Machine learning

5. SVM

V. Lefieux

April 2019

Table of contents

Table of contents

SVM ?

- Support Vector Machine (large machine classifiers) are supervised learning algorithms.
- Available for classification and regression.
- These algorithms are required to have good generalization properties: compromise between estimation and prediction.

Presentation

- In the case of a binary classification $\mathcal{Y} = \{-1, 1\}$ with $\mathcal{X} = \mathbb{R}^d$.
- The SVM (Vapnik) approach can be seen as a generalization of "optimal hyperplane search".

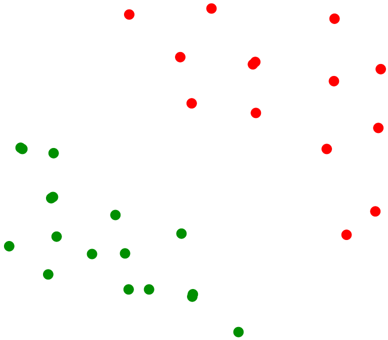
Simple case

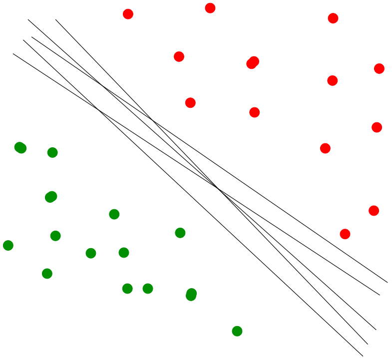
Data $(x_1, y_1), \ldots, (x_n, y_n)$ are linearly separable if there exists $(w, b) \in \mathbb{R}^d \times \mathbb{R}$ such that for all i:

- $y_i = 1$ if $\langle w, x_i \rangle + b > 0$,
- $y_i = -1$ if $\langle w, x_i \rangle + b < 0$.

i.e.:

$$\forall i \in \{1, \ldots, n\} : y_i \left(\langle w, x_i \rangle + b \right) > 0 \; .$$



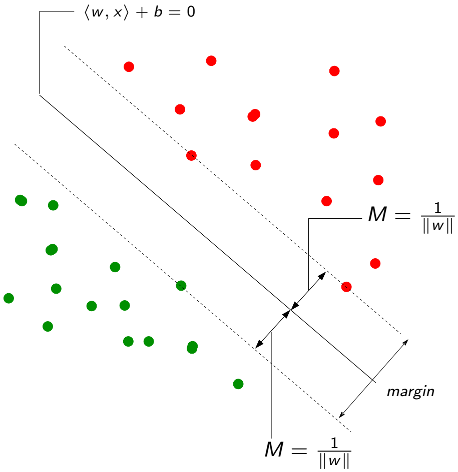


Problem

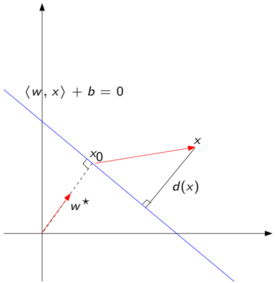
There is an infinite of infinite of separating hyperplanes so an infinite of potential classification rules.

Solution

Vapnik proposes to choose the hyperplane with the largest margin.



Margin



- $d(x) = \frac{\|\langle w,x\rangle + b\|}{\|w\|} = x^\top w + b$ if $\|w\| = 1$.
- if (w, b) is the optimal separating hyperplan, its margin is:

$$\min_{i\in\{1,\ldots,n\}} y_i \left(w_i^\top x_i + b\right) .$$

Optimal separating hyperplan

Solution of the constrained optimization problem:

$$\max_{w,b,\|w\|=1} M$$

$$\text{under constraint} \quad \forall i \in \{1, \ldots, n\} : y_i \left(w^\top x_i + b\right) \geq M .$$

Rewriting

- If we remove the constraint $\|w\| = 1$, then we seek the largest margin such that:

$$\forall i \in \{1, \ldots, n\} : y_i \left(w^\top x_i + b \right) \geq \|w\| \, M \ .$$

- With the new constraint $\|w\| = \frac{1}{M}$,optimization problem becomes:

Primal problem

$$\min_{w,b} \frac{1}{2} \|w\|^2$$

under constraint $\quad \forall i \in \{1, \ldots, n\} : y_i \left(w^\top x_i + b \right) \geq 1 \ .$

  - Problem of convex optimization under linear constraints.
  - Existence of a optimum global obtained by resolution of the dual problem.

Resolution by lagrangian method I

- The lagrangien of the primal problem is:

$$L_p(w, b; \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i \left[y_i(x_i^\top w + b) - 1 \right] .$$

- Considering partial derivatives with respect of w and b:

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i \quad \text{et} \quad \sum_{i=1}^{n} \alpha_i y_i = 0 .$$

Dual problem

By substituting these two equations in L_p, we obtain the dual problem which consists in maximizing:

$$L_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{n} \alpha_i \alpha_k y_i y_k x_i^\top x_k$$

under contraints $\alpha_i \geq 0$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$.

Resolution by lagrangian method II

- We note $\alpha^\star$ the dual problem solution.
- We then deduce:

$$w^\star = \sum_{i=1}^{n} \alpha_i^\star y_i x_i \ .$$

Karush-Kuhn-Tucker conditions (KKT)

- $\forall i \in \{1, \ldots, n\} : \alpha_i^\star \geq 0 \ ,$
- $\forall i \in \{1, \ldots, n\} : \alpha_i^\star \left[y_i \left(x_i^\top w + b \right) - 1 \right] = 0 \ .$

- We obtain $b^\star$ by solving:

$$\alpha_i^\star \left[y_i \left(x_i^\top w + b \right) - 1 \right] = 0$$

for $\alpha_i^\star$ non-zero.

Support vectors

- x_i such that $\alpha_i^\star > 0$ check:

$$y_i \left(x_i^\top w^\star + b^\star \right) = 1 \ .$$

- So they are on the border with the maximal margin. These points are called support vectors.

Support vectors

- x_i such that $\alpha_i^\star > 0$ check:

$$y_i \left(x_i^\top w^\star + b^\star \right) = 1 \ .$$

- So they are on the border with the maximal margin. These points are called support vectors.

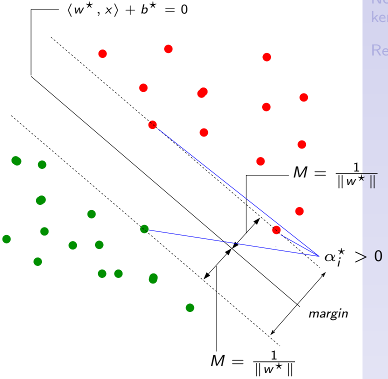
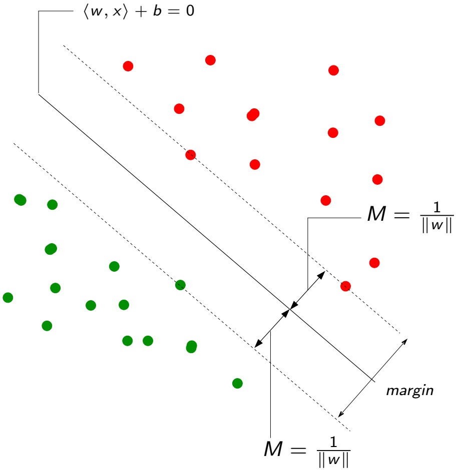


Table of contents

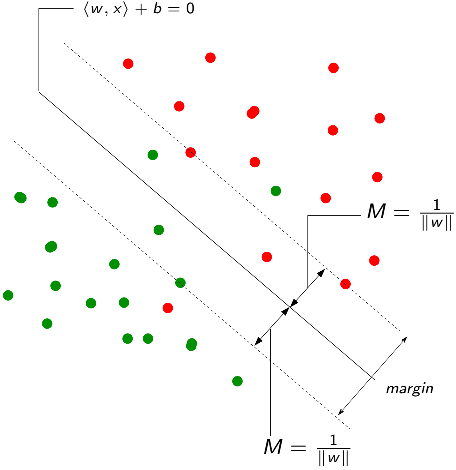
Problem

In the majority of cases, data aren't linearly separable. . .



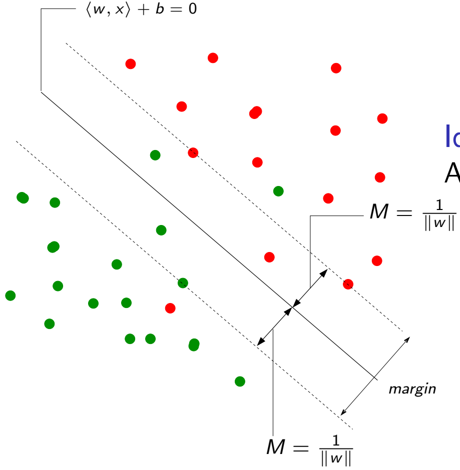
Problem

In the majority of cases, data aren't linearly separable...



Problem

In the majority of cases, data aren't linearly separable...



Idea

Allow certain points

1. to be well classified but within the margin,
2. and/or to be misclassified.

Slack variables

Reminder: separable case

$$\min_{w,b} \frac{1}{2} \|w\|^2$$

under constraint $\quad \forall i \in \{1, \dots, n\} : y_i \left(w^\top x_i + b \right) \geq 1 \ .$

- The contraints $y_i \left(w^\top x_i + b \right) \geq 1$ mean that all points are outside the boundary defined by the margin.

Slack variables

Reminder: separable case

$$\min_{w,b} \frac{1}{2} \|w\|^2$$

under constraint $\quad \forall i \in \{1, \ldots, n\} : y_i \left(w^\top x_i + b\right) \geq 1 \ .$

- The contraints $y_i \left(w^\top x_i + b\right) \geq 1$ mean that all points are outside the boundary defined by the margin.
- Non separable case: the problem below does not admit a solution !

Slack variables creation !

Slack variables

To allow certain points to be "in the margin" and/or misclassified, we define positive slack variables $(\xi_1, \ldots, \xi_n)$ such as $y_i \left(w^\top x_i + b \right) \geq 1 - \xi_i$. 2 cases are to be distinguished:

1. $\xi_i \in [0, 1] \Rightarrow$ well classified but in the region defined by the margin.
2. $\xi_i > 1 \Rightarrow$ misclassified.

Slack variables

- Of course, we would like to have the maximum of slack variables ξ_i zero.
- When $\xi_i > 0$, one expect ξ_i to be the smallest possible.

Non separable case: primal optimization problem

- We want to find the (w, b, ξ) value that minimize:

$$\frac{1}{2}\|w\|^2$$

under the constraints:

$$\begin{cases} y_i\left(w^\top x_i + b\right) \geq 1 \quad , \; i = 1, \ldots, n. \\ \end{cases}$$

Slack variables

- Of course, we would like to have the maximum of slack variables ξ_i zero.
- When $\xi_i > 0$, one expect ξ_i to be the smallest possible.

Non separable case: primal optimization problem

- We want to find the (w, b, ξ) value that minimize:

$$\frac{1}{2}\|w\|^2$$

under the constraints:

$$\begin{cases} y_i\left(w^\top x_i + b\right) \geq 1 \quad , \; i = 1, \ldots, n. \\ \xi_i \geq 0, i = 1, \ldots, n. \end{cases}$$

Slack variables

- Of course, we would like to have the maximum of slack variables ξ_i zero.
- When $\xi_i > 0$, one expect ξ_i to be the smallest possible.

Non separable case: primal optimization problem

- We want to find the (w, b, ξ) value that minimize:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$$

under the constraints:

$$\begin{cases} y_i\left(w^\top x_i + b\right) \geq 1 - \xi_i, \ i = 1, \ldots, n. \\ \xi_i \geq 0, i = 1, \ldots, n. \end{cases}$$

- $C > 0$ must be calibrated (cost parameter): $C \nearrow \Longrightarrow$ adjustment $\nearrow$.

Slack variables

- Of course, we would like to have the maximum of slack variables ξ_i zero.
- When $\xi_i > 0$, one expect ξ_i to be the smallest possible.

Non separable case: primal optimization problem

- We want to find the (w, b, ξ) value that minimize:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$$

under the constraints:

$$\begin{cases} y_i\left(w^\top x_i + b\right) \geq 1 - \xi_i, \ i = 1, \ldots, n. \\ \xi_i \geq 0, i = 1, \ldots, n. \end{cases}$$

- $C > 0$ must be calibrated (cost parameter): $C \nearrow \Longrightarrow$ adjustment $\nearrow$.
- The separable case corresponds to $C = +\infty$.

Optimization solution

- The solutions of this new optimization problem are obtained in the same way: lagrangian, dual problem...

KKT conditions

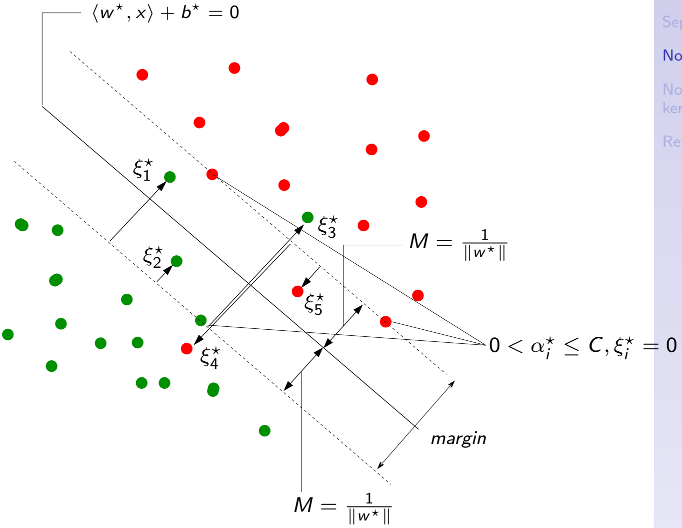
1. $0 \leq \alpha_i^\star \leq C$.
2. $y_i(\langle w^\star, x_i\rangle + b^\star) \geq 1 - \xi_i^\star$.
3. $\alpha_i^\star(y_i(\langle w^\star, x_i\rangle + b^\star) + \xi_i^\star - 1) = 0$.
4. $\xi_i^\star(\alpha_i^\star - C) = 0$.

Support vectors

- x_i such as $\alpha_i^\star > 0$ are support vectors.
- There are two types:
  1. those on the border defined by the margin: $\xi_i^\star = 0$;
  2. those outside: $\xi_i^\star > 0$ and $\alpha_i^\star = C$.
- Non support vectors check $\alpha_i^\star = 0$ et $\xi_i^\star = 0$.

Classification rule

- As in the separable case, the optimal hyperplane is defined by:
$$w^\star = \sum_{i=1}^{n} \alpha_i^\star y_i x_i$$
and $b^\star$ is the solution of $y_i \left(\langle w^\star, x_i \rangle + b^\star \right) = 1$ for any i such as $0 < \alpha_i^\star < C$.
- We deduce the SVM classification rule:
$$g(x) = \mathbb{1}_{\langle w^\star, x \rangle + b^\star \geq 0} - \mathbb{1}_{\langle w^\star, x \rangle + b^\star < 0} \ .$$
- The maximum margin is $\frac{1}{\|w^\star\|} = \left[\sum_{i=1}^{n} \left(\alpha_i^\star \right)^2 \right]^{-1/2}$.



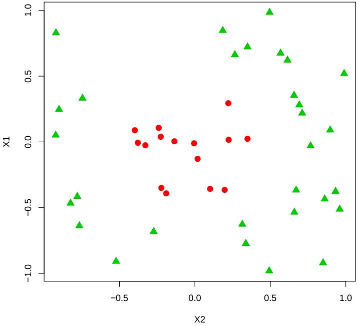
$\langle w^\star, x\rangle + b^\star = 0$
$\xi_1^\star$
$\xi_2^\star$
$\xi_3^\star$
$\xi_4^\star$
$\xi_5^\star$
$M = \frac{1}{\|w^\star\|}$
$0 < \alpha_i^\star \leq C, \xi_i^\star = 0$
margin
$M = \frac{1}{\|w^\star\|}$

Choice of C

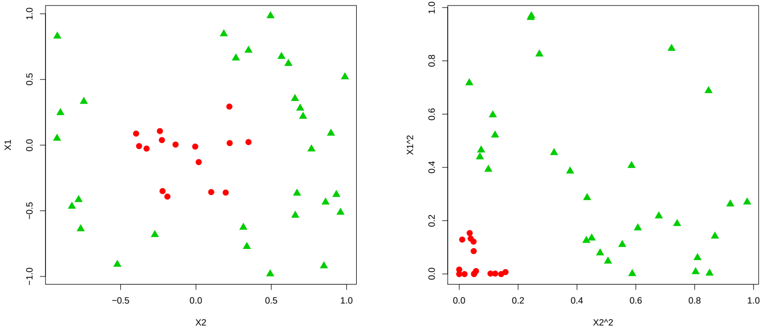
- It's crucial to the performance of the SVM.
- Il est le plus souvent choisi de faon "classique":
  1. We consider a performance criterion (e. g. misclassification rate).
  2. We estimate the value of the criterion for different values of C.
  3. We choose the value of C whoc minimizes the criterion.

Table of contents

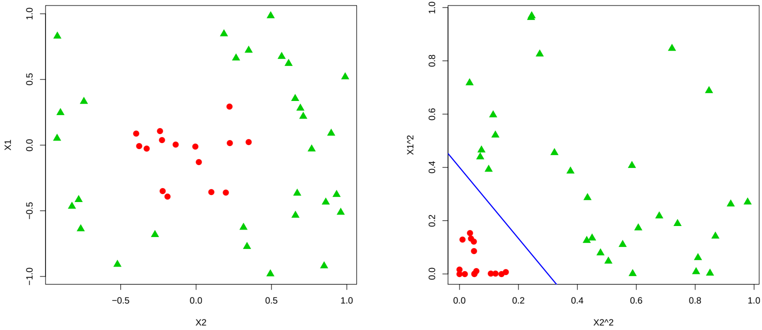
- Linear solutions are not always the best...



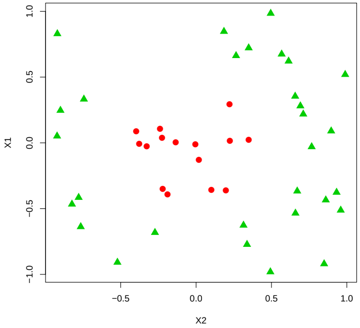
- Linear solutions are not always the best. . .

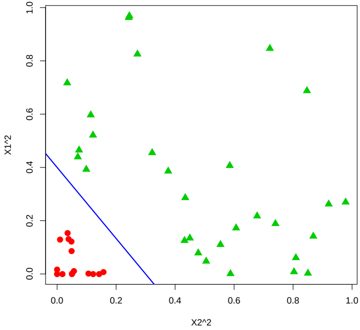


- Linear solutions are not always the best...



- Linear solutions are not always the best...





Idea

Find a data transformation such that the transformed data are linearly separable.

Kernel

Let $\Phi : \mathcal{X} \to \mathcal{H}$ be an application from $\mathcal{X}$ to a Hilbert space$\mathcal{H}$.
The kernel K at values x and x', associated to Φ, is the inner product of $\Phi(x)$ and $\Phi(x')$:

$$\begin{aligned} K : \mathcal{X} \times \mathcal{X} &\to \mathbb{R} \\ (x, x') &\mapsto \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}. \end{aligned}$$

Kernel

Let $\Phi : \mathcal{X} \to \mathcal{H}$ be an application from $\mathcal{X}$ to a Hilbert space$\mathcal{H}$.
The kernel K at values x and x', associated to Φ, is the inner product of $\Phi(x)$ and $\Phi(x')$:

$$\begin{aligned} K : \mathcal{X} \times \mathcal{X} &\to \mathbb{R} \\ (x, x') &\mapsto \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}. \end{aligned}$$

Example

If $\mathcal{X} = \mathcal{H} = \mathbb{R}^2$ and $\Phi(x_1, x_2) = (x_1^2, x_2^2)$ then:

$$K(x, x') = (x_1)^2 (x_1')^2 + (x_2)^2 (x_2')^2 \ .$$

Kernel trick

- The trick consists in transforming data x_i in a Hilbert space $\mathcal{H}$ called feature space...

Kernel trick

- The trick consists in transforming data x_i in a Hilbert space $\mathcal{H}$ called feature space...
- ...expecting that $(\Phi(x_1), y_1), \ldots, (\Phi(x_n), y_n)$ are (almost) linearly separables in order to apply SVM on transformed data.

Kernel trick

- The trick consists in transforming data x_i in a Hilbert space $\mathcal{H}$ called feature space...
- ...expecting that $(\Phi(x_1), y_1), \ldots, (\Phi(x_n), y_n)$ are (almost) linearly separables in order to apply SVM on transformed data.

1. A lot of linear algorithms (in particular SVM) can be applied to $\Phi(x)$ without ever computing Φ. All we have to do is to compute $K(x, x')$.

Kernel trick

- The trick consists in transforming data x_i in a Hilbert space $\mathcal{H}$ called feature space...
- ...expecting that $(\Phi(x_1), y_1), \ldots, (\Phi(x_n), y_n)$ are (almost) linearly separables in order to apply SVM on transformed data.

1. A lot of linear algorithms (in particular SVM) can be applied to $\Phi(x)$ without ever computing Φ. All we have to do is to compute $K(x, x')$.
2. We don't need to know the space $\mathcal{H}$ or the function Φ, we just need to consider a kernel$K$!

SVM in the inital space

- The dual problem maximises:

$$L_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{n} \alpha_i \alpha_k y_i y_k \langle x_i, x_k \rangle$$

under the constraints:

$$\begin{cases} \forall i \in \{1, \dots, n\} : 0 \leq \alpha_i \leq C \\ \sum_{i=1}^{n} \alpha_i y_i = 0. \end{cases} .$$

- The decision rule is obtained by calculating the sign of:

$$f(x) = \sum_{i=1}^{n} \alpha_i^\star y_i \langle x_i, x \rangle + b^\star \ .$$

SVM in the feature space

- The dual problem maximises:

$$L_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{n} \alpha_i \alpha_k y_i y_k \langle \Phi(x_i), \Phi(x_k) \rangle$$

under the consraints:

$$\begin{cases} \forall i \in \{1, \dots, n\} : 0 \leq \alpha_i \leq C \\ \sum_{i=1}^{n} \alpha_i y_i = 0. \end{cases} .$$

- The decision rule is obtained by calculating the sign of:

$$f(x) = \sum_{i=1}^{n} \alpha_i^\star y_i \langle \Phi(x_i), \Phi(x) \rangle + b^\star \ .$$

SVM in the feature space with kernels

- The dual problem maximises:

$$L_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{n} \alpha_i \alpha_k y_i y_k K(x_i, x_k)$$

under the consraints:

$$\begin{cases} \forall i \in \{1, \ldots, n\} : 0 \leq \alpha_i \leq C \\ \sum_{i=1}^{n} \alpha_i y_i = 0. \end{cases} .$$

- The decision rule is obtained by calculating the sign of:

$$f(x) = \sum_{i=1}^{n} \alpha_i^\star y_i K(x_i, x) + b^\star \ .$$

Conclusion

- To compute SVM, we don't need to know $\mathcal{H}$ or Φ, we just need to know K!

Conclusion

- To compute SVM, we don't need to know $\mathcal{H}$ or Φ, we just need to know K!
- Questions: What's a kernel ? How to build a kernel ?

Conclusion

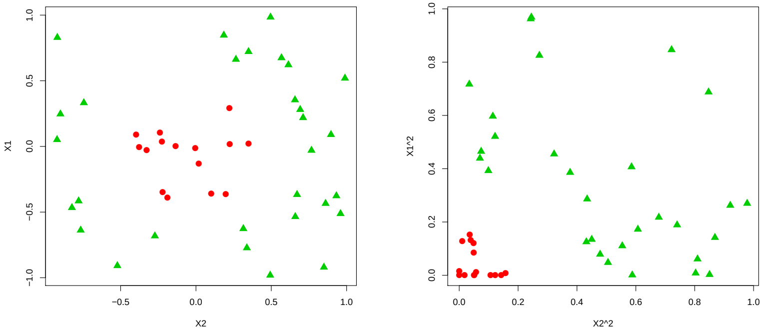
- To compute SVM, we don't need to know $\mathcal{H}$ or Φ, we just need to know K!
- Questions: What's a kernel ? How to build a kernel ?

$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel if and only if it's a symetric and positive definite function :

1. $\forall (x, x') \in \mathcal{X}^2 : K(x, x') = K(x', x)$.
2. $\forall (x_1, \ldots, x_N) \in \mathcal{X}^N$ et $\forall (a_1, \ldots, a_N) \in \mathbb{R}^N$:

$$\sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j K(x_i, x_j) \geq 0 .$$

Example



If:

$$\begin{aligned}\Phi: \quad \mathbb{R}^2 &\to \mathbb{R}^3 \\ (x_1, x_2) &\mapsto (x_1^2, \sqrt{2}x_1x_2, x_2^2)\end{aligned}$$

then $K(x, x') = (x^\top x')^2$.

Kernel examples

1. Linear (on $\mathbb{R}^d$): $K(x, x') = x^\top x'$.
2. Polynomial (on $\mathbb{R}^d$): $K(x, x') = (x^\top x' + 1)^d$.
3. Gaussian (Gaussian radial basis function ou RBF) (on $\mathbb{R}^d$):
$$K(x, x') = \exp\left(-\frac{\|x - x'\|}{2\sigma^2}\right) .$$
4. Laplace (sur $\mathbb{R}$): $K(x, x') = \exp(-\gamma |x - x'|)$.
5. Min (sur $\mathbb{R}^+$): $K(x, x') = \min(x, x')$.

Kernel examples

1. Linear (on $\mathbb{R}^d$): $K(x, x') = x^\top x'$.
2. Polynomial (on $\mathbb{R}^d$): $K(x, x') = (x^\top x' + 1)^d$.
3. Gaussian (Gaussian radial basis function ou RBF) (on $\mathbb{R}^d$):
$$K(x, x') = \exp\left(-\frac{\|x - x'\|}{2\sigma^2}\right) .$$
4. Laplace (sur $\mathbb{R}$): $K(x, x') = \exp(-\gamma\,|x - x'|)$.
5. Min (sur $\mathbb{R}^+$): $K(x, x') = \min(x, x')$.

Any positive definite function works... It's possible to build kernels (and so ti apply SVM) on more complex objects (curves, images, texts...).

References

Vapnik, V. N. (1998). *Statistical learning theory*. Adaptive and learning systems for signal processing, communications, and control. Wiley.