rangle = 0$ implies $f(x) = 0$ for any $x \in X$, hence $f = 0$.

The proof follows using a continuity argument, that is by taking $\mathcal{H}$ to be the completion of $\mathcal{H}_{pre}$. □

Reproducing Kernel Hilbert Space

Definition. Let $X \neq \emptyset$ and $\mathcal{H}$ be a $\mathbb{K}$-Hilbert function space over X, i.e., a Hilbert space that consists of functions with domain in X and range into $\mathbb{K}$ (e.g., $\mathbb{K} = \mathbb{R}$ or $= \mathbb{C}$).

- A function $k : X \times X \to \mathbb{K}$ is called a **reproducing kernel** of $\mathcal{H}$ if $k(\cdot, x) \in \mathcal{H}$ for all $x \in X$ and it satisfies the **reproducing property**

  $$f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}$$

  for all $f \in \mathcal{H}$ and all $x \in X$.
- The space $\mathcal{H}$ is called a **reproducing kernel Hilbert space (RKHS)** over X if for all $x \in X$ the *Dirac functional* $\delta_x : \mathcal{H} \to \mathbb{K}$ defined by

  $$\delta_x(f) = f(x), \quad f \in \mathcal{H},$$

  is continuous.

Reproducing Kernel Hilbert Space

Remark. A RKHS is a space of functions, hence $L^2(\mathbb{R}^d)$ is not a RKHS.

If $\mathcal{H}$ is a RKHS, then *norm convergence implies pointwise convergence.*

To show that this is the case, let $(f_n) \in \mathcal{H}$ be such that $\|f_n - f\|_{\mathcal{H}} \to 0$ as $n \to \infty$ with $f \in \mathcal{H}$.
It follows that for any $x \in X$ there is a constant c such that

$$|\delta_x(f_n) - \delta_x(f)| \leq c\|f - f_n\|_{\mathcal{H}}.$$

Hence

$$\lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \delta_x(f_n) = \delta_x(f) = f(x).$$

Reproducing Kernel Hilbert Space

Proposition (Reproducing kernels are kernels).
Let $\mathcal{H}$ be a Hilbert function space over X that has a reproducing kernel k.
Then $\mathcal{H}$ is a RKHS and $\mathcal{H}$ is also a feature space of k, where the feature map $\Phi : X \to \mathcal{H}$ is given by

$$\Phi(x) = k(\cdot, x), \quad x \in X.$$

We call Φ the **canonical feature map** of kernel k.

Reproducing Kernel Hilbert Space

Proof. First we show that δ_x is continuous so that $\mathcal{H}$ is a RKHS. Since k is a reproducing kernel in $\mathcal{H}$, for any $f \in \mathcal{H}$,

$$|\delta_x(f)| = |f(x)| = |\langle f, k(\cdot, x)\rangle_H| \leq \|f\|\|k(\cdot, x)\|.$$

This shows that δ_x is continuous for any $x \in X$.
Therefore, $\mathcal{H}$ is a *RKHS*.

Next , we show that $\mathcal{H}$ is a feature space of k with feature map Φ.
For a fixed $x' \in X$, let $f = k(\cdot, x')$.
Then, for any $x \in X$,

$$\langle \Phi(x'), \Phi(x)\rangle = \langle k(\cdot, x'), k(\cdot, x)\rangle = \langle f, k(\cdot, x)\rangle = f(x) = k(x, x').$$

Therefore, $\mathcal{H}$ is a feature space of k with a feature map Φ. □

Reproducing Kernel Hilbert Space

We have just seen that every Hilbert function space with a reproducing kernel is a RKHS.

We next show that, conversely, every RKHS has a (unique) reproducing kernel over X and that this kernel can be determined by the Dirac functionals δ_x, $x \in X$.

Theorem (Every RKHS has a unique reproducing kernel). *Let $\mathcal{H}$ be a RKHS over X and $\mathcal{H}'$ be the dual space of $\mathcal{H}$. Then $k : X \times X \to \mathbb{K}$ defined by*

$$k(x, x') = \langle \delta_x, \delta_{x'} \rangle_{\mathcal{H}'}, \quad x, x' \in X,$$

is the only reproducing kernel of $\mathcal{H}$.

Furthermore, if $(e_i)_{i \in I}$ is an orthonormal basis of $\mathcal{H}$, then for all $x, x' \in X$, we have

$$k(x, x') = \sum_{i \in I} e_i(x) \overline{e_i(x')}.$$

Reproducing Kernel Hilbert Space

Proof.

First, we show that k is a reproducing kernel by showing that the reproducing property holds.

By Riesz representation theorem, there exists an isometric anti-linear isomorphisim $I : \mathcal{H}' \to \mathcal{H}$ that assigns to any $g' \in \mathcal{H}'$ a representing element in $\mathcal{H}$; that is

$$g'(f) = \langle f, Ig' \rangle, \text{ for all } f \in \mathcal{H}, g' \in \mathcal{H}'.$$

In particular, for $g' = \delta_x \in \mathcal{H}'$, $f = I\delta_{x'} \in \mathcal{H}$, then

$$\langle I\delta_{x'}, I\delta_x \rangle_{\mathcal{H}} = \delta_x(I\delta_{x'}).$$

With this observation, for all $x, x' \in X$,

$$k(x, x') \overset{\text{def}}{=} \langle \delta_x, \delta_{x'} \rangle_{\mathcal{H}'} \overset{\text{Riesz}}{=} \langle I\delta_{x'}, I\delta_x \rangle_{\mathcal{H}} = \delta_x(I\delta_{x'}) \overset{\text{def}}{=} I\delta_{x'}(x).$$

This shows that $k(\cdot, x') = I\delta_{x'}$ for all $x' \in X$. Hence,

$$f(x') \overset{\text{def}}{=} \delta_{x'}(f) = \langle f, I\delta_{x'} \rangle_{\mathcal{H}} = \langle f, k(\cdot, x') \rangle, \text{ for all } x' \in X.$$

This shows that k has the reproducing property.

Reproducing Kernel Hilbert Space

To show uniqueness, let $\tilde{k}$ be an arbitrary reproducing kernel on $\mathcal{H}$.

For any $x' \in X$, given a basis $(e_i)_{i\in I} \in \mathcal{H}$, we have

$$\tilde{k}(\cdot, x') = \sum_{i\in I} \langle \tilde{k}(\cdot, x'), e_i \rangle e_i = \sum_{i\in I} \overline{\langle e_i, \tilde{k}(\cdot, x') \rangle} e_i.$$

By the reproducing property of $\tilde{k}$, we have

$$\overline{\langle e_i, \tilde{k}(\cdot, x') \rangle} e_i = \overline{e_i(x')} e_i.$$

Therefore,

$$\tilde{k}(\cdot, x') = \overline{e_i(x')} e_i.$$

Since $\tilde{k}$ and $(e_i)_{i\in I}$ are arbitrarily chosen, we find $\tilde{k} = k$.

Therefore, k is the only reproducing kernel of $\mathcal{H}$. □

Reproducing Kernel Hilbert Space

The theorem above shows that a RKHS uniquely determines its reproducing kernel (which is also a kernel).
The following theorem now shows that, conversely, every kernel has a unique RKHS. Thus, it establishes a one-to-one relation between a kernel and a RKHS.

Theorem *Let $X \neq \emptyset$ and k be a kernel over X with feature space $\mathcal{H}_0$ and feature map $\Phi_0 : X \to \mathcal{H}_0$. Then*

$$\mathcal{H} := \{f : X \to \mathbb{K} \mid \exists w \in \mathcal{H}_0 \text{ with } f(x) = \langle w, \Phi_0(x)\rangle_{\mathcal{H}_0} \text{ for any } x \in X\}$$

equipped with the norm

$$\|f\|_{\mathcal{H}} := \inf \{\|w\|_{\mathcal{H}_0} : \; w \in \mathcal{H}_0 \text{ with } f = \langle w, \Phi_0(\cdot)\rangle_{\mathcal{H}_0}\}$$

is the only RKHS for which k is a reproducing kernel.

Reproducing Kernel Hilbert Space

Moreover, the operator $V : \mathcal{H}_0 \to \mathcal{H}$ defined by

$$Vw = \langle w, \Phi_0(\cdot)\rangle_{\mathcal{H}_0} \text{ for } w \in \mathcal{H}_0$$

is a metric surjection. That is, $VB_{\mathcal{H}_0} = B_{\mathcal{H}}$, where $B_{\mathcal{H}_0}$ and $B_{\mathcal{H}}$ are the open unit balls of $\mathcal{H}_0$ and $\mathcal{H}$. Also, the set

$$\mathcal{H}_{pre} := \{\sum_{i=1}^{n} \alpha_i\, k(\cdot, x_i) : n \in \mathbb{N}, \alpha_1, \ldots, \alpha_n \in \mathbb{K}, x_1, ..., x_n \in X\}$$

is dense in $\mathcal{H}$ and, for $f := \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) \in \mathcal{H}_{pre}$, we have

$$\|f\|^2_{\mathcal{H}} = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \overline{\alpha_j} \langle k(\cdot, x_i), k(\cdot, x_j)\rangle_{\mathcal{H}} = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \overline{\alpha_j} k(x_j, x_i).$$

Reproducing Kernel Hilbert Space

Proof. We first show that $\mathcal{H}$ is a Hilbert space.
From the definition, we have that $\mathcal{H}$ is a vector space of functions from $X \to \mathbb{K}$ and $V : \mathcal{H}_0 \to \mathcal{H}$ is a surjective linear operator.

Using V, for any $f \in \mathcal{H}$, we write

$$\|f\|_{\mathcal{H}} = \inf_{w \in V^{-1}(f)} \|w\|_{\mathcal{H}_0}.$$

To show that $\|\cdot\|_{\mathcal{H}}$ is a Hilbert space norm, let $(w_n) \subset \ker V = \{w \in \mathcal{H}_0 | Vw = 0\}$ with the property that $\lim_{n\to\infty} w_n = w$. Then

$$\langle w, \Phi(x) \rangle_{\mathcal{H}_0} = \lim_{n\to\infty} \langle w_n, \Phi(x) \rangle = 0, \text{ for any } x \in X,$$

showing that $w \in \ker V$ and, hence, $\ker V$ is closed.

Denoting $\tilde{\mathcal{H}} = (\ker V)^{\perp}$, we can write $\mathcal{H}_0 = \ker V \oplus \tilde{\mathcal{H}}$.
Then, by construction, the restriction $V_{|\tilde{\mathcal{H}}} : \tilde{\mathcal{H}} \to \mathcal{H}$ of V to $\tilde{\mathcal{H}}$ is injective.

Reproducing Kernel Hilbert Space

We will show that $V_{|\tilde{\mathcal{H}}}$ is also surjective.
Let $f \in \mathcal{H}$ and $w \in \mathcal{H}_0$ with $f(x) = \langle w, \Phi_0(x)\rangle_{\mathcal{H}_0} = Vw(x)$.
We can write $w = w_0 + \tilde{w}$ with $w_0 \in \ker V$ and $\tilde{w} \in (\ker V)^{\perp} = \tilde{\mathcal{H}}$.
Then $f = V(w_0 + \tilde{w}) = V\tilde{w} = V_{|\tilde{\mathcal{H}}}\tilde{w}$.
This shows that $V_{|\tilde{\mathcal{H}}}$ is surjective and, thus, $V_{|\tilde{\mathcal{H}}}$ is also bijective.

Let $(V_{|\tilde{\mathcal{H}}})^{-1}$ be the inverse operator of $V_{|\tilde{\mathcal{H}}}$. Then we have

$$\|f\|^2_{\mathcal{H}} = \inf_{w \in V^{-1}(\{f\})} \|w\|^2_{\mathcal{H}_0} = \inf_{w_0 \in \ker V, \tilde{w} \in \tilde{H}, w_0 + \tilde{w} \in V^{-1}(\{f\})} \|w_0 + \tilde{w}\|^2_{\mathcal{H}_0}$$

$$= \inf_{w_0 \in \ker V, \tilde{w} \in \tilde{H}, w_0 + \tilde{w} \in V^{-1}(\{f\})} \|w_0\|^2_{\mathcal{H}_0} + \|\tilde{w}\|^2_{\mathcal{H}_0} = \|(V_{|\tilde{H}})^{-1} f\|^2_{\tilde{H}}.$$

Since $\tilde{\mathcal{H}}$ is a Hilbert space norm, then $\|\cdot\|_{\mathcal{H}}$ is also a Hilbert space norm. Hence we have shown that $V_{|\tilde{\mathcal{H}}}$ is an isometric isomorphism from $\tilde{\mathcal{H}}$ to $\mathcal{H}$.

Reproducing Kernel Hilbert Space

To show that k is a reproducing kernel of $\mathcal{H}$, note that, for any $x \in X$, by definition

$$k(\cdot, x) = \langle \Phi_0(x), \Phi_0(\cdot) \rangle = V\Phi_0(x) \in \mathcal{H}.$$

Since $\langle w, \Phi_0(x) \rangle = Vw(x) = 0$ for any $w \in \ker V$, then

$$\Phi_0(x) \in (\ker V)^\perp = \tilde{\mathcal{H}}.$$

Since $V_{|\tilde{H}} : \tilde{\mathcal{H}} \to \mathcal{H}$ is isometric, we obtain that

$$f(x) = \langle (V_{|\tilde{\mathcal{H}}})^{-1} f, \Phi_0(x) \rangle_{\mathcal{H}_0} = \langle f, V_{|\tilde{\mathcal{H}}} \Phi_0(x) \rangle_{\mathcal{H}} = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}$$

for all $f \in \mathcal{H}$, $x \in X$, which is the reproducing property of k. Therefore, $\mathcal{H}$ is a RKHS by Proposition above.

Reproducing Kernel Hilbert Space

We next show that

$$\mathcal{H}_{pre} := \{\sum_{i=1}^{n} \alpha_i \, k(\cdot, x_i) : n \in \mathbb{N}, \alpha_1, \ldots, \alpha_n \in \mathbb{K}, x_1 \ldots, x_n \in X\}$$

is dense in any RKHS $\hat{\mathcal{H}}$ with k as the reproducing kernel. By the definition of reproducing kernel, we observe that $k(\cdot, x) \in \hat{\mathcal{H}}$ for all $x \in X$. Hence, $\mathcal{H}_{pre} \subset \hat{\mathcal{H}}$.

Now we suppose that $\mathcal{H}_{pre}$ is not dense in $\hat{\mathcal{H}}$. Then, $(\mathcal{H}_{pre})^{\perp} \neq \{0\}$. Therefore, there exists a function $g \in (\mathcal{H}_{pre})^{\perp}$ and a $x \in X$ with $g(x) \neq 0$. Since $g \in (\mathcal{H}_{pre})^{\perp}$ and $k(\cdot, x) \in \hat{\mathcal{H}}$, $\langle g, k(\cdot, x)\rangle = 0$. By the reproducing property of k, $\langle g, k(\cdot, x)\rangle = g(x) \neq 0$. This is a contradiction. Therefore, $\mathcal{H}_{pre}$ is dense in any RHKS.

Now, for any $f := \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) \in \mathcal{H}_{pre}$, by the reproducing property,

$$\|f\|_{\hat{\mathcal{H}}}^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \overline{\alpha_j} \langle k(\cdot, x_i), k(\cdot, x_j)\rangle_{\hat{\mathcal{H}}} = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \overline{\alpha_j} k(x_j, x_i).$$

Reproducing Kernel Hilbert Space

Finally we prove that k has only one RKHS.
Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two *RKHSs* of k.
We just proved that $\mathcal{H}_{pre}$ is dense in both $\mathcal{H}_1$ and $\mathcal{H}_2$ and that the norms of $\mathcal{H}_1$ and $\mathcal{H}_2$ coincide on $\mathcal{H}_{pre}$.
Choose $f \in \mathcal{H}_1$. There exists a sequence $(f_n) \subset \mathcal{H}_{pre}$ with $\|f_n - f\|_{\mathcal{H}_1} \to 0$. Since $\mathcal{H}_{pre} \subset \mathcal{H}_2$, the sequence (f_n) is also contained in $\mathcal{H}_2$, and since the norms of $\mathcal{H}_1$ and $\mathcal{H}_2$ coincide on $\mathcal{H}_{pre}$, the sequence (f_n) is a Cauchy sequence in $\mathcal{H}_2$. Therefore, there exists a $g \in \mathcal{H}_2$ with $\|f_n - g\|_{\mathcal{H}_2} \to 0$.
Since the convergence with respect to a RHKS norm implies pointwise convergence, then $f(x) = g(x)$ for all $x \in X$, hence $f \in \mathcal{H}_2$. Furthermore, $\|f_n - f\|_{\mathcal{H}_1} \to 0$ and $\|f_n - f\|_{\mathcal{H}_2} \to 0$ imply

$$\|f\|_{\mathcal{H}_1} = \lim_{n\to\infty} \|f_n\|_{\mathcal{H}_1} = \lim_{n\to\infty} \|f_n\|_{\mathcal{H}_{pre}} = \lim_{n\to\infty} \|f_n\|_{\mathcal{H}_2} = \|f\|_{\mathcal{H}_2}.$$

Therefore, $\mathcal{H}_1$ is isometrically included in $\mathcal{H}_2$.
Similarly, we can prove that $\mathcal{H}_2 \subset \mathcal{H}_1$. So the reproducing kernel k has a unique RKHS. □

Reproducing Kernel Hilbert Space

Remarks. The theorem describes the RKHS $\mathcal{H}$ of a given kernel k as the 'smallest' feature space of k in the sense that there exists a canonical metric surjection V from any other feature space $\mathcal{H}_0$ of k onto $\mathcal{H}$.

Recall that the nonlinear SVM approach produces decision functions of the form $x \mapsto \langle w, \Phi_0(x) \rangle$, where $\Phi_0 : X \to \mathcal{H}_0$ is a fature map of k and $w \in \mathcal{H}_0$ is an appropriate weight vector. The space

$$\mathcal{H} := \{ f : X \to \mathbb{K} \mid \exists w \in \mathcal{H}_0 \text{ with } f(x) = \langle w, \Phi_0(x) \rangle_{\mathcal{H}_0} \text{ for any } x \in X \}$$

given in the Theorem states that the RKHS associated with k consists exactly of all possible functions of such form.

Moreover, by the theorem, this set of functions dose not change if we consider different feature spaces or feature maps of k.

Properties of RKHSs - Boundedness

Recall that, for a function f on a topological space Z, the uniform norm of f is given by $\|f\|_b = \sup_{z \in Z} |f(z)|$.
A function f defined on Z is **bounded** if $\|f\|_b < \infty$.

For kernel functions k on $X \times X$, we introduce the norm $\| \cdot \|_\infty$

$$\|k\|_\infty := \sup_{x \in X} \sqrt{k(x,x)}.$$

Note that in general, $\|k\|_b \neq \|k\|_\infty$.
However, for a kernel k on X with RKHS $\mathcal{H}$, we have some remarkable properties.

Properties of RKHSs - Boundedness

Lemma *Let $k : X \times X \to \mathbb{K}$ be a kernel on a reproducing kernel Hilbert space $\mathcal{H}$ with the feature map $\Phi : X \to \mathcal{H}$. Then k is bounded iff*

$$\|k\|_\infty = \sup_{x \in X} \sqrt{k(x,x)} < \infty.$$

Proof.

By the reproducing property of $\mathcal{H}$, for any $x, x' \in X$,

$$k(x,x') = \langle \Phi(x'), \Phi(x) \rangle = \langle k(\cdot,x), k(\cdot,x') \rangle .$$

Hence, by the Cauchy-Schwarz inequality, we get that

$$\begin{aligned} |k(x,x')|^2 &= |\langle k(\cdot,x), k(\cdot,x') \rangle|^2 \\ &\leq \|k(\cdot,x)\|_{\mathcal{H}}^2 \cdot \|k(\cdot,x')\|_{\mathcal{H}}^2 \\ &= k(x,x) \cdot k(x',x'). \end{aligned}$$

Conversely

$$\sup_{(x,x') \in X \times X} |k(x,x')| \geq \sup_{(x,x) \in X \times X} k(x,x)$$

and the proof follows by choosing $x' = x$. □

Properties of RKHSs - Boundedness

We can relate the boundedness of k to the boundedness of its feature map Φ.
Lemma *Let $k : X \times X \to \mathbb{K}$ be a kernel on a reproducing kernel Hilbert space $\mathcal{H}$ with feature space $\mathcal{H}_0$ and feature map $\Phi : X \to \mathcal{H}_0$. Then k is bounded iff Φ is bounded.*

Proof. Since $\Phi : X \to \mathcal{H}_0$ is a feature map for k, by the reproducing property

$$\|\Phi(x)\|_{\mathcal{H}_0}^2 = \langle \Phi(x), \Phi(x) \rangle_{\mathcal{H}_0} = \langle k(\cdot, x), k(\cdot, x) \rangle_{\mathcal{H}} = k(x, x).$$

Taking the supremum over X on both sides gives $\|\Phi\|_b^2 = \|k\|_\infty^2$. Thus, $\|\Phi\|_b^2 < \infty$ iff $\|k\|_\infty^2 < \infty$.

□

Properties of RKHSs - Boundedness

We can now characterize the boundedness of the reproducing kernel in terms of the feature space elements $f \in \mathcal{H}$.

Lemma *Let* $k : X \times X \to \mathbb{K}$ *be a kernel on* X *with a RKHS* $\mathcal{H}$. *Then* k *is bounded iff every* $f \in \mathcal{H}$ *is bounded.*

Moreover, in this case the induction map $(id) : \mathcal{H} \to \ell^\infty(X)$ *is continuous, with*

$$\|(id) : \mathcal{H} \to \ell^\infty(X)\| = \|k\|_\infty.$$

Proof. ($\implies$) Assume k is bounded. By virtue of the properties of RKHS's for $\mathcal{H}$ and the Cauchy-Schwarz inequality, we have for all $x \in X$ and $f \in \mathcal{H}$,

$$|f(x)|^2 = |\langle f, k(\cdot, x)\rangle_{\mathcal{H}}|^2 \leq \|f\|_{\mathcal{H}}^2 \ k(x, x).$$

Taking the supremum over X gives

$$\|f\|_b \leq \|f\|_{\mathcal{H}} \ \|k\|_\infty.$$

Since k is assumed bounded and $f \in \mathcal{H} \implies \|f\|_{\mathcal{H}} < \infty$, then we have $\|f\|_b < \infty$, showing boundedness.

Properties of RKHSs - Boundedness

This also shows that $(id) : \mathcal{H} \to \ell^\infty(X)$ is well-defined, and that

$$\|(id) : \mathcal{H} \to \ell^\infty(X)\| \leq \|k\|_\infty.$$

$(\Longleftarrow)$ If every $f \in \mathcal{H}$ is bounded, the inclusion $(id) : \mathcal{H} \to \ell^\infty(X)$ is well-defined.
(id) is a linear map since, for any $\alpha \in \mathbb{K}$ and $f, g \in \mathcal{H}$,

$$(id)(\alpha f + g)(x) = (\alpha f + g)(x) = \alpha f(x) + g(x) = \alpha (id)(f)(x) + (id)(g)(x).$$

We will use the Closed Graph Theorem to prove that (id) is bounded. For that, let $(f_n)_{n=1}^\infty \subset \mathcal{H}$ be such that

$$\lim_{n\to\infty} \|f_n - f\|_{\mathcal{H}} = 0$$
$$\text{and} \quad \lim_{n\to\infty} \|id(f_n) - g\|_\infty = \lim_{n\to\infty} \|f_n - g\|_\infty = 0$$

for some $f \in \mathcal{H}$ and $g \in \ell^\infty(X)$.

Properties of RKHSs - Boundedness

Then, we have that for any $x \in X$,

$$\begin{aligned}
&\lim_{n\to\infty} |f_n(x) - f(x)|^2 \\
&= \lim_{n\to\infty} |\langle f_n, k(\cdot, x)\rangle_{\mathcal{H}} - \langle f, k(\cdot, x)\rangle_{\mathcal{H}}|^2 && \text{(reproducing kernel property)} \\
&= \lim_{n\to\infty} |\langle f_n - f, k(\cdot, x)\rangle_{\mathcal{H}}|^2 && \text{(inner product)} \\
&\leq \lim_{n\to\infty} \|f_n - f\|_{\mathcal{H}}^2 \|k(\cdot, x)\|_{\mathcal{H}}^2 && \text{(Cauchy-Schwarz inequality)} \\
&= 0. && \text{(by hypothesis)}
\end{aligned}$$

Also, since $|f_n(x) - g(x)| \leq \|f_n - g\|_\infty$ for any $x \in X$, it follows from our assumption on (f_n) that $\lim_{n\to\infty} |f_n(x) - g(x)| = 0$ for every $x \in X$, implying that $f(x) = g(x)$ for all $x \in X$ iff $f = (id)(f) = g$.

Thus, $(id) : \mathcal{H} \to \ell^\infty(X)$ has a closed graph and hence is bounded. Since (id) is linear, it is also continuous.

Properties of RKHSs - Boundedness

Finally, for any $x \in X$ we have that

$$\begin{aligned} |k(x,x)| \leq \|k(\cdot,x)\|_\infty &\leq \|(id) : \mathcal{H} \to \ell^\infty(X)\| \|k(\cdot,x)\|_{\mathcal{H}} \\ &= \|(id) : \mathcal{H} \to \ell^\infty(X)\| \sqrt{(k(x,x)} \end{aligned}$$

which implies that

$$\sqrt{k(x,x)} \leq \|(id) : \mathcal{H} \to \ell^\infty(X)\|.$$

Since this holds for every $x \in X$, taking the sup over X on both sides gives that

$$\|k\|_\infty \leq \|(id) : \mathcal{H} \to \ell^\infty(X)\|. \tag{6}$$

By a Lemma above, this shows that k is bounded. □

Properties of RKHSs - Measurability

The measurability of a kernel k can be characterized in terms of the measurability of the functions in the associated RKHS.

Lemma *Let (X, μ) be a measurable space and k be a kernel on X with reproducing kernel Hilbert space $\mathcal{H}$. Every $f \in \mathcal{H}$ is μ-measurable iff the restricted kernel function $k(\cdot, x') : X \to \mathbb{R}$ is μ-measurable for all $x' \in X$.*

Proposition. *Let (X, μ) be a measurable space and k be a kernel on X with reproducing kernel Hilbert space $\mathcal{H}$ such that the restricted kernel function $k(\cdot, x') : X \to \mathbb{R}$ is μ-measurable for all $x' \in X$. If $\mathcal{H}$ is separable, then*

(i) *the canonical feature map $\Phi : X \to \mathcal{H}$ is μ-measurable,*

(ii) *the full kernel $k : X \times X \to \mathbb{R}$ is $\mu \times \mu$-measurable on the product space $X \times X$.*

Properties of RKHSs - Integrability

Theorem *Let (X, μ) be a measurable space, μ be a σ-finite measure on X, and $\mathcal{H}$ be a separable RKHS over X with measurable kernel $k : X \times X \to \mathbb{R}$.*
If there exists $p \in [1, \infty)$ such that

$$\|k\|_{L^p} := \left(\int_X k(x,x)^{p/2} \, d\mu(x) \right)^{1/p} < \infty,$$

then the following holds:

(i) *$\mathcal{H}$ consists of $L^p(\mu)$-integrable functions.*

(ii) *The inclusion map $(id) : \mathcal{H} \to L^p(\mu)$ is continuous.*

(iii) *The adjoint of the inclusion map exists. It is the operator $S_k : L^{p'} \to \mathcal{H}$ given by*

$$S_k g(x) = \int_X k(x,x') g(x') \, d\mu(x')$$

for $g \in L^{p'}$, $x \in X$, and conjugate exponents $\frac{1}{p} + \frac{1}{p'} = 1$.

Note: The L^p norm notation here is not the standard one.

Properties of RKHSs - Integrability

Proof. (i),(ii) Fix $f \in \mathcal{H}$. Since $\|k(\cdot, x)\|_{\mathcal{H}} = \sqrt{k(x,x)}$ then

$$\begin{aligned} \|f\|_{L^p}^p &= \int_X |f(x)|^p \, d\mu(x) \\ &= \int_X |\langle f, k(\cdot, x)\rangle|^p \, d\mu(x) \\ &\leq \|f\|_{\mathcal{H}}^p \int_X (k(x,x))^{p/2} \, d\mu(x) \\ &= \|f\|_{\mathcal{H}}^p \, \|k\|_{L^p}^p \end{aligned}$$

This shows that $f \in L^p(\mu)$ and that $(id) : \mathcal{H} \to L^p(\mu)$ is continuous with

$$\|(id) : \mathcal{H} \to L^p(\mu)\| \leq \|k\|_{L^p}.$$

Properties of RKHSs - Integrability

(iii) For $g \in L^{p'}$, using Cauchy-Schwartz and Hölder's inequalities we have that

$$\begin{aligned}\int_X |k(x,x')g(x')|\, d\mu(x') &\leq \sqrt{k(x,x)} \int_X \sqrt{k(x',x')}\left|g(x')\right|\, d\mu(x') \\ &\leq \sqrt{k(x,x)}\|k\|_{L^p}\|g\|_{L^{p'}}\end{aligned}$$

This shows the integrability of $k(x,x')g(x')$ and thus the existence of the integral defining $S_k g(x)$ for all $x \in X$.
Since $\sqrt{k(x',x')} = \|\Phi(x')\|_{\mathcal{H}}$, the last inequality shows that $x' \to \|\Phi(x')g(x')\|_{\mathcal{H}}$ is integrable, Finally, we have

$$\begin{aligned}S_k g(x) &= \int_X \langle \Phi(x'),\ \Phi(x) \rangle_{\mathcal{H}}\, g(x')\, d\mu(x') \\ &= \left\langle \int_X g(x')\Phi(x')\, d\mu(x'),\ \Phi(x) \right\rangle_{\mathcal{H}}.\end{aligned}$$

This shows that $S_k g := \bar{g} = \int_X g(x')\Phi(x')\, d\mu(x') \in \mathcal{H}$. □

Properties of RKHSs - Integrability

Remark: Under the conditions of Theorem above, using the fact that a bounded linear operator has a dense image if and only if its adjoint is injective, one can also derive the following properties for the feature space $\mathcal{H}$ in terms of the adjoint map S_k

1. $\mathcal{H}$ is dense in L^p iff the adjoint operator $S_k : L^{p'} \to \mathcal{H}$ is injective.
2. The adjoint $S_k : L^{p'} \to \mathcal{H}$ has a dense image $S_k(L^{p'})$ iff the inclusion $(id) : \mathcal{H} \to L^p$ is injective.

Properties of RKHSs - Integrability

Theorem *Let (X, μ) be a measurable space, μ be a σ-finite measure on X, and $\mathcal{H}$ be a separable RKHS over X with measurable kernel $k : X \times X \to \mathbb{R}$ such that*

$$\|k\|_{L^2} = \left(\int_X k(x,x) \, d\mu(x) \right)^{1/2} < \infty.$$

Then

(i) *$S_k : L^2 \to \mathcal{H}$ given by*

$$S_k g(x) = \int_X k(x,x') g(x') \, d\mu(x')$$

for $g \in L^2$, $x \in X$, is a Hilbert-Schmidt operator with

$$\|S_k\|_{HS} = \|k\|_{L^2};$$

(ii) *the integral operator $T_k = S_k^* S_k : L^2(\mu) \to L^2(\mu)$ is compact, positive, self-adjoint.*

Recall that $\|S\|_{HS}^2 := \sum_i \|Se_i\|_{L^2}^2$ where $\{e_i\} \subseteq \mathcal{H}$ is an ONB of $\mathcal{H}$.

Properties of RKHSs - Continuity

We define a (pseudo)-metric in terms of k and use this to characterize continuity.

Definition *Let X be a topological vector space. A kernel k on X is* **separately continuous** *if $k(\cdot, x) : X \to \mathbb{R}$ is continuous for all $x \in X$.*

Lemma *Let X be a topological space and k a kernel on X with reproducing kernel Hilbert space $\mathcal{H}$.*
Then k is bounded and separately continuous iff every $f \in \mathcal{H}$ is bounded and continuous. In this case, the inclusion map $id : \mathcal{H} \to C_b(X)$ is continuous and

$$\|id : \mathcal{H} \to C_b(X)\| = \|k\|_\infty.$$

Properties of RKHSs - Continuity

Definition. *Let k be a kernel on X with a feature map $\Phi : X \to \mathcal{H}$.*
The **kernel metric** *is given by;*

$$d_k(x, x') = \|\Phi(x) - \Phi(x')\|_{\mathcal{H}} \qquad x,\ x' \in X.$$

We remark that d_k is a *pseudo-metric* in general since $d_k(x, x') = 0$ for not imply that $x = x'$ in general.
It is a metric if Φ is injective.

Furthermore, we have

$$d_k(x, x') = \sqrt{k(x, x) - 2k(x, x') + k(x', x')}$$

showing that the definition of d_k is independent of Φ.

Properties of RKHSs - Continuity

The following act shows how the kernel metric can be used to characterize the continuity of the kernel k.

Proposition. *Let (X, τ) be a topological vector space and k a kernel on X with feature space $\mathcal{H}$ and feature map Φ. The following are equivalent:*

i. *k is continuous.*

ii. *k is separately continuous and $x \mapsto k(x, x)$ is continuous.*

iii. *Φ is continuous.*

iv. *The map $id : (X, \tau) \to (X, d_k)$ is continuous.*

Properties of RKHSs - Continuity

Proof.

(i) $\Longrightarrow$ *(ii).* Trivial.

(ii) $\Longrightarrow$ *(iv).* By the formula of d_k in terms of the kernel and the assumption, we see that $d_k(\cdot, x) : (X, \tau) \to \mathbb{R}$ is continuous for every $x \in X$.

Consequently, $\{x' \in X : d_k(x', x) < \epsilon\}$ is open with respect to τ and therefore $id : (X, \tau) \to (X, d_k)$ is continuous.

(iv) $\Longrightarrow$ *(iii).* This follows from the fact that $\Phi : (X, d_k) \to \mathcal{H}$ is continuous.

(iii) $\Longrightarrow$ *(i).* Fix $x_1,\ x_1' \in X$ and $x_2,\ x_2' \in X$. Then we have

$$|k(x_1, x_1') - k(x_2, x_2')| \leq |\langle \Phi(x_1'), \Phi(x_1) - \Phi(x_2) \rangle| + |\langle \Phi(x_1') - \Phi(x_2'), \Phi(x_2) \rangle|$$

$$\leq \|\Phi(x_1')\| \cdot \|\Phi(x_1) - \Phi(x_2)\| + \|\Phi(x_2)\| \cdot \|\Phi(x_1') - \Phi(x_2')\|.$$

From this we conclude that k is continuous. □

Properties of RKHSs - Compactness

We have seen above that a RKHS over X is continuously contained in ℓ^∞ if it has a bounded kernel.

The following proposition provides an additional condition so that this inclusion is compact

Proposition. *Let k be a kernel on a space X with RKHS $\mathcal{H}$ and canonical feature map $\Phi : X \to \mathcal{H}$. If $\Phi(X)$ is compact in $\mathcal{H}$ then the inclusion map given by*

$$id : \mathcal{H} \to \ell^\infty(X)$$

is also compact.

Properties of RKHSs - Compactness

Proof. Since $\Phi(X)$ is compact, then k is bounded and the space (X, d_k) is compact with respect to the kernel metric d_k.

Let $C(X, d_k)$ be the space of functions $f : X \to \mathbb{R}$ that are continuous with respect to d_k.
For $x,\ x' \in X$ and $f \in$ we have

$$|f(x) - f(x')| = |\langle f, \Phi(x) - \Phi(x')\rangle| \leq \|f\|_{\mathcal{H}} \cdot d_k(x, x'),$$

showing that f is continuous on (X, d_k).
It follows that the unit ball $B_{\mathcal{H}} \subset \mathcal{H}$ is equicontinuous and bounded.

By Arzela-Ascoli Theorem, $\overline{B_{\mathcal{H}}}$ is compact in $C(X, d_k)$ and, hence, in $\ell^\infty(X)$ since $C(X, d_k) \subset \ell^\infty(X)$.
This shows that $id : \mathcal{H} \to \ell^\infty(X)$ is compact. □

Properties of RKHSs - Compactness

We establish a sufficient condition for the separability of a RKHS.

Proposition. *Let X be a separable topological space and k a continuous kernel on X.*
Then the RKHS $\mathcal{H}$ of k is separable.

Proof. By the Proposition about compactness above, the canonical feature map $\Phi : X \to \mathcal{H}$ is continuous, thereby implying that $\Phi()$ is separable. It follows that vector space

$$\mathcal{H}_{pre} := \{\sum_{i=1}^{n} \alpha_i \, k(\cdot, x_i) : n \in \mathbb{N}, \alpha_1, \ldots, \alpha_n \in k, x_1, \ldots, x_n \in X\}$$

is also separable.
We observed in the proof of a previous theorem that $\mathcal{H}_{pre}$ is dense in $\mathcal{H}$. Hence, the separability of $\mathcal{H}$ follows by completion. □

Properties of RKHSs - Mercer's Theorem

The celebrated Mercer's Theorem shows the existence of a series representation for continuous kernels that are defined on a compact domain.

This series representation can be used to characterize the corresponding RKHSs.

Recall: we proved that the integral operator $T_k = S_k^* S_k$, where $S_k : L^2 \to \mathcal{H}$ is given by

$$S_k g(x) = \int_X k(x, x') g(x') \, d\mu(x')$$

is compact, positive and self-adjoint.
By the Spectral Theorem, there exists a countable ONS $(e_i) \subset L^2$ and a family $(\lambda_i) \subset \mathbb{R}$ converging to 0 such that, for $f \in L^2$

$$T_k f = \sum_i \lambda_i \langle f, e_i \rangle \, e_i .$$

Properties of RKHSs - Mercer's Theorem

In addition, $\{\lambda_i : i \in I\}$ is a set of non-zero eigenvalues of T_k.

Set $\tilde{e}_i := \lambda_i^{-1} S_k e_i \in \mathcal{H}$, $i \in I$.
It follows that $S_k^* \tilde{e}_i = \lambda_i^{-1} T_k e_i = e_i$ (using the fact that λ_i is an eigenvalue of T_k) and $\lambda_i \tilde{e}_i = S_k e_i$ for all $i \in I$.
From this, we have that

$$\begin{aligned} \lambda_i \lambda_j \langle \tilde{e}_i, \tilde{e}_j \rangle_{\mathcal{H}} = \langle S_k e_i, S_k e_j \rangle_{\mathcal{H}} = \langle e_i, S_k^* S_k e_j \rangle_{L^2} &= \langle e_i, T_k e_j \rangle_{L^2} \\ &= \lambda_j \langle e_i, e_j \rangle_{L^2} \end{aligned}$$

This shows that the set $(\sqrt{\lambda_i}\tilde{e}_i)_{i \in I}$ is an ONS in $\mathcal{H}$.

Mercer's Theorem shows that, under certain conditions, this set is an ONB of $\mathcal{H}$.

Properties of RKHSs - Mercer's Theorem

Theorem (Mercer's). *Let X be compact metric space and $k : X \times X \to \mathbb{R}$ be continuous. Let μ be a finite Borel measure with supp$(\mu) = X$.*
Then there exists a countable orthonormal sequence $(e_i)_{i \in I} \subset \mathcal{H}$ and a family $(\lambda_i)_{i \in I} \subset \mathbb{R}$ converging to 0 such that

$$k(x, x') = \sum_{i \in I} \lambda_i \, e_i(x) \, e_i(x') \quad x, x' \in X$$

with absolute and uniform convergence.
Note that above we assumed

$$|\lambda_1| \geq |\lambda_2| \geq |\lambda_3| \geq \dots$$

Properties of RKHSs - Mercer's Theorem

Remarks:

- Mercer's Theorem implies that $\Phi : X \mapsto \ell^2$ given by

  $$\Phi(x) = (\sqrt{\lambda_i} e_i(x))_{i \in I}, x \in X,$$

  is a feature map of k with $k(x, x') = \langle \Phi(x'), \Phi(x) \rangle$.
- With the assumptions of Mercer's theorem, if $(a_i)_{i \in I} \subset \ell^2(I)$ and $x \in X$, $J \subset I$, then

  $$\sum_{i \in J} |a_i \sqrt{\lambda_i} e_i(x)| \leq \sqrt{\sum_{i \in J} a_i^2} \sqrt{\sum_{i \in J} \lambda_i e_i^2(x)} = \|(a_i)\|_{\ell^2(I)} \cdot \sqrt{k(x, x)}.$$

Properties of RKHSs - Mercer's Theorem

Theorem (Mercer's Representation theorem for RKHS)

With the assumptions from previous theorem, let

$$H := \left\{ \sum_{i \in I} a_i \sqrt{\lambda_i}\; e_i : (a_i) \in \ell^2(I) \right\}.$$

For

$$f = \sum a_i \sqrt{\lambda_i} e_i \in H, \quad g = \sum b_i \sqrt{\lambda_i} e_i \in H$$

set

$$\langle f, g \rangle_H = \sum_{i \in I} a_i b_i .$$

Then H equipped with $\langle \cdot, \cdot \rangle_H$ is the RKHS of the kernel of k. Furthermore, $T_k^{1/2} : L^2(\mu) \to H$ given by

$$T_k^{1/2} f = \sum_{i \in I} \langle f, e_i \rangle \sqrt{\lambda_i} e_i$$

is an isometric isomorphism.

Properties of RKHSs - Mercer's Theorem

Proof. A direct and straightforward argument shows that H is a complete inner product space under the norm $\langle\cdot,\cdot\rangle_H$, hence H is a Hilbert Space.

For $x \in X$, by Mercer's Theorem, we have that

$$k(\cdot,x) = \sum_{i\in I} \sqrt{\lambda_i} e_i(x) \sqrt{\lambda_i} e_i(\cdot)$$

showing that $k(\cdot,x) \in H$.

Additionally, for $f = \sum_{i\in I} a_i \sqrt{\lambda_i} e_i \in H$, we have

$$\langle f, k(\cdot,x)\rangle_H = \sum_{i\in I} a_i \sqrt{\lambda_i} e_i(x) = f(x), \qquad x \in X$$

showing that k is the reproducing kernel of H.

Properties of RKHSs - Mercer's Theorem

Let us next examine $T_k^{1/2}$.

Fix $f \in L^2(\mu)$.
Since $(e_i)_{i \in I}$ is an orthonormal basis of $L^2(\mu)$, we can write

$$f = \sum_{i \in I} \langle f, e_i \rangle_{L^2(\mu)} \, e_i .$$

By Parseval's formula,

$$\|f\|_{L^2(\mu)}^2 = \sum_{i \in I} |\langle f, e_i \rangle|^2,$$

showing that $(\langle f, e_i \rangle)_{i \in I} \subset \ell^2(I)$.
It follows that

$$T_k^{1/2} f = \sum_{i \in I} \langle f, e_i \rangle \sqrt{\lambda_i} \, e_i \in H.$$

Properties of RKHSs - Mercer's Theorem

Moreover,

$$\| T_k^{1/2} f \|^2 = \sum_{i \in I} |\langle f, e_i \rangle|^2 = \|f\|^2_{L^2}(\mu),$$

implying that $T_k^{1/2}$ is an isometry on H and hence injective.

To show that $T_k^{1/2}$ is surjective, fix $f \in H$.
By the definition of H, there is a sequence $(a_i)_{i \in I} \subset \ell^2(I)$ such that $f(x) = \sum_{i \in I} a_i \sqrt{\lambda_i}\, e_i(x)$.
Also, we have that $g := \sum_{i \in I} a_i e_i \in L^2(\mu)$ and, thus, $a_i = \langle g, e_i \rangle_{L^2}$.
Therefore,

$$T_k^{1/2} g(x) = \sum_{i \in I} \langle g, e_i \rangle_{L^2} \sqrt{\lambda_i}\, e_i(x) = \sum_{i \in I} a_i \sqrt{\lambda_i}\, e_i(x) = f(x),$$

proving that $T_k^{1/2}$ is surjective. □

Universal Kernels

The 'size' of the RKHS is a critical issue on the generalization ability of an SVM since we typically desire a solution space large enough to give accurate solutions, yet not too large to avoid over-fitting.

Definition. A continuous kernel k on a compact metric space X is **universal kernel** if the RKHS $\mathcal{H}$ of k is dense in $C(X)$, i.e., for every $g \in C(X)$ and all $\epsilon > 0$, there exists an $f \in \mathcal{H}$ such that

$$\|f - g\|_\infty \leq \epsilon.$$

Definition. Let k be a kernel on a metric space X with RKHS $\mathcal{H}$. We say that k **separates** the disjoint sets $A,\ B \subset X$, if there exists an $f \in \mathcal{H}$ such that $f(x) > 0$ for all $x \in A$, and $f(x) < 0$ for all $x \in B$. We say that k separates all finite (or compact) sets if k separates all finite (or compact) disjoints sets $A,\ B \subset X$.

Universal Kernels

Theorem. Let X be a compact metric space and k a universal kernel on X. Then k separates all compact sets in X.

Theorem (Test for universality). Let X be a compact metric space and k a continuous kernel on X with the property that $k(x, x) > 0$ for all $x \in X$.
Suppose that we have an injective feature map $\Phi X :\to \ell^2$ and denote $\Phi(x) = (\phi_1(x), \phi_2(x), ..., \phi_k(x), ...)$, $x \in X$. If $\mathcal{A} = span\{\phi_n : n \in \mathbb{N}\}$ is an algebra, then k is universal.

Universal Kernels

Examples of universal Kernels

- Polynomial: $k(x, x') = f(\langle x, x' \rangle)$ where $f(t) = \sum_{k=0}^{\infty} a_k t^k, \quad a_k > 0$.
- Exponential: $k(x, x') = exp(\langle x, x' \rangle)\rangle$.
- Gaussian RBF: $k_\gamma(x, x') = exp(-\gamma^2 \|x - x'\|_L^2)$