Machine Learning in Finance

Pierre Brugière

Machine Learning in Finance

Pierre Brugiere

University Paris 9 Dauphine

pierre.brugiere@dauphine.fr

November 1, 2016

Overview

Calibration versus Prediction

Background

We distinguish several type of statistical problems :

- Regression problems where Y and X are quantitative variables and where Y is inferred by a function $f(X)$
- Classification problems where Y is a qualitative variable and where the class of Y is inferred from X
- Clusterization problems where a quantitative variable X is observed and classified into groups of similar features.

Remarks: Often a qualitative variable will be "coded" for modelisation purposes into a quantitative variable but usually without any implicit order relationship or proximity notion between the values coded, and this contrarily to what would happen for "native" quantitative variables.

Calibration versus Prediction

We will focus mainly on classification problems where:

- Y is a binary variable and X is a quantitative variable in $\mathbb{R}^d$.
- $(X^1, Y^1), (X^2, Y^2), \cdots (X^n, Y^n)$ are observations

The issue is to choose:

- a particular class of models $\mathcal{F} \in \{\mathcal{F}_\alpha\}$
- a function f within $\mathcal{F}$ to estimate Y by $f(X)$

We define a measure of error between Y and $f(X)$ as :

- $\|Y - f(X)\|$ for a regression problem
- $1_{\{Y \neq f(X)\}}$ for a classification problem

Mathematically in a classification problem the goal is to find f which minimizes the risk $E[1_{f(X) \neq Y}]$

Calibration versus Prediction - Risk Measure

Definition

For any f in $\mathcal{F}$ we note:

- $R(f) := E[1_{f(X) \neq Y}]$
- $R_n(f) := \frac{1}{n} \sum_{i=1}^{i=n} 1_{\{f(X_i) \neq Y_i\}}$

Calibration associates to a sample $(X_i, Y_i)_{i \in [\![1,n]\!]}$ an element f_n of $\mathcal{F}$.
In this case f_n is a random variable taking its value in $\mathcal{F}$ and we note:

- $R(f_n) := E[1_{f_n(X_{i+1}) \neq Y_{i+1}}]$
- $R_n(f_n) := \frac{1}{n} \sum_{i=1}^{i=n} 1_{\{f_n(Xn+1) \neq Y_{n+1}\}}$

Remarks: As f_n is chosen in $\mathcal{F}$ to satisfy $R_n(f_n) = \min_{f \in \mathcal{F}} R_n(f)$ then in most models we will have $E[R_n(f_n)] < R(f_n)$

Calibration versus Prediction - Risk Measure

Example: Let (X, Y) be random variables with $X \sim \mathcal{U}([0,1])$ and $Y = 1_{X \leq a}$ with $a \in]0,1[$.
We assume that we do not know the existing relationship between X and Y but want to build a classifier based on some sampling $(X_i, Y_i)_{i \in [\![1,n]\!]}$ and a machine $\mathcal{F} = \{1_{x \leq \alpha}, 1_{x \geq \alpha}\}_{\alpha \in \mathbb{R}}$.
If we assume that when observing $(X_i, Y_i)_{i \in [\![1,n]\!]}$ we choose the classifier $1_{x \leq Max(X_1 1_{\{X_1 < a\}}, X_2 1_{\{X_2 < a\}}, .. X_n 1_{\{X_n < a\}}}$ of $\mathcal{F}$ then show that:

- $R_n(f_n) = 0$
- $R(f_n) = \frac{1-(1-a)^{n+1}}{n+1}$

Hint: $R(f_n) = \int_0^a P\left(\max_{i \in [\![1,n]\!]} X_i 1_{\{X_i < a\}} < u \right) du$

Calibration versus Prediction

Our goal is:

- not so much to explain perfectly what has happened (calibration) but
- to be as precise as possible in the prediction

So we face a dilemma as:

- a model which has too many parameters may enable perfect calibration but lead to over-fitting and a poor quality of prediction
- a too simplistic model which fits only very poorly the sample data has no chance to predict accurately

The Vapnik Chernovenkis theorem enables to control $R(f_n)$ based on:

- $R_n(f_n)$
- the complexity, noted $VC(\mathcal{F})$, of the model $\mathcal{F}$

Calibration versus Prediction - VC dimension

Remarks: in Machine Learning

- each class of estimator $\mathcal{F}_\alpha$ is called a machine
- the phase of calibration is called the learning phase
- if the Y_i are known in the sample and thus an error of calibration can be calculated, the learning is said to be supervised

Definition: VC dimension of $\mathbb{R}^d$ classifiers

Let $\mathcal{F} = \{f\}_{\alpha \in \mathcal{E}}$ be a family of classificators, each f_α being a function from $\mathbb{R}^d$ to $\{0, 1\}$.
The Vapnik Chervonenkis dimension of $\mathcal{F}$ noted $VC(\mathcal{F})$ is the maximum number of points of $\mathbb{R}^d$ that can be classified in all possible different ways by some classificators of $\mathcal{F}$.

Calibration versus Prediction - VC dimension

Remarks: $VC(\mathcal{F}) = k$ if and only if it is possible to find k points $(x_i)_{i\in\{1,k\}}$ in $\mathbb{R}^d$ such that for any of the 2^k possible labelling $(y_i)_{i\in\{1,k\}}$ in $\{0,1\}^k$ it is possible to find f in $\mathcal{F}$ such that $\forall i \in [\![1,k]\!], f(x_i) = y_i$.

VC Theorem (admitted): Confidence interval for the risk of prediction

We note $(X_i, Y_i)_{i\in[\![1,n]\!]}$ a i.i.d sample of (X, Y)
Let $\mathcal{F}_d = \{f\}_{\alpha\in\mathcal{E}}$ be a machine with $VC(\mathcal{F}_d) < n$
Let $R_n(f_n) = \min_{f\in\mathcal{F}_d} R_n(f_n)$ for $(X_i, Y_i)_{i\in[\![1,n]\!]}$, then

$$\forall\eta \in [0,1],\ P\left(R(f_n) > R_n(f_n) + \phi_{n,\eta}\left(\frac{VC(\mathcal{F}_d)}{n}\right)\right) \leq \eta$$

where $\phi_{n,\eta}(x) = \sqrt{x(1 + ln(\frac{2}{x})) + \frac{1}{n}ln(\frac{4}{\eta})}$ and $x = \frac{VC(\mathcal{F}_d)}{n}$ so

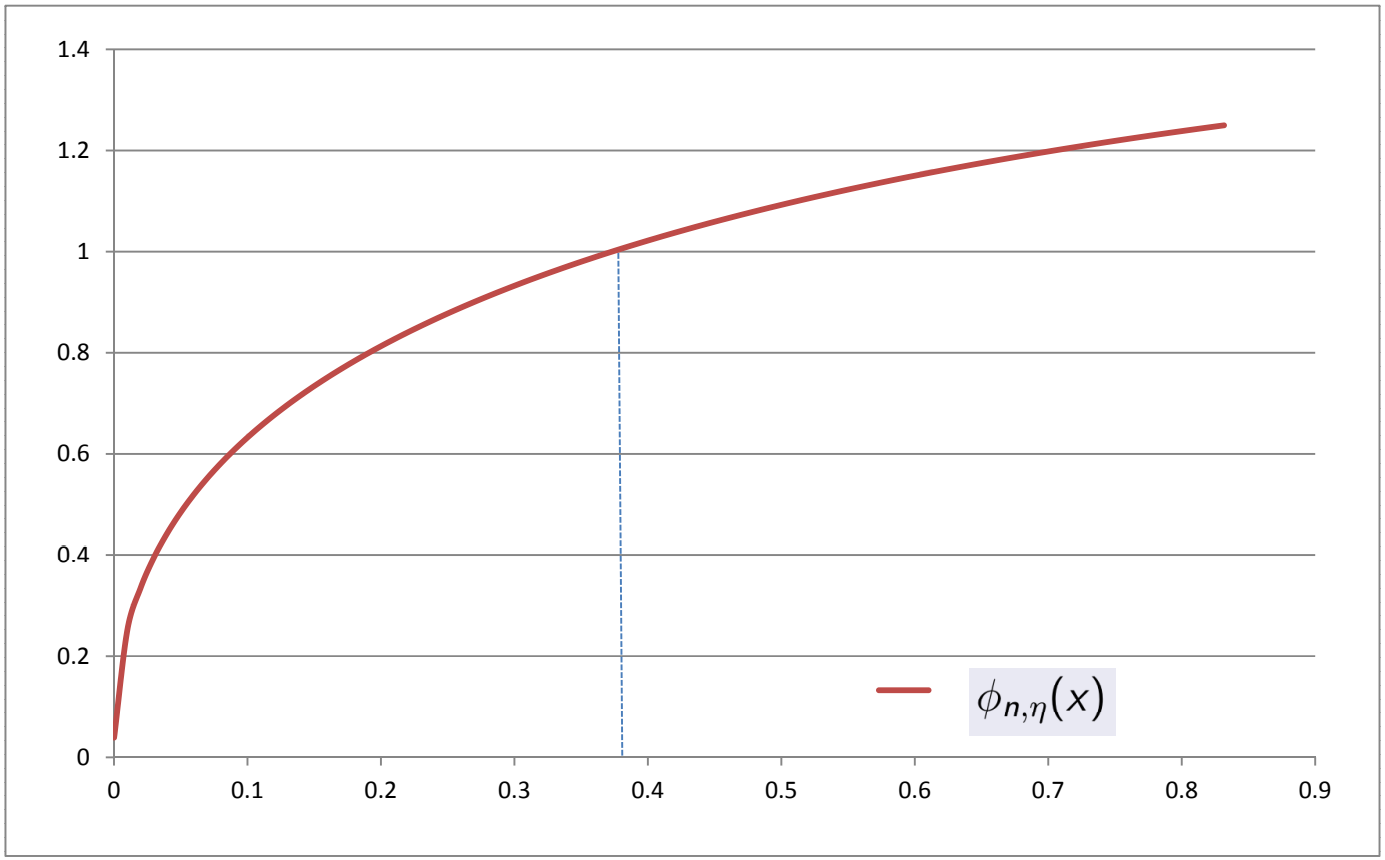
$\left[0, R_n(f_n) + \phi_{n,\eta}(\frac{VC(\mathcal{F}_d)}{n})\right]$ is an interval at confidence level $1-\eta$ for $R(f_n)$

Calibration versus Prediction - VC dimension

Example 1: If we assume $VC(\mathcal{F}_d) = 20$, $n = 10,000$
then with $\eta = 1\%$ we obtain $P(R(f_n) > R_n(f_n) + 12.81\%) \leq 1\%$

Example 2: In the previous example of classification with $\mathcal{F} = \{1_{x<\alpha}\}_{\alpha \in \mathbb{R}}$ it is easy to check that $VC(\mathcal{F}) = 2$.
With 10,000 observations the VC-theorem then guarantees that at 95% confidence level $R(f_n)$ (for estimators with minimum empirical risks) should be within the interval $[0, 4.98\%]$ (as $R_n(f_n) = 0$ and $\phi_{10,000,5\%}(\frac{2}{10,000}) = 4.98\%$).
We note that the estimation of the confidence interval for this particular problem is quite loose because as seen previously
$R(f_n) = \frac{1-(1-a)^{n+1}}{n+1} \leq \frac{1}{n+1} = 0.01\%$.

Calibration versus Prediction - VC dimension



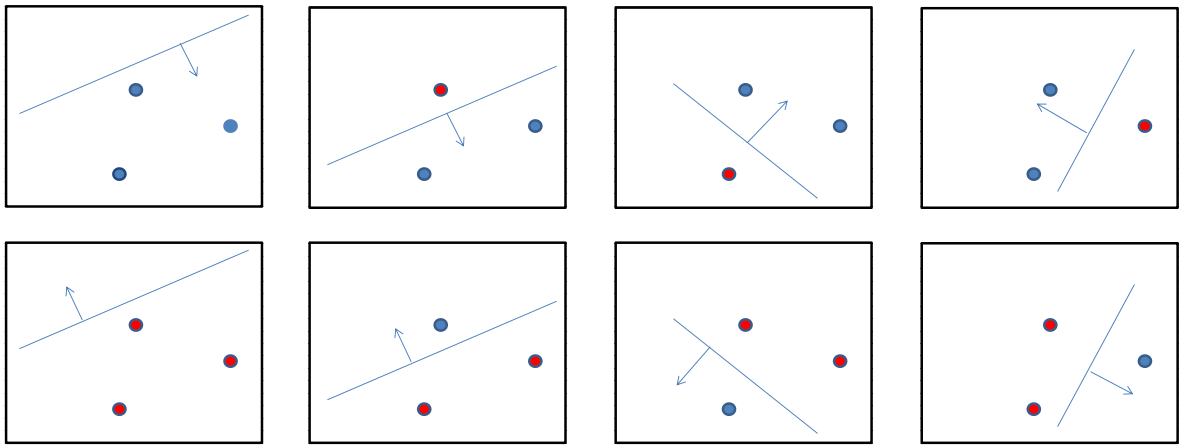
Calibration versus Prediction - VC dimension

Exemple: We consider the following machine (of $\{0,1\}$-classifiers) in $\mathbb{R}^2$:
$\mathcal{F} = \{1_{ax+by+c\geq 0}, 1_{ax+by+c\leq 0}, (a,b) \in \mathbb{R}^2 \setminus \{0\}, c \in \mathbb{R}\}$.
Each classifier, classifies points in $\mathbb{R}^2$ according to their positions relatively to the line $ax + by + c = 0$.
We notice that:

- we can find 3 points in $\mathbb{R}^2$ that can be $\{0,1\}$-classified in all possible ways with classifiers from $\mathcal{F}$
- it seems impossible to find 4 points in $\mathbb{R}^2$ that can be $\{0,1\}$-classified in all possible ways

If the later assumption is true, it will prove that $VC(\mathcal{F}) = 3$.
We are going to prove this result as a particular case of a more general result.

Calibration versus Prediction - VC dimension



Three points from $\mathbb{R}^2$ being $\{0,1\}$-classified in all possible ways by the machine $\mathcal{F}$ (blue=1,red=0)

Calibration versus Prediction - VC dimension

Theorem : VC dimension of oriented hyperplanes of $\mathbb{R}^d$

Let $x_1, x_2, \cdots, x_n$ be n points of $\mathbb{R}^d$
Let $\mathcal{F}_d = \{1_{\{<w,x>+c\geq 0\}}, w \in \mathbb{R}^d \setminus \{0\}, c \in \mathbb{R}\}$ be the family of $\{0,1\}$-classifiers defined by the oriented hyperplanes of $\mathbb{R}^d$.
Then, $x_1, x_2, \cdots, x_n$ can be $\{0,1\}$-classified in all possible ways by $\mathcal{F}_d$ if and only if $x_2 - x_1, x_3 - x_1, \cdots, x_n - x_1$ are linearly independent.

Corollary

$VC(\mathcal{F}_d) = d + 1$

Remarks: From the corollary we can say that for an "affine classifier" in $\mathbb{R}^d$ the VC dimension is simply the number of parameters.

Calibration versus Prediction - VC dimension

Demonstration theorem:

let $(y_i)_{i\in[\![1,n]\!]}$ be a $\{0,1\}$-classification of the $(x_i)_{i\in[\![1,n]\!]}$

let I_1 be the indices of the x_i with the same classification as x_1

let I_2 be the indices of the x_i with a different classification from x_1

we want to prove that we can separate the $\{x_i\}_{i\in I_1}$ and the $\{x_i\}_{i\in I_2}$

Let $\mathcal{C}_1$ (resp $\mathcal{C}_2$) be the convex envelope of the $\{x_i\}_{i\in I_1}$ (resp $\{x_i\}_{i\in I_2}$)

Let's start proving that $\mathcal{C}_1 \cap \mathcal{C}_2 = \varnothing$

If this was not the case we could find $(\lambda_i)_{i\in I_1}$ $(\lambda_j)_{j\in I_2}$ such that:

$\forall i \in I_1\ \lambda_i \geq 0, \forall j \in I_2\ \lambda_j \geq 0, \sum_{i\in I_1} \lambda_i = 1, \sum_{j\in I_2} \lambda_j = 1$

and $\sum_{i\in I_1} \lambda_i x_i = \sum_{j\in I_2} \lambda_j x_j$ (1)

by substracting x_1 from both terms of (1) we would have :

$\sum_{i\in I_1\setminus\{1\}} \lambda_i(x_i - x_1) = \sum_{j\in I_2} \lambda_j(x_j - x_1)$ which would be in contradiction with the assumption of independence in the theorem

Calibration versus Prediction - VC dimension

So necessarily $\mathcal{C}_1 \cap \mathcal{C}_2 = \varnothing$.
By compacity we deduct that we can find $z_1 \in \mathcal{C}_1$ and $z_2 \in \mathcal{C}_2$ such that $|z_1 - z_2| = \textit{distance}(\mathcal{C}_1, \mathcal{C}_2) > 0$. If now we consider the hyperplane orthogonal to $z_2 - z_1$ and containing $\frac{z_1+z_2}{2}$ it is easy to check that:

- this hyperplane separates $\mathcal{C}_1$ and $\mathcal{C}_2$ and has for equation $\langle x, z_2 - z_1 \rangle = \langle \frac{z_1+z_2}{2}, z_2 - z_1 \rangle$
- the points of $\mathcal{C}_1$ satisfy $\langle x, z_2 - z_1 \rangle \leq \langle z_1, z_2 - z_1 \rangle$
- the points of $\mathcal{C}_1$ satisfy $\langle x, z_2 - z_1 \rangle \geq \langle z_2, z_2 - z_1 \rangle$

So $\mathcal{C}_1$ and $\mathcal{C}_2$ are separated by an hyperplane and so the $(x_i)_{i \in I_1}$ $(x_i)_{i \in I_2}$.
So the independence condition shows that the points can be classified in all possible ways.

Calibration versus Prediction - VC dimension

Let's prove now that:
(the points can be classified in all possible ways) $\Rightarrow$
($x_2 - x_1, x_3 - x_1, \cdots, x_n - x_1$ are linearly independent).
For this we show the transpose.
If we assume that $x_2 - x_1, x_3 - x_1, \cdots, x_n - x_1$ are linearly dependent then we can find $(\lambda_i)_{i \in [\![2,n]\!]} \in \mathbb{R}^{n-1} \setminus \{0\}$ such that $\sum_{i=2}^{i=n} \lambda_i (x_i - x_1) = 0$ (2)
we then note:
$I = \{i \in [\![2,n]\!], \lambda_i \geq 0\}$ $J = \{i \in [\![2,n]\!], \lambda_i > 0\}$
$\lambda_i = \lambda_i^+ if \lambda_i \geq 0$ and $\lambda_i = -\lambda_i^- if \lambda_i < 0$ and we can rewrite (2) as
$\sum_{i \in I} \lambda_i^+ (x_i - x_1) - \sum_{j \in J} \lambda_j^- (x_j - x_1) = 0$ (3)
a) We assume in a first case that the λ_i are not all of the same sign and without loss of generality that $\sum_{i \in I} \lambda_i^+ \geq \sum_{j \in J} \lambda_j^-$

Calibration versus Prediction - VC dimension

If the $(x_i)_{i\in[\![2,n]\!]}$ can be separated with $\mathcal{F}_d$ we can find w and c such that: $\forall i \in I$, $\langle w, x_i\rangle \geq c$ and $\forall j \in J$, $\langle w, x_j\rangle < c$ but from (3):

$$\sum_{i\in I}\lambda_i^+\langle w, x_i\rangle - \sum_{j\in J}\lambda_j^-\langle w, x_j\rangle = (\sum_{i\in I}\lambda_i^+ - \sum_{j\in J}\lambda_j^-)\langle w, x_1\rangle \quad (4)$$

implies that x_1 cannot be separated from the $(x_i)_{i\in I}$ as

$$\sum_{i\in I}\lambda_i^+\langle w, x_i\rangle - \sum_{j\in J}\lambda_j^-\langle w, x_j\rangle \geq (\sum_{i\in I}\lambda_i^+ - \sum_{j\in J}\lambda_j^-)c$$

implies from (4) that $\langle w, x_1\rangle \geq c$ as well. Q.E.D

b) If we assume now that the λ_i are all of the same sign and without loss of generality that this sign is positive then (2) can be rewritten as

$\sum_{i=1}^{i=n}\lambda_i x_i = (\sum_{i=1}^{i=n}\lambda_i)x_1$ (5) which proves that no classifier in $\mathcal{F}_d$ can separate the $(x_i)_{i\in[\![2,n]\!]}$ from x_1 as:

$\forall i \in [\![2, n]\!], \langle w, x\rangle \geq c \Rightarrow \sum_{i=1}^{i=n}\lambda_i\langle x, x_i\rangle \geq (\sum_{i=1}^{i=n}\lambda_i)c$ and from (5) this implies $\langle w, x_1\rangle \geq c$ as well. Q.E.D

Calibration versus Prediction - VC dimension

Demonstration corollary:
In $\mathbb{R}^d$ if we take d vectors $x_1, x_2, \cdots x_d$ independent then according to the theorem, the vectors: $0, x_1, x_2, \cdots x_d$ can be classified in all possible ways by $\mathcal{F}_d$. This proves that $VC(\mathcal{F}_d) \geq d + 1$.
Conversely we know that if $x_1, x_2, \cdots x_n$ can be classified in all possible different ways by $\mathcal{F}_d$ then the $n-1$ vectors $x_d - x_1$ must be independent and therefore $n - 1 \leq d$ and $VC(\mathcal{F}_d) - 1 \leq d$.
Consequently $VC(\mathcal{F}_d) = d + 1$. Q.E.D

Calibration versus Prediction - VC dimension

Remarks: Even if with hyperplane classifiers $VC(\mathcal{F}_d)$ is the number of parameters of the hyperplanes, in general the VC dimension is something different from the number of parameters of the model.

Exercice:

We consider on $\mathbb{R}$ the machine $\mathcal{F} = \{1_{sin(\alpha x)>0}, \alpha \in \mathbb{R}\}$

We consider $(x_i)_{i\in[\![1,l]\!]}$ defined by $x_i = 10^{-i}$.

Show that for any $\{0,1\}$-classification $(y_i)_{i\in[\![1,l]\!]}$ of the $(x_i)_{i\in[\![1,l]\!]}$ the classifier $1_{sin(\alpha x)>0}$ with $\alpha = \pi\left(1 + \sum_{i=1}^{i=l}(1-y_i)10^i\right)$ classifies perfectly all the points. Conclude that $VC(\mathcal{F}) = +\infty$

Demonstration:

For any indice $1 < j < l$ we have:

$$\alpha x_j = \pi\left(1 + \sum_{i=1}^{i=j-1}(1-y_i)10^i\right)10^{-j} + \pi(1-y_j) + \pi\sum_{i=j+1}^{i=l}(1-y_i)10^{i-j}$$

Calibration versus Prediction - VC dimension

We notice that the last term is a multiple of 2π and thus can be noted $2k\pi$ and that the first term is always between 0 and π and thus can be noted $\beta\pi$ with $0 < \beta < 1$ so:
if $y_j = 1$, $sin(\alpha x) = sin(\beta\pi + 0 + 2k\pi) = sin(\beta\pi) > 0$
if $y_j = 0$, $sin(\alpha x) = sin(\beta\pi + \pi + 2k\pi) = sin(\beta\pi + \pi) < 0$
so $1_{sin(\alpha x)>0}$ classifies x_j correctly.
We can prove the same for x_1 and x_l which proves that whatever the labels are for the $(x_i)_{i\in[\![1,l]\!]}$ we can classify them correctly.
Now $\forall l$, $VC(\mathcal{F}) \geq l \Rightarrow VC(\mathcal{F}) = +\infty$. Q.E.D

Remarks: In the exercise above the classifiers depends only on one parameter but the VC dimension of the machine is infinite. So the complexity of a model, as measured by its VC dimension, and the number of parameters can be quite different in the non-linear case.

Maximum Margin Classifiers

Maximum Margin Classifiers

Definition

Let $(x_i, y_i)_{i \in [\![1,n]\!]}$ be a sample of (X, Y) with $x_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$
Let $H_{w,b} = \{x \in \mathbb{R}^d, \langle w, x \rangle + b = 0\}$
We say that $H_{w,b}$ separates totally the $(x_i, y_i)_{i \in [\![1,n]\!]}$ iff
for one class of points $\langle w, x \rangle + b \geq 0$ while for the other class
$\langle w, x \rangle + b < 0$.

Proposition

Let $H_{w,b}$ be an hyperplane of $\mathbb{R}^d$ then for any $x \in \mathbb{R}^d$,
$d(x, H_{w,b}) = \frac{|\langle wx \rangle + b|}{\|w\|}$

Notation:
We note $\mathcal{X}_0 = \{x_i, i \in [\![1, n]\!] \text{ such that } y_i = 0\}$,
$\mathcal{X}_1 = \{x_i, i \in [\![1, n]\!] \text{ such that } y_i = 1\}$ and $\mathcal{S} = \{(x_i, y_i)\}_{i \in [\![1,n]\!]}$

Maximum Margin Classifiers

Demonstration:
Let $y = p_{H_{w,b}}(x)$ be the othogonal projection of x onto $H_{w,b}$ then,
$\exists \lambda \in R$, $y - x = \lambda w$ and $d(x, H_{w,b}) = |\lambda| \|w\|$
but, $y - x = \lambda w \Rightarrow \langle w, y - x \rangle = \lambda \|w\|^2 \Rightarrow -b - \langle w, x \rangle = \lambda \|w\|^2$
this implies $\lambda = \frac{-b - \langle w,x \rangle}{\|w\|^2}$ and $|\lambda| = \frac{|b + \langle w,x \rangle|}{\|w\|}$ Q.E.D

Exercise: Show that

- $d(H_{w,b_1}, H_{w,b_2}) = \frac{|b_2 - b_1|}{\|w\|}$
- $H_{w,b} = H_{-w,-b}$
- $d(H_{w,b_1}, H_{-w,-b_2}) = \frac{|b_2 - b_1|}{\|w\|}$

Maximum Margin Classifiers

Definition: Margin, Maximum Margin Hyperplane

Let $(x_i, y_i)_{i\in[\![1,n]\!]}$ be a sample of (X, Y) with $x_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$.
if $H_{w,b}$ separates totally the $(x_i, y_i)_{i\in[\![1,n]\!]}$.

- We call margin of $H_{w,b}$ and note $\Delta(H_{w,b})$ the quantity :
$$\begin{cases} \max\limits_{c_1,c_2} d(H_{w,c_1}, H_{-w,-c_2}) \\ H_{w,c_1}, H_{-w,-c_2} \text{ separates totally the } (x_i, y_i)_{i\in[\![1,n]\!]} \end{cases}$$
- We say that $H_{w,b}$ has maximum margin iff any other hyperplane H separating totally the $(x_i, y_i)_{i\in[\![1,n]\!]}$, verifies $\Delta(H) \leq \Delta(H_{w,b})$

Exercise: Show that if the $(x_i, y_i)_{i\in[\![1,n]\!]}$ are a sample of (X, Y) separable by an hyperplane, then the margin of the maximum margin hyperplane is $d(\mathcal{C}_0, \mathcal{C}_1)$ where the $\mathcal{C}_0$ and $\mathcal{C}_1$ are the convex envelopes of the two classes.

Maximum Margin Classifiers

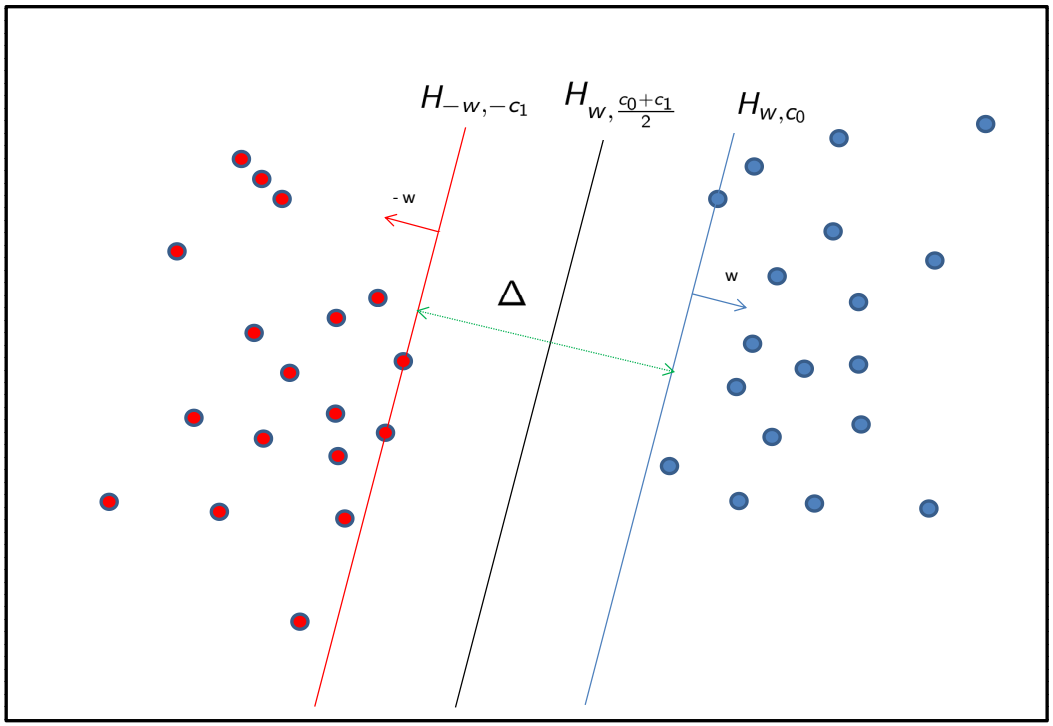
Exercise: Let $(x_i, y_i)_{i \in [\![1,n]\!]}$ be a sample of (X, Y). Let $H_{w,c}$ be an hyperplane which separates the convex envelopes $\mathcal{C}_0$ and $\mathcal{C}_1$.
a) show that $\exists c_0$ and c_1,

- $\forall x \in \mathcal{C}_0$, $\langle w, x \rangle + c_0 \geq 0$
- $\forall x \in \mathcal{C}_1$, $\langle -w, x \rangle - c_1 \geq 0$ and
- $\Delta(H_{w,c}) = \frac{|c_1 - c_0|}{\|w\|}$

b) show that,

- $d(H_{w, \frac{c_0+c_1}{2}}, H_{w,c_0}) = \frac{|\frac{c_0+c_1}{2} - c_0|}{\|w\|} = \frac{|\frac{c_1-c_0}{2}|}{\|w\|}$
- $d(H_{w, \frac{c_0+c_1}{2}}, H_{-w,-c_1}) = \frac{|\frac{c_0+c_1}{2} - c_1|}{\|w\|} = \frac{|\frac{c_0-c_1}{2}|}{\|w\|}$
- $\forall x \in \mathcal{C}_0$, $\langle w, x \rangle + \frac{c_0+c_1}{2} \geq \frac{\Delta(H_{w,c})}{2}\|w\|$ (5)
- $\forall x \in \mathcal{C}_1$, $\langle w, x \rangle + \frac{c_0+c_1}{2} \leq -\frac{\Delta(H_{w,c})}{2}\|w\|$ (6)

Maximum Margin Classifiers



Maximum Margin Hyperplane H_w

Maximum Margin Classifiers

Remarks:

- $H_{w,\frac{c_0+c_1}{2}}$ lies at equal distance from the two hyperplanes, orthogonal to w, separating, with maximum distance between them, $\mathcal{C}_0$ and $\mathcal{C}_1$. We note this hyperplane H_w
- If we define $\omega = \frac{w}{\|w\|}\frac{2}{\Delta}$ and $b = \frac{c_0+c_1}{\Delta\|w\|}$
  we can write (5) and (6) in the standard form:
  $\forall x \in \mathcal{C}_0,\ \langle \omega, x\rangle + b \geq 1$ (5)
  $\forall x \in \mathcal{C}_1,\ \langle \omega, x\rangle + b \leq -1$ (6)
  The three (parallel) hyperplanes defined previously can now be noted $H_{\omega,b-1}$, $H_{-\omega,-b-1}$ and $H_{\omega,b}$ and $d(H_{-\omega,-b-1}, H_{\omega,b-1}) = \frac{2}{\|\omega\|}$
- Therefore, in practice to search for an hyperplane with maximum margin search for ω and b which solve:

$$(P)\begin{cases} \max\limits_{\omega,b} \frac{2}{\|\omega\|} \\ \forall x_i \in \mathcal{X}_0,\ \langle \omega, x\rangle + b \geq 1 \\ \forall x_i \in \mathcal{X}_1,\ \langle \omega, x\rangle + b \leq -1 \end{cases}$$

Maximum Margin Classifiers

Remarks: ω and b also solve:

$$(P)\begin{cases} \min\limits_{\omega,b} \|\omega\|^2 \\ \forall x_i \in \mathcal{X}_0,\ \langle \omega, x \rangle + b \geq 1 \\ \forall x_i \in \mathcal{X}_1,\ \langle \omega, x \rangle + b \leq -1 \end{cases}$$

which is a quadratic problem with affine constraints, which can be solved using the Karush-Kuhn-Tucker theorem.

Structural Risk Minimization and Gap Tolerant Classifiers

SRM and Gap Tolerant Classifiers

In the Structural Risk Minimization method:

- we define nested ensembles of classifiers (machines), $\mathcal{F}_1 \subset \mathcal{F}_2 \cdots \subset \mathcal{F}_k \cdots$, with $VC(\mathcal{F}_1) < VC(\mathcal{F}_2) < \cdots \leq VC(\mathcal{F}_k)$
- for the sample $(x_i, y_i)_{i \in [\![1,n]\!]}$, we calculate for each machine $\mathcal{F}_k$ the best classifier $f_{n,k}$ and its empirical risk $R_n(f_{n,k})$
- to control in the best possible way the error of prediction at confidence level 5%, we pick the estimator $f_{n,k}$ which minimizes $R_n(f_{n,k}) + \phi_{n,5\%}(\frac{VC(\mathcal{F}_k)}{n})$

SRM and Gap Tolerant Classifiers

Definition: Δ-Gap Tolerant Classifier of Diameter D

For $w \in R^d$, $b \in \mathbb{R}$ and B_D being a ball in $\mathbb{R}^d$ of diameter D we define $h_{w,b}^{B_D,\Delta}$ as:

$h_{w,b}^{B_D,\Delta}(x) = 1$ iff $x \in B_D$ and $\langle w, x \rangle + b \geq \frac{\Delta \|w\|}{2}$

$h_{w,b}^{B_D,\Delta}(x) = 0$ iff $x \in B_D$ and $\langle w, x \rangle + b \leq -\frac{\Delta \|w\|}{2}$

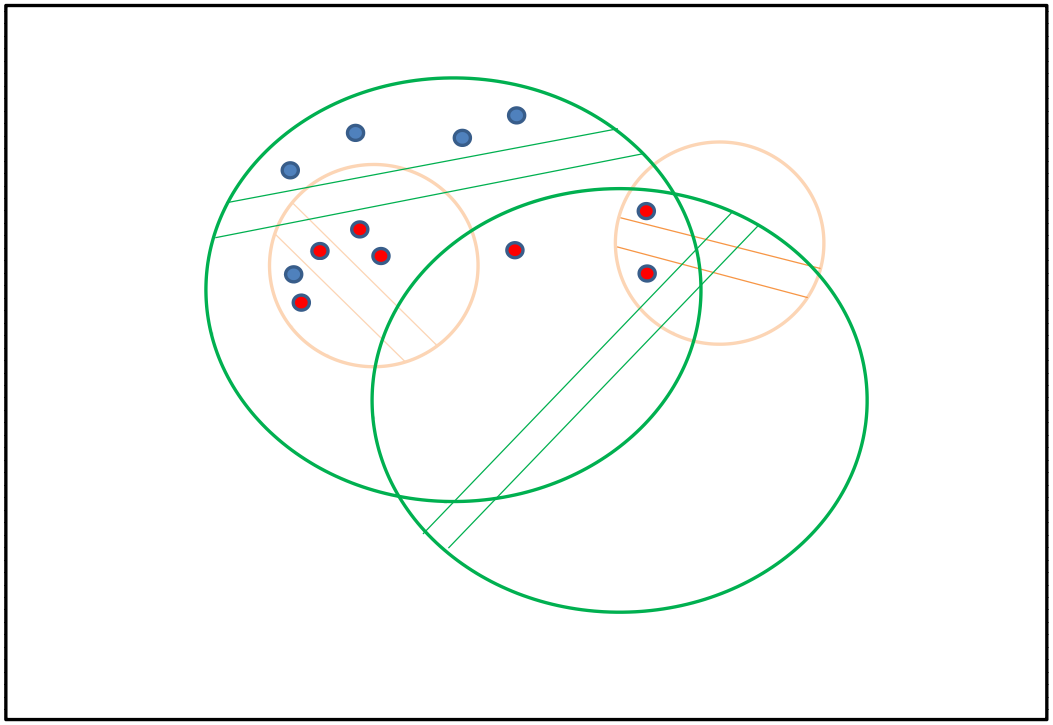
if $x \notin B_D$ or $|\langle w, x \rangle + b| < \frac{\Delta \|w\|}{2}$ then $h_{w,b}^{B_D,\Delta}$ is not defined

Such a classifier is called a Δ-Gap tolerant, $\{0,1\}$-classifier of diameter D

Remarks:

- For the Δ-Gap tolerant, $\{0,1\}$-classifier of diameter D the two hyperplanes $H_{w,b-\frac{\Delta \|w\|}{2}}$ and $H_{w,b+\frac{\Delta \|w\|}{2}}$ which separates two distinct classes of points are distant of Δ.
- We allow in the definition that the classifier may classify some points incorrectly

SRM and Gap Tolerant Classifiers



Only one Gap Tolerant Classifier classifies all the points here

SRM and Gap Tolerant Classifiers

Theorem admitted: VC of Δ-Gap Tolerant Classifier of Diameter D

Let $\mathcal{F}_{\Delta,D} = \{h_{w,b}^{B_D,\Delta}, w \in \mathbb{R}^d, b \in \mathbb{R}$ and B_D is a ball of diameter $D\}$
then $VC(\mathcal{F}_{\Delta,D}) \leq 1 + Min(\frac{D^2}{\Delta^2}, d)$

Remark:

- the notion of margin was introduced to classify as robustly as possible (i.e to minimize the risk of misclassification in case of a small errors in the measurements).
- using classifiers with a fixed margin may reduce significantly the VC dimension of the Machine when observing data in large dimension. For example if $d = 1,000,000$, $D = 1$, $\Delta = 0.1$, the VC dimension of hyperplane classifiers is $1,000,001$ while the same hyperplane classifiers with a margin of 0.1 and a diameter of 1 have a VC dimension of no more than 101.

SRM and Gap Tolerant Classifiers

Theorem admitted: Max Margin

The margin Δ at which a family of $k+1$ points within a ball of radius 1 can be classified in all possible ways by a family of Δ- Gap tolerant classifiers of radius 1 cannot be more than $\sqrt{\frac{k+1}{k}}\sqrt{\frac{1}{[\frac{k+1}{2}]}+\frac{1}{k+1-[\frac{k+1}{2}]}}$ where $[\frac{k+1}{2}]$ denotes the integer part of $\frac{k+1}{2}$. This maximum can be attained for some particular choices of families of $k+1$ points.

Remarks:

We know that is is possible to find $k+1$ points of $\mathbb{R}^k$ that can be classified in all possible ways by hyperplane classifiers. By renormalizing these points we can put them inside a ball of radius 1 and the hyperplanes renormalized will continue to classify them in all possible ways. This family of classifiers exhibits a certain margin and the theorem above gives us a limit in terms of the maximum we can expect. Later on we will show that the maximum margin is attained when the points form a simplex of the affine space $\mathbb{R}^k$ i.e can be seen as an othonormal family of vectors of $\mathbb{R}^{k+1}$.

SRM and Gap Tolerant Classifiers

The strategy to predict with Gap Tolerant Classifiers after observing a (learning) sample $(x_i, y_i)_{i \in [\![1,n]\!]}$ is as follows:

- define a ball B_D such that all the $x_i, i \in [\![1, n]\!]$ are inside B_D. Note $\mathcal{S} = \{x_i, i \in [\![1, n]\!]\}$
- find α such that the $\{x_i, i \in [\![1, n]\!]\}$ can be classified by some elements of $\mathcal{F}_{\alpha,D}$. Note $\mathcal{F}^{\mathcal{S}}_{\alpha,D}$ the elements of $\mathcal{F}_{\alpha,D}$ which classifies all the $\{x_i, i \in [\![1, n]\!]\}$ (correctly or incorrectly).
- build a nested set of machines $\mathcal{F}^{\mathcal{S}}_{\alpha_0,D} \subset \mathcal{F}^{\mathcal{S}}_{\alpha_1,D} \subset \cdots \subset \mathcal{F}^{\mathcal{S}}_{\alpha_n,D}$ with decreasing margins $\alpha = \alpha_0 > \alpha_1 > \cdots > \alpha_n$
- for each machine consider a gap tolerant classifier f_{n,α_n} with minimum empirical error $R(f_{n,\alpha_n})$
- using the fact that $VC(\mathcal{F}^{\mathcal{S}}_{\alpha_i,D}) \leq VC(\mathcal{F}_{\alpha_i,D}) \leq 1 + Min(\frac{D^2}{\alpha_i^2}, d)$ choose and estimator for which the error of calibration $R(f_{n,\alpha_n})$ and the complexity term, estimated by $1 + Min(\frac{D^2}{\alpha_i^2}, d)$ are providing the best control on the error of prediction.

Trade-off between Margin and Errors

Trade-off between Margin and Errors

Theorem and Definition:

The set $\{x \in \mathbb{R}^d, |\langle wx + b \rangle| \leq 1\}$ consists of R^d points between $H_{w,b-1}$ and $H_{-w,-b-1}$.
As $d(H_{w,b-1}, H_{-w,-b-1}) = \frac{2}{\|w\|}$ this ensemble is called hyperplan of thickness $\frac{2}{\|w\|}$ and is noted $H_{w,b}^{\frac{2}{\|w\|}}$.

When the sample points $\{(x_i, y_i)\}_{i \in [\![1,n]\!]}$ are separable (with $y_i \in \{-1, 1\}$) we search for an hyperplane of maximum thickness separating the points and solve

$$(P) \begin{cases} \min\limits_{w,b} \|w\|^2 \\ \forall x_i \in \mathcal{X}_1, \ \langle w, x_i \rangle + b \geq 1 \\ \forall x_i \in \mathcal{X}_{-1}, \ \langle w, x_i \rangle + b \leq -1 \end{cases}$$

Trade-off between Margin and Errors

(P) can also be written as:

$$\begin{cases} \min\limits_{w,b} \|w\|^2 \\ \forall (x_i, y_i) \in \mathcal{S}, y_i \left[\langle w, x_i \rangle + b\right] \geq 1 \end{cases}$$

When the points cannot be totally separated (i.e the domain of (P) is $\varnothing$) we search for w, b and $\xi = (\xi_i)_{i \in [\![1,n]\!]} \in \mathbb{R}^n$ solutions of:

$$(P_C) \begin{cases} \min\limits_{w,b,\{\xi_i\}_{i \in [\![1,n]\!]}} \|w\|^2 + C \sum\limits_{i=1}^{i=n} \xi_i \\ \forall (x_i, y_i) \in \mathcal{S},\ y_i [\langle w, x_i \rangle + b] \geq 1 - \xi_i \\ \forall i \in [\![1, n]\!], \xi_i \geq 0 \end{cases}$$

Trade-off between Margin and Errors

Remarks:

- the $(\xi_i)_{i\in[\![1,n]\!]}$ are (slack) variables which enable the relaxation of the constraints of strict separability of the $(x_i, y_i) \in \mathcal{S}$
- the parameter $C \geq 0$ is a cost of not separating a point correctly, based on the distance between this point and the frontier of the hyperplane defining its class
- other cost functions could have been used for mis-classification such as $C\sum_{i=1}^{i=n}\xi_i^2$ or $C\sum_{i=1}^{i=n}1_{\xi_i>0}$ but with slightly different solutions and interpretations for (P_C)
- at this point we have not defined what would be the strategy of classification for a new observation x lying between $H_{w,b-1}$ and $H_{-w,-b-1}$ and the function $C\sum_{i=1}^{i=n}\xi_i^2$ is linked to the error of classification but not to an empirical risk of classification.

Trade-off between Margin and Errors: Resolution

to solve:

$$(P_C)\begin{cases} \min\limits_{w,b,\xi\in\mathbb{R}^n} \|w\|^2 + C\sum\limits_{i=1}^{i=n}\xi_i \\ \forall (x_i,y_i)\in\mathcal{S},\ y_i[\langle w,x_i\rangle + b]\geq 1-\xi_i\ (1) \\ \forall i\in[\![1,n]\!],\xi_i\geq 0\ (2) \end{cases}$$

we consider the Lagrangian:

$$L(w,b,\xi,\alpha,\mu) = \|w\|^2 + C\sum_{i=1}^{i=n}\xi_i - \sum_{i=1}^{i=n}\alpha_i\left(y_i[\langle w,x_i\rangle + b] - 1 + \xi_i\right) - \sum_{i=1}^{i=n}\mu_i\xi_i$$

with $\xi = (\xi_1,\xi_2,\cdots,\xi_n)'$, $\alpha = (\alpha_1,\alpha_2,\cdots,\alpha_n)'$ and $\mu = (\mu_1,\mu_2,\cdots,\mu_n)'$

Lemma 1: (property of the Lagrangian)

$\max\limits_{\alpha\in(\mathbb{R}^+)^n\mu\in(\mathbb{R}^+)^n} L(w,b,\xi,\alpha,\mu)$ equals:

$$\begin{cases} +\infty \text{ if either (1) or (2) are not satisfied} \\ \|w\|^2 + C\sum\limits_{i=1}^{i=n}\xi_i \text{ if both (1) and (2) are satisfied} \end{cases}$$

Trade-off between Margin and Errors: Resolution

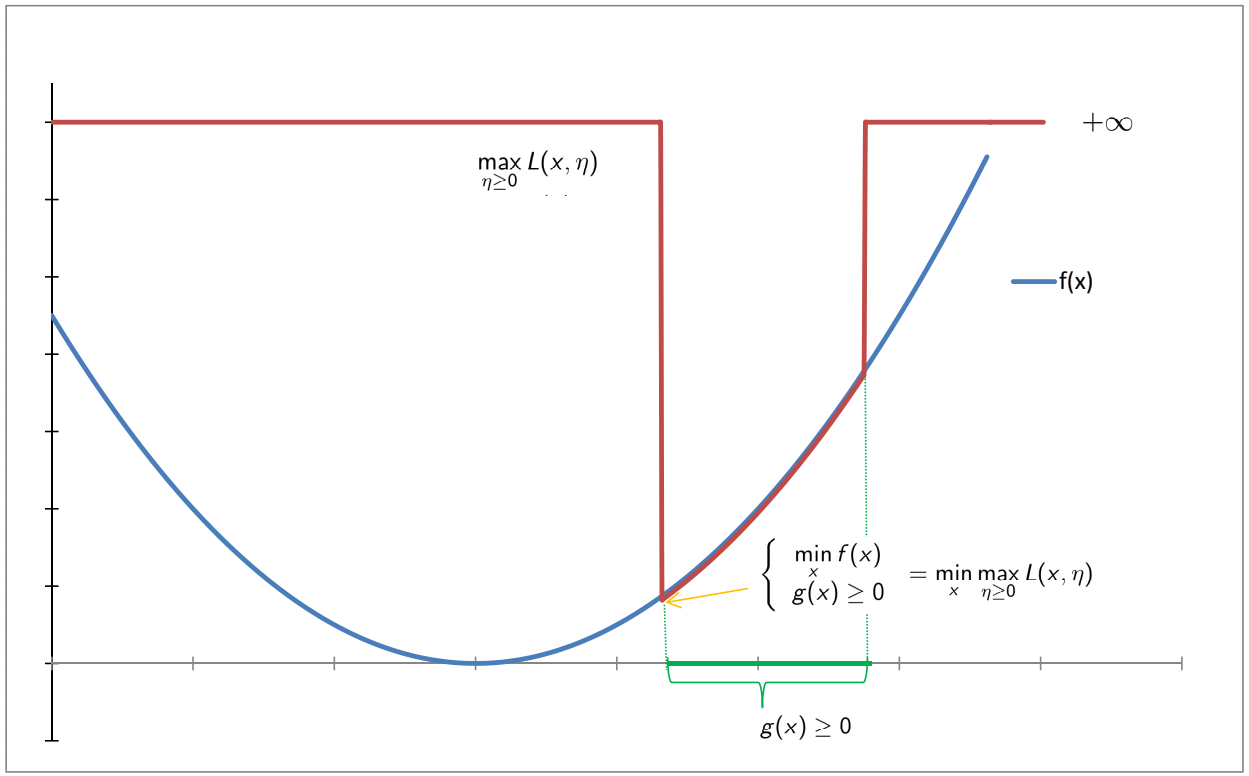
Demonstration:

$y_i[\langle w, x_i\rangle + b - 1 + \xi_i] < 0 \Rightarrow \lim_{\alpha_i \to +\infty} L(w, b, \xi, \alpha, \mu) = +\infty$ and $\xi_i < 0 \Rightarrow \lim_{\mu_i \to +\infty} L(w, b, \xi, \alpha, \mu) = +\infty$. This proves the first part.

Now, if both (1) and (2) are satisfied then $\forall \alpha \in (\mathbb{R}^+)^n, \forall \mu \in (\mathbb{R}^+)^n$ $-\sum_{i=1}^{i=n} \alpha_i \left(y_i[\langle w, x_i\rangle + b - 1 + \xi_i]\right) - \sum_{i=1}^{i=n} \mu_i \xi_i \geq 0$ and so the minimum (of zero) is attained for $\alpha = \mu = 0$.

As $L(w, b, \xi, 0, 0) = \|w\|^2 + C \sum_{i=1}^{i=n} \xi_i$ this proves the result.

Trade-off between Margin and Errors: Resolution



Lagrangian principle illustrated

Trade-off between Margin and Errors: Resolution

Lemma 2: (mini-max theorem)

For any domains $\mathcal{Y}$ and $\mathcal{Z}$ and real function g defined on $\mathcal{Y}x\mathcal{Z}$:

$$\max_{z\in\mathcal{Z}}\left[\min_{y\in\mathcal{Y}} g(y,z)\right] \leq \min_{y\in\mathcal{Y}}\left[\max_{z\in\mathcal{Z}} g(y,z)\right]$$

Demonstration:

$$\min_{y\in\mathcal{Y}} g(y,z) \leq g(y,z) \Rightarrow \max_{z\in\mathcal{Z}}\left[\min_{y\in\mathcal{Y}} g(y,z)\right] \leq \max_{z\in\mathcal{Z}} g(y,z) \ (1)$$

As (1) is true for all y the inequality stands for the *min* of the right term of (1). So, $\max_{z\in\mathcal{Z}}\left[\min_{y\in\mathcal{Y}} g(y,z)\right] \leq \min_{y\in\mathcal{Y}}\left[\max_{z\in\mathcal{Z}} g(y,z)\right]$ Q.E.D

Trade-off between Margin and Errors: Resolution

Table: example for mini-max

g(y,z)	y=1	y=2	y=3
z=3	3	3	1
z=2	2	1	3
z=1	1	2	3

for the example here:

$$\max_{z\in\mathcal{Z}}\left[\min_{y\in\mathcal{Y}} g(y,z)\right] = 1$$

$$\min_{y\in\mathcal{Y}}\left[\max_{z\in\mathcal{Z}} g(y,z)\right] = 3$$

Trade-off between Margin and Errors: Resolution

Lemma 3: (Lagrangian method)

Solving P_C is equivalent to solving:

$$\min_{w,b,\xi\in\mathbb{R}^n}\left[\max_{\alpha\in(\mathbb{R}^+)^n\mu\in(\mathbb{R}^+)^n} L(w,b,\xi,\alpha,\mu)\right]$$

Demonstration: this follows directly from lemma 1

Definition: Duality

$$\min_{w,b,\xi\in\mathbb{R}^n}\left[\max_{\alpha\in(\mathbb{R}^+)^n\mu\in(\mathbb{R}^+)^n} L(w,b,\xi,\alpha,\mu)\right] \text{ is called the primal problem}$$

$$\max_{\alpha\in(\mathbb{R}^+)^n\mu\in(\mathbb{R}^+)^n}\left[\min_{w,b,\xi\in\mathbb{R}^n} L(w,b,\xi,\alpha,\mu)\right] \text{ is called the dual problem}$$

The dual problem is noted P_C^*

Trade-off between Margin and Errors: Resolution

Remarks:

- if d is the value obtained for the dual problem and p for the primal problem then according to the mini-max lemma $d \leq p$
- according to the KKT theorem, a way to guarantee that $d = p$ is, when solving the dual problem, $\min_{w,b,\xi \in R^n} L(w, b, \xi, \alpha, \mu)$ to impose some additional constraints, known as the "complementary slackness" conditions, defined here by:
  (KKT1): $\forall i \in [\![1, n]\!]$, $\alpha_i \left(y_i[\langle w, x_i \rangle + b] - 1 + \xi_i\right) = 0$
  (KKT2): $\forall i \in [\![1, n]\!]$, $\mu_i \xi_i = 0$.
  The effect of these complementary constraints is to increase d up to d^* such that $d^* = p$.
- some convex analysis results guarantee that for the (P_C) here (convex function optimized under affine constraints defining a domain of non empty interior) $d = d^* = p$. Based on this (KKT1) and (KKT2) will not be added to the constraints for P_C^* but will just be used as auxiliary equations when useful.

SVM and C-SVM

SVM and C-SVM: Solving the Dual Problem

to solve $\max_{\alpha\in(\mathbb{R}^+)^n \mu\in(\mathbb{R}^+)^n} \left[\min_{w,b,\xi\in\mathbb{R}^n} L(w,b,\xi,\alpha,\mu) \right]$ we first solve $\min_{w,b,\xi\in\mathbb{R}^n} L(w,b,\xi,\alpha,\mu)$ as a function of α and μ.

$$L(w,b,\xi,\alpha,\mu) =$$

$$= \|w\|^2 + \sum_{i=1}^{i=n} \xi_i(C-\alpha_i-\mu_i) - \langle w, \sum_{i=1}^{i=n} \alpha_i y_i x_i \rangle - b \sum_{i=1}^{i=n} \alpha_i y_i + \sum_{i=1}^{i=n} \alpha_i$$

$\frac{\partial L}{\partial w}$ is defined as $(\frac{\partial L}{\partial w_1}, \frac{\partial L}{\partial w_2}, \cdots, \frac{\partial L}{\partial w_n})$

$$\frac{\partial L}{\partial w} = 2w' - \sum_{i=1}^{i=n} \alpha_i y_i x_i \Rightarrow w = \frac{1}{2} \sum_{i=1}^{i=n} \alpha_i y_i x_i \ (C1)$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{i=n} \alpha_i y_i = 0 \ (C2)$$

$$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow C - \alpha_i - \mu_i = 0 \ (C3)$$

SVM and C-SVM: Solving the Dual Problem

so by duality:

$$(P_C) \Leftrightarrow \begin{cases} \max\limits_{\alpha \in (\mathbb{R}^+)^n \mu \in (\mathbb{R}^+)^n} -\frac{1}{4}\| \sum\limits_{i=1}^{i=n} \alpha_i x_i \|^2 + \sum\limits_{i=1}^{i=n} \alpha_i \\ C - \alpha_i - \mu_i = 0 \\ \sum\limits_{i=1}^{i=n} \alpha_i y_i = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \max\limits_{\alpha \in \mathbb{R}^n} -\frac{1}{4}\| \sum\limits_{i=1}^{i=n} \alpha_i x_i \|^2 + \sum\limits_{i=1}^{i=n} \alpha_i \\ 0 \leq \alpha_i \leq C \\ \sum\limits_{i=1}^{i=n} \alpha_i y_i = 0 \end{cases}$$

which can be solved numerically

we note α^* the solution of this system

SVM and C-SVM: Interpretation

Remarks:

- from $(C1)$: $w^* = \frac{1}{2}\sum_{i=1}^{i=n} y_i \alpha_i^* x_i$
- from $(KKT1)$, $(KKT2)$ and $(C3)$:
  $\begin{cases} \forall i \in [\![1,n]\!],\ (C-\alpha_i^*)\xi_i = 0 \text{ (as } \mu_i^* = C-\alpha_i^*) \\ \forall i \in [\![1,n]\!],\ \alpha_i^*\left(y_i[\langle w^*, x_i\rangle + b] - 1 + \xi_i\right) = 0 \end{cases}$
  so b^* can be determined by picking indices i for which $0 < \alpha_i^* < C$ as in this case: $\xi_i = 0$ and consequently $y_i[\langle w^*, x_i\rangle + b^*] - 1 = 0$,
  leading to : $b^* = y_i - \langle w^*, x_i\rangle$.
  Note that in practice, as in the determination of α^* there be some approximation errors, b^* is calculated as the average of $y_i - \langle w^*, x_i\rangle$ for the indices i for which $0 < \alpha_i^* < C$.
- if we assume that we have at least one represent of each class in the sample, the condition $\sum_{i=1}^{i=n} \alpha_i y_i = 0$ implies that there are some α_i^* satisfying $0 < \alpha_i^* < C$.

SVM and C-SVM: Interpretation

Remarks:

two types of vectors x_i are used to determine w^*

- the x_i for which $0 < \alpha_i^* < C$
  in this case $(KKT2)$ and $(C3) \Rightarrow \xi_i = 0$ and $y_i[\langle w^*, x_i \rangle + b] - 1 = 0$
  and these x_i are well classified and on the two separating hyperplanes $H_{w,b-1}$ and $H_{-w,-b-1}$
- the x_i for which $\alpha_i^* = C$ which can be misclassified as in this case there is no constraint of nullity on ξ_i derived from $(KKT2)$ and $(C3)$

Definition : Support Vector, Support Vector Machines

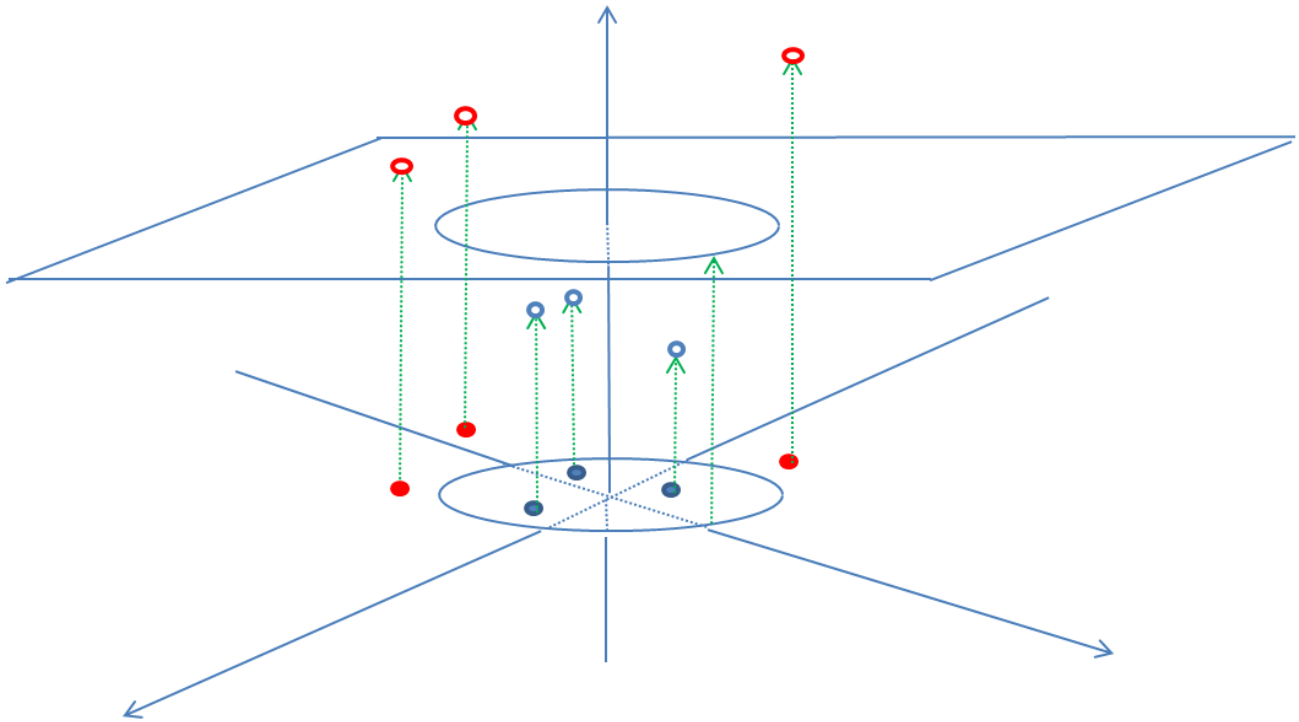
The vectors x_i for which $\alpha_i^* \neq 0$ and which are used for the expression of w^* are called "support vectors". The method of classification is then called "Support Vector Machines" and noted "SVM".
When some errors are permitted in the classification, with the introduction of the "slack variables" ξ and the cost C the method is called C-SVM.

SVM and C-SVM: Interpretation

Remarks:
When the points from the sample are perfectly separable the solution of (P) correspond to the solutions of (P_C) for C large enough. Indeed if the cost C is large enough the solution of (P_C) will maximize the margin while constricting the ξ_i to zero.

The Kernel Trick

The Kernel Trick



Classification after a change of variable

The Kernel Trick

In some situations the $(x_i, y_i)_{i\in[\![1,n]\!]}$ cannot be separated by an hyperplane in $\mathbb{R}^d$ but it is possible to find a transformation ϕ such that the $(\phi(x_i), y_i)_{i\in[\![1,n]\!]}$ are separable.

Example: Consider in $\mathbb{R}^2$ the classification of (X, Y) where $X = (X^1, X^2)'$ and $Y = 1_{(X^1)^2+(X^2)^2\leq 1}$. In the graph we represent a sample of 6 points $(x_i, y_i)_{i\in[\![1,6]\!]}$. The blue points are the points for which $Y_i = 1$ and the red points the points for which $Y_i = 0$.
It appears that we cannot separate correctly these points in $\mathbb{R}^2$.
If we consider now,

$$\phi : \begin{pmatrix}\alpha\\ \beta\end{pmatrix} \longrightarrow \begin{pmatrix}\alpha\\ \beta\\ \alpha^2+\beta^2\end{pmatrix}$$ then the points $(\phi(x_i), y_i)_{i\in[\![1,n]\!]}$ can be separated by the hyperplane H of $\mathbb{R}^3$ defined by:

$$H = \left\{ \begin{pmatrix}\alpha\\ \beta\\ \gamma\end{pmatrix} \in \mathbb{R}^3, \gamma = 1 \right\}$$

The Kernel Trick

We now consider immersions ϕ: $R^d \longrightarrow l_2(\mathbb{R}, \langle , \rangle_{\mathbb{N}})$, where:
$l_2(\mathbb{R}, \langle , \rangle_{\mathbb{N}})$ is the vector space of sequences $(z_i)_{i\in\mathbb{N}}$ such that $\sum_{i\in\mathbb{N}} z_i^2 < +\infty$
and $\langle , \rangle_{\mathbb{N}}$ is defined by $\langle (z_i)_{i\in\mathbb{N}}, (t_j)_{j\in\mathbb{N}} \rangle_{\mathbb{N}} = \sum_{i\in\mathbb{N}} z_i t_i$

In the space $Vect\{\phi(x_i)\}_{i\in\mathbb{N}}$ a C-SVM classifies a new observation $\phi(x)$ based on the values of $\sum_{i=1}^{i=n} \alpha_i^* y_i \langle \phi(x_i^*), \phi(x) \rangle_{\mathbb{N}} + b^*$ (1)

We can write (1) as $\sum_{i=1}^{i=n} \alpha_i^* y_i K_\phi(x_i^*, x) + b^*$ where K_ϕ: $R^d x \mathbb{R}^d \longrightarrow \mathbb{R}$ is defined by $K_\phi(x, z) = \langle \phi(x), \phi(z) \rangle_{\mathbb{N}}$

To determine what flexibility we earn by using classificators based on functions K_ϕ we are going to determine what the set of functions $\{K_\phi\}$ is. For this purpose we use Mercer's theorem.

The Kernel Trick

Theorem and Definition : Mercer's Theorem

Let $K : \mathbb{R}^d x \mathbb{R}^d \longrightarrow \mathbb{R}$ be such that,

- $\forall x, y \in \mathbb{R}^d$, $K(x, y) = K(y, x)$
- $\forall f \in L^2(\mathbb{R}^d, \mathbb{R})$, $\int K(x, y) f(y) dy \in L^2(\mathbb{R}^d, \mathbb{R})$

If we define $\langle ., . \rangle_K : L^2(\mathbb{R}^d, \mathbb{R}) x L^2(\mathbb{R}^d, \mathbb{R}) \longrightarrow \mathbb{R}$ by,
$\langle f, g \rangle_K = \int K(x, y) f(x) g(y) dx dy$
then the two following propositions are equivalent:

- (P1): $\exists \phi : \mathbb{R}^d \longrightarrow \mathbb{R}^{\mathbb{N}}$ such that $K(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathbb{N}}$
- (P2): the bilinear symmetric form $\langle ., . \rangle_K$ is positive on $L^2(\mathbb{R}^d, \mathbb{R})$.

A function K satisfying these properties will be called a Kernel.

The Kernel Trick

Demonstration:

We assume (P1), then $\forall f \in L^2(\mathbb{R}^d, \mathbb{R})$:

$\langle f, f \rangle_K = \int \int K(x,y) f(x) f(y) dx dy = \int \int \langle \phi(x), \phi(y) \rangle_{\mathbb{N}} f(x) f(y) dx dy$

$= \langle \int \phi(x) f(x) dx, \int \phi(y) f(y) dy \rangle_{\mathbb{N}} = \| \int \phi(x) f(x) dx \|^2_{\mathbb{N}}$.

So $\langle ., . \rangle_K$ is positive

We assume (P2). As $\langle ., . \rangle_K$ is symmetric it can be diagonalised so $\exists (e_i)_{i \in \mathbb{N}} \in L^2(\mathbb{R}^d, \mathbb{R})$ and $(\lambda_i)_{i \in \mathbb{N}}$ elements of $\mathbb{R}$ such that:

- $\langle e_i, e_j \rangle_{L^2} = \delta_{i,j}$
- $\forall f \in L^2(\mathbb{R}^d, \mathbb{R}), \langle f, e_i \rangle_K = \lambda_i \langle f, e_i \rangle_{L^2}$

based on this, if f and g are in $L^2(\mathbb{R}^d, \mathbb{R})$, after decomposing f and g on the orthonormal basis $(e_i)_{i \in \mathbb{N}}$ we get:

$\langle f, g \rangle_K = \langle \sum_{i \in \mathbb{N}} \langle f, e_i \rangle_{L^2} e_i, \sum_{j \in \mathbb{N}} \langle g, e_j \rangle_{L^2} e_j \rangle_K = \sum_{i,j \in \mathbb{N}} \langle f, e_i \rangle_{L^2} \langle g, e_j \rangle_{L^2} \langle e_i, e_j \rangle_K$

$= \sum_{i \in \mathbb{N}} \langle f, e_i \rangle_{L^2} \langle g, e_i \rangle_{L^2} \lambda_i$.

The Kernel Trick

we want to equate this expression to:
$\int\int \langle \phi(x), \phi(y) \rangle_{\mathbb{N}} f(x) g(y) dx dy = \langle \int \phi(x) f(x) dx, \int \phi(y) g(y) dy \rangle_{\mathbb{N}}$
$= \langle \langle \phi, f \rangle_{L^2}, \langle \phi, g \rangle_{L^2} \rangle_{\mathbb{N}}$
As $\langle ., . \rangle_K$ is assumed to be positive $\lambda_i \geq 0$.
Taking $\phi(x) = (\sqrt{\lambda_i} e_i(x))_{i \in \mathbb{N}}$ we get:
$\langle \phi, f \rangle_{L^2} = (\langle f, \sqrt{\lambda_i} e_i \rangle_{L^2})_{i \in \mathbb{N}}$ and
$\langle \langle \phi, f \rangle_{L^2}, \langle \phi, g \rangle_{L^2} \rangle_{\mathbb{N}} = \sum_{i \in \mathbb{N}} \langle f, \sqrt{\lambda_i} e_i \rangle_{L^2} \langle g, \sqrt{\lambda_i} e_i \rangle_{L^2}$ which equates $\langle f, g \rangle_K$.
As this is true for all f and g, $K(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathbb{N}}$. Q.E.D

Remarks:

based on Mercer's theorem we know that

- $K(x, y) = \sum_{i \in \mathbb{N}} \lambda_i e_i(x) e_i(y)$ with $\lambda_i \geq 0$
- $\|\phi(x)\|_N^2 = \sum_{i \in \mathbb{N}} \lambda_i e_i^2(x)$ and $\int \|\phi(x)\|_{\mathbb{N}}^2 dx = \sum_{i \in \mathbb{N}} \lambda_i$

The Kernel Trick: Radial Basis Functions

Theorem: Example of Kernels

a) $\forall k \in \mathbb{N} : (x, y) \longrightarrow \langle x, y \rangle_d^k$ is a kernel
b) $(x, y) \longrightarrow exp(-\|x - y\|_d^2)$ is a kernel

Demonstration:

$$\int \int \langle x, y \rangle_d^k f(x) f(y) dx dy = \int \int (\sum_{i=1}^{i=d} x^i y^i)^k f(x) f(y) dx dy$$

$$= \int \int \sum_{i_1, i_2, \cdots, i_k} x_{i_1} x_{i_2} \cdots x_{i_k} y_{i_1} y_{i_2} \cdots y_{i_k} f(x) f(y) dx dy$$

$$= \sum_{i_1, i_2, \cdots, i_k} \left(\int x_{i_1} x_{i_2} \cdots x_{i_k} f(x) dx \right) \left(\int y_{i_1} y_{i_2} \cdots y_{i_k} f(y) dy \right)$$

$$= \sum_{i_1, i_2, \cdots, i_k} \left(\int x_{i_1} x_{i_2} \cdots x_{i_k} f(x) dx \right)^2 \geq 0$$

As the form is positive, according to Mercer's Theorem $\langle x, y \rangle_d^k$ is a kernel.
Q.E.D

The Kernel Trick: Radial Basis Functions

$$\int\int exp(-\|x-y\|_d^2) f(x) f(y) dxdy$$
$$= \int\int exp(\langle x, y\rangle_d) exp(-\|x\|_d^2) exp(-\|y\|_d^2) f(x) f(y) dxdy$$
$$= \sum_{k\in\mathbb{N}} \int\int \frac{\langle x,y\rangle_d^k}{k!} \left[exp(-\|x\|_d^2) f(x)\right] \left[exp(-\|y\|_d^2) f(y)\right] dxdy$$

as $\langle x, y\rangle_d^k$ is a kernel, each of the terms are positive, so the sum is positive, so $exp(-\|x-y\|_d^2)$ is a kernel. Q.E.D

Remarks:

- $\forall \sigma \in \mathbb{R}$, $exp(-\frac{\|x-y\|_d^2}{2\sigma^2})$ is a kernel, called the "Gaussian Kernel"
- $h(x, y)$ is called radial basis function i.i.f we can find ψ such that $h(x, y) = \psi(\|x - y\|_d)$
- the Gaussian kernel is a radial basis function

The Kernel Trick: Radial Basis Functions

Proposition

When a (supervised) learning is made on the sample $(x_i, y_i)_{i\in\langle 1,n\rangle}$ of $\mathbb{R}^d$ using a kernel K, a new point x is classified based on the value of $\sum_{i=1}^{i=d} \alpha_i^* y_i K(x_i^*, x) + b^*$

Demonstration: if ϕ is the immersion associated to K, by using the Kernel method we (SVM or C-SVM) classify the points $(\phi(x_i), y_i)_{i\in\langle 1,n\rangle}$ with an hyperplane from $Vect\{\phi(x_i)\}_{i\in[\![1,n]\!]}$ whose equation for any point z of $\mathbb{R}^{\mathbb{N}}$ is $\sum_{i=1}^{i=n} \alpha_i^* y_i \langle \phi(x_i^*), z\rangle_{\mathbb{N}} + b^* = 0$. Any new observation $x \in R^d$ is classified depending on the value of $\sum_{i=1}^{i=n} \alpha_i^* y_i \langle \phi(x_i^*), \phi(x)\rangle_{\mathbb{N}} + b^*$

$= \sum_{i=1}^{i=n} \alpha_i^* y_i K(x_i^*, x) + b^*$.Q.E.D

The Kernel Trick: Radial Basis Functions

Remark 1: for a C-SVM the margin of the hyperplane in $\phi(\mathbb{R}^d)$ is defined by $\frac{2}{\|\omega^*\|_{\mathbb{N}}}$ where $\omega^* = \sum_{i=1}^{i=n} \alpha_i^* y_i \phi(x_i^*)$. This quantity can be calculated from the kernel K as $\| \sum_{i=1}^{i=n} \alpha_i^* y_i \phi(x_i^*) \|_{\mathbb{N}} = \alpha^{*\prime}[K(x_i, x_j)]\alpha^*$ where $[K(x_i, x_j)]$ is the matrix of $\mathbb{R}^n x \mathbb{R}^n$ formed by the $\{K(x_i, x_j)\}_{i,j \in [\![1,n]\!]}$.

Remark 2: it is easy to verify that in $\mathbb{R}^{\mathbb{N}}$ the hyperplanes of equation $\langle \omega, x \rangle_{\mathbb{N}} + b = 0$ have an infinite VC dimension therefore we may wonder if the Kernel method is going to lead to some over-fitting. For a Gaussian Kernel $K_\sigma(x, y) = exp(-\frac{\|x-y\|}{2\sigma^2})$ (linked to an immersion ϕ_σ that we do not need to calculate) we can make the following remarks:

- $\forall x \in \mathbb{R}^d$, $\|\phi(x)\|_{\mathbb{N}} = 1$ because $\|\phi(x)\|_{\mathbb{N}}^2 = K(x, x) = exp(-\frac{\|0\|}{2\sigma^2}) = 1$ so the points to be classified are localized on the surface of the sphere centred on zero and of radius 1 of $\mathbb{R}^{\mathbb{N}}$, which is quite restrictive.

The Kernel Trick: Radial Basis Functions

- $\forall x, y \in \mathbb{R}^d$, $\langle \phi(x), \phi(y) \rangle_{\mathbb{N}} \geq 0$ because $K_\sigma(x, y) \geq 0$. so, all the points to classify are situated in the same orthant of $\mathbb{R}^{\mathbb{N}}$ (which restricts further where the points can lay).
- A $C - SVM$ classifier of $\mathbb{R}^d$ of Kernel $K_\sigma(x, y)$ can be seen as a Δ-GAP-tolerant classifiers of $R^{\mathbb{N}}$ of radius 1 with $\Delta = \frac{2}{\alpha^{*\prime}[K(x_i, x_j)]\alpha^*}$. As the VC dimension of Δ-GAP-tolerant classifiers of radius 1 is bounded by $1 + \frac{4}{\Delta^2}$ (and is not infinite as the dimension of $\mathbb{R}^{\mathbb{N}}$) we end up controlling the Vapnik dimension if $\Delta = \frac{2}{\alpha^{*\prime}[K(x_i, x_j)]\alpha^*}$ is not too small.

The Kernel Trick: Radial Basis Functions

Exercice:

Let $\mathcal{F}_\sigma^n = \left\{ \begin{array}{c} h : \mathbb{R}^d \longrightarrow \{-1,1\}, h(x) = \Theta\left(\sum_{i=1}^{i=n} \mu_i K_\sigma(x, z_i) + b\right), \\ z_i \in \mathbb{R}^d, \mu_i \in \mathbb{R}, b \in \mathbb{R}, \mu'[K(z_i, z_j)]\mu \neq 0 \end{array} \right\}$

where, $\Theta(u) = 1$ if $u \geq 0$ and otherwise $\Theta(u) = -1$ be a machine of $\{-1,1\}$-classifiers and let $(x_i)_{i\in[\![1,n]\!]}$ be n distinct points of $\mathbb{R}^d$, and d be the minimum distance between the points i.e $d = \min_{i\neq j} \|x_i - x_j\|_d$.

Let σ be such that $(n-1)exp(-\frac{d}{2\sigma^2}) < 1$.

Let $(y_i)_{i\in[\![1,n]\!]}$ be a $\{-1,1\}$ labelling of the $(x_i)_{i\in[\![1,n]\!]}$.

a) show that $\Theta\left(\sum_{i=1}^{i=n} y_i K_\sigma(x, z_i)\right)$ classifies correctly the $(x_i, y_i)_{i\in[\![1,n]\!]}$

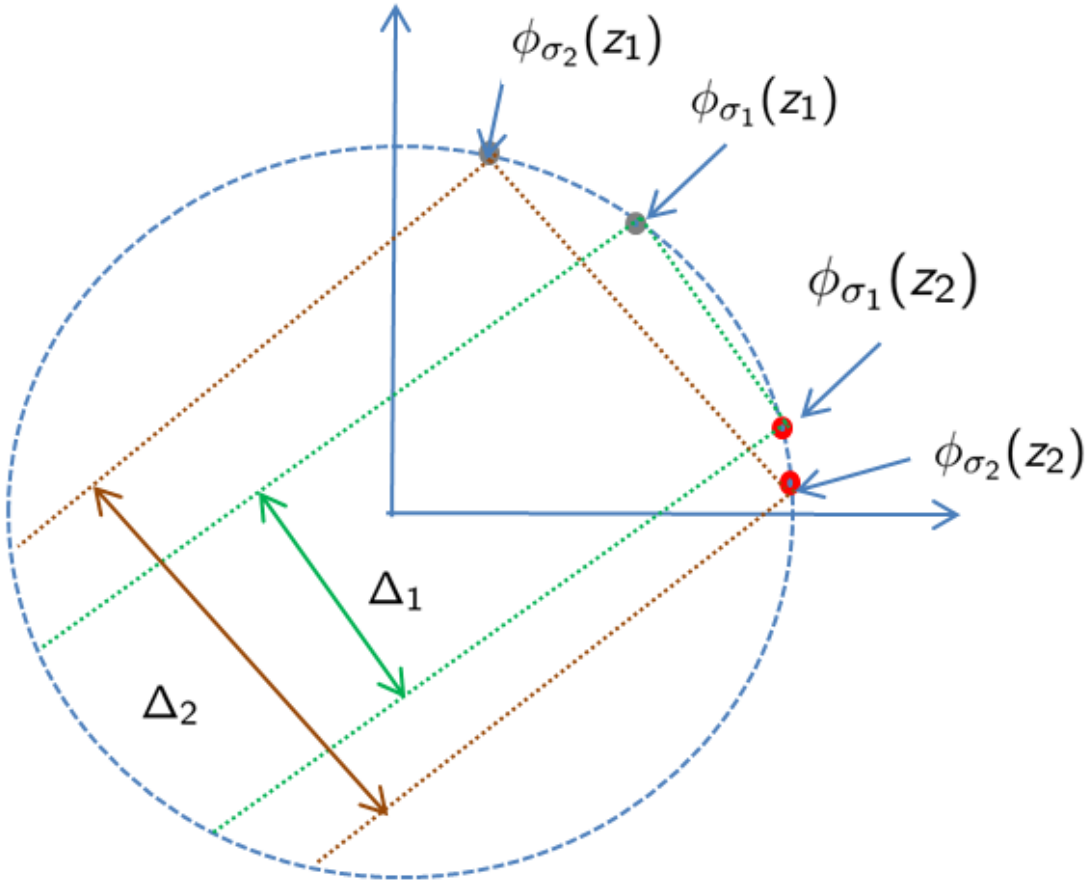
b) deduct from a) that $VC(\mathcal{F}_\sigma^n) \geq n$

The Kernel Trick: Radial Basis Functions

Remarks:

- Let $(x_i)_{i\in[\![1,n]\!]}$ be n distinct points of $\mathbb{R}^d$, then $\forall i \neq j$ $\langle \phi_\sigma(x_i), \phi_\sigma(x_j)\rangle_{\mathbb{N}} \underset{\sigma\to 0}{\longrightarrow} 0$ so in the limit the $(\phi_\sigma(x_i))_{i\in[\![1,n]\!]}$ are orthonormal in $\mathbb{R}^{\mathbb{N}}$ and thus independent and therefore separable by an hyperplane of the vector space they generate in $\mathbb{R}^{\mathbb{N}}$. So if σ is small enough the $(x_i)_{i\in[\![1,n]\!]}$ can be labelled as desired in $\mathbb{R}^d$.
- as we will see later if n^+ points are labelled 1 and n^- are labelled -1 on a sphere of radius 1 and are orthogonal they can be separated by an hyperplane of margin $\sqrt{\frac{1}{n^+}+\frac{1}{n^-}}$. So if σ is very small it is not an achievement to be able to separate the points with such a margin as random orthogonal points on the sphere with random labelling could be classified with this margin.
- If there is a real structure the classes should be separable without having to totally "orthogonalize" the observations.
- In general cross-validation will be used to justify that the model is adequate.

The Kernel Trick: classifications for various parameters



The margin increases as σ decreases ($\sigma_2 < \sigma_1$) and the points on the sphere are "orthogonalized"

The Kernel Trick: Radial Basis Functions

Example:

We consider the Δ-classifier h in $\mathbb{R}^2$ defined by:

$$h(x) = 1 \Leftrightarrow \sum_{i=1}^{i=4} \alpha_i K_\sigma(x, z_i) + b_1 \geq 0$$

$$h(x) = -1 \Leftrightarrow \sum_{i=1}^{i=4} \alpha_i K_\sigma(x, z_i) + b_2 \leq 0 \text{ with}$$

$$z_1 = \begin{pmatrix} 0.2 \\ 0.2 \end{pmatrix} \; z_2 = \begin{pmatrix} 0.2 \\ 0.8 \end{pmatrix} \; z_3 = \begin{pmatrix} 0.8 \\ 0.2 \end{pmatrix} \; z_4 = \begin{pmatrix} 0.8 \\ 0.8 \end{pmatrix}$$

$$\alpha_1 = 1, \alpha_2 = -1, \alpha_3 = -1, \alpha_4 = 2$$

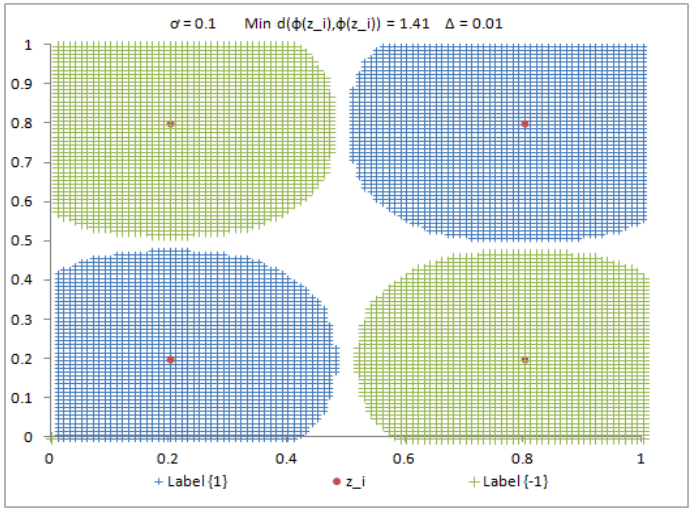
We colour in green the region classified $\{-1\}$, in blue the region classified $\{1\}$ and leave in white the rest of the space.

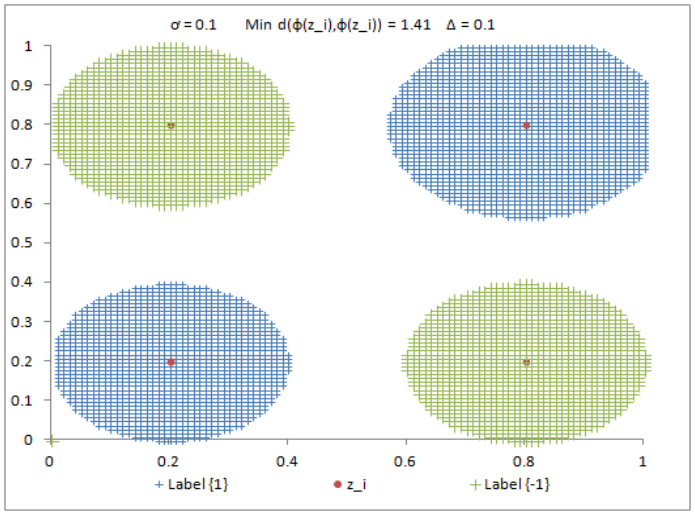
The Kernel Trick: Radial Basis Functions

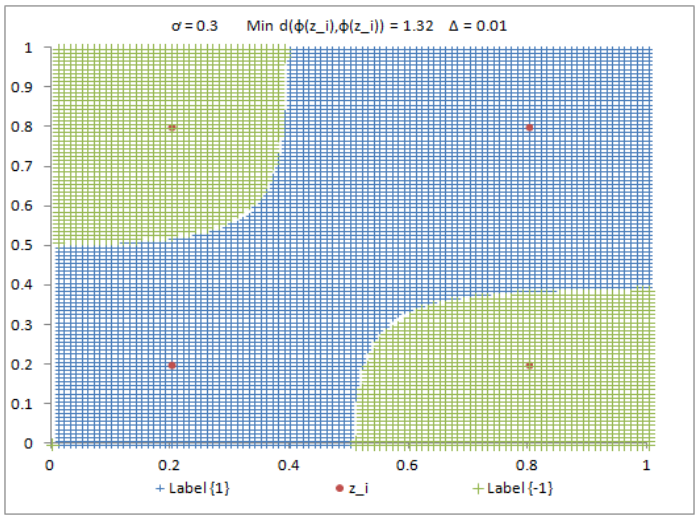
If we take $\sigma = 1$, $b_1 = -0.132$, $b_2 = 0.132$

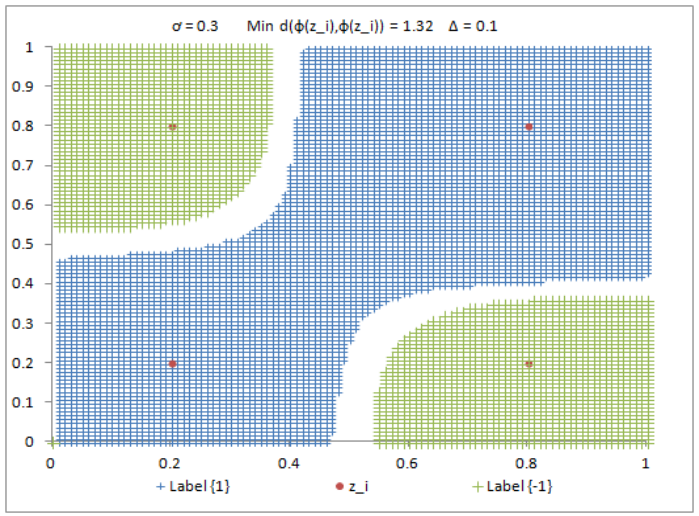
- the 4 points z_i are classified correctly
- as σ is small the classification of z_i is not impacted a lot by the classification of the other points $\{z_j \neq z_i\}$
- to understand how $\phi_\sigma(.)$ spread apart the points z_i (to facilitate their separation) we calculate $\min_{i \neq j} d(\phi_\sigma(z_i), \phi_\sigma(z_j)) = \min_{i \neq j} K_\sigma(z_i, z_j)$.
  Here $\min_{i \neq j} d(\phi_\sigma(z_i), \phi_\sigma(z_j)) = 1.41$, which means that the $\phi_\sigma(z_i)$ are "almost" orthogonal (they would be orthogonal for the value $\sqrt{1^2 + 1^2} = \sqrt{2}$).
- if we note $\omega = \sum_{i=1}^{i=4} \phi_\sigma(z_i)$ then $\|\omega\|_{\mathbb{N}} = \alpha'[K(z_i, z_j)]\alpha = 2.65$ here
- here $\Delta = \frac{|b_2 - b_1|}{||\omega||_{\mathbb{N}}} = 0.1$ (which is not too big compared to 1.41 ....)

The Kernel Trick: classifications for various parameters



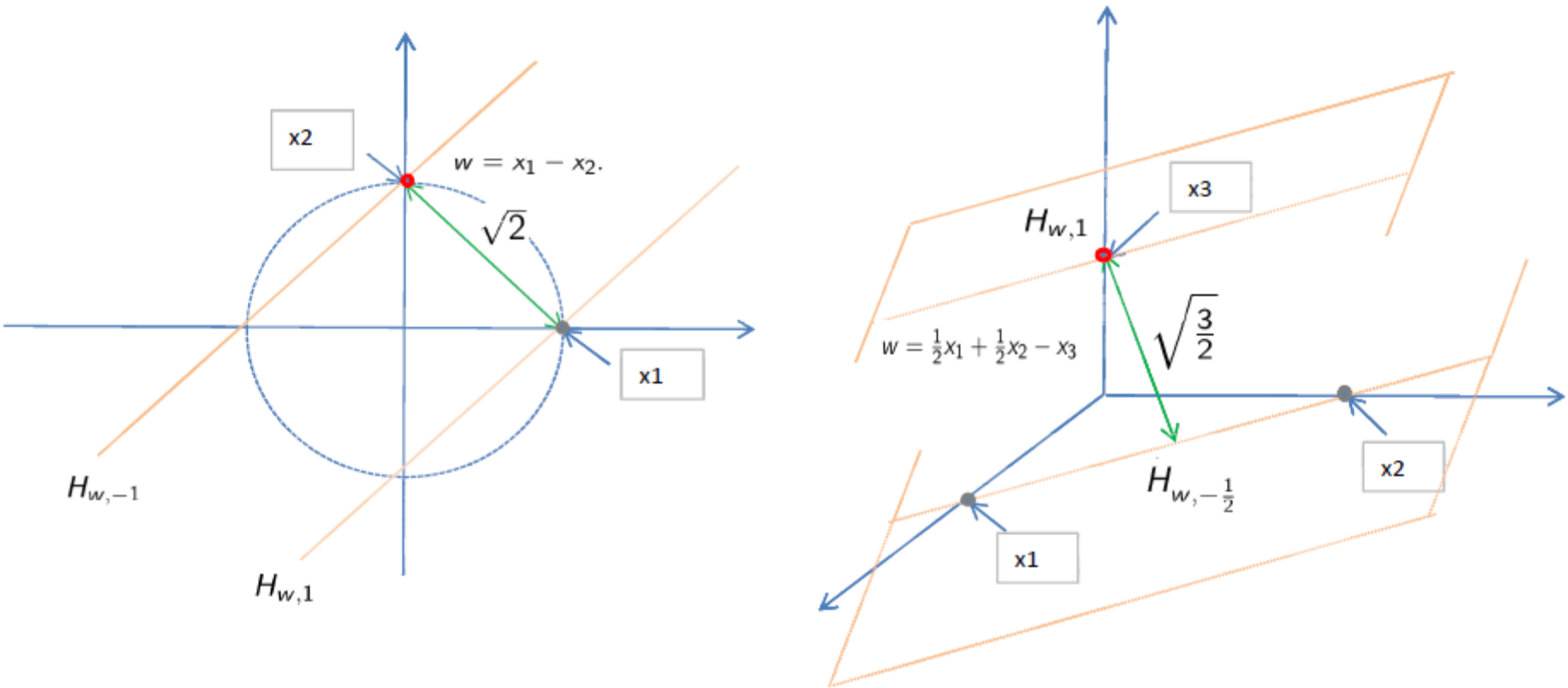






Shattering Orthogonal Vectors

Shattering Orthogonal Vectors



Maximum margin for separation of orthogonal points of $S^1_{\mathbb{N}}$

Shattering Orthogonal Vectors

Remarks:

- For two vectors x_1, x_2 of $S^1_{\mathbb{N}}$ orthogonal and classified 1 and -1, the maximum margin of an hyperplane separating them is $\sqrt{2}$. The hyperplanes forming the borders of the separation set are: $H_{w,-1}$ and $H_{w,1}$ with $w = x_1 - x_2$.
- For three vectors x_1, x_2, x_3 of $S^1_{\mathbb{N}}$ orthogonal and classified -1 for x_3 and 1 for others, the maximum margin of an hyperplane separating them is $\sqrt{\frac{3}{2}}$. The hyperplanes forming the borders of the separation set are: $H_{w,-\frac{1}{2}}$ and $H_{w,1}$ with $w = \frac{1}{2}x_1 + \frac{1}{2}x_2 - x_3$

Shattering Orthogonal Vectors

Proposition: Maximum Margin on $S^1_{\mathbb{N}}$

Let $\{x_i\}_{i\in[\![1,n^+]\!]}$ be n^+ vectors of $S^1_{\mathbb{N}}$ labelled 1 and $\{y_j\}_{j\in[\![1,n^-]\!]}$ be n^- vectors of $S^1_{\mathbb{N}}$ labelled -1. If we assume that the $\{x_i, y_j\}$ form a family of orthogonal vectors and define $w = \frac{1}{n^+}\sum_{i=1}^{i=n^+} x_i - \frac{1}{n^-}\sum_{j=1}^{j=n^-} y_j$ then :

- Any hyperplane of margin Δ which separates the $\{x_i\}$ from the $\{y_j\}$ satisfies $\Delta \leq \sqrt{\frac{1}{n^-}+\frac{1}{n^+}}$
- $H_{w,\frac{1}{n^+}}$ and $H_{w,-\frac{1}{n^-}}$ are the borders of the maximum margin hyperplane classifier which separates the $\{x_i\}$ from the $\{y_j\}$. $d(H_{w,\frac{1}{n^+}}, H_{w,-\frac{1}{n^-}}) = \sqrt{\frac{1}{n^-}+\frac{1}{n^+}}$

Shattering Orthogonal Vectors

Demonstration:

$w^+ = \frac{1}{n^+} \sum_{i=1}^{i=n^+} x_i$ belongs to the convex envelope of the $\{x_i\}$ and

$w^- = \frac{1}{n^-} \sum_{j=1}^{j=n^-} y_j$ belongs to the convex envelope of the $\{y_j\}$.

As the maximum margin is the distance between the two convex envelopes we have: $\Delta \leq MaxMargin = d(\mathcal{C}_x, \mathcal{C}_y) \leq d(w^+, w^-)$ and $d(w^+, w^-) = \sqrt{\frac{1}{n^-} + \frac{1}{n^+}}$ which proves the first bullet point.

As the vectors are orthogonal we have:

$\forall x_i \ \langle w, x_i \rangle = \frac{1}{n^+}$ and $\forall y_j = \langle w, y_j \rangle = -\frac{1}{n^-}$ so $H_{w, \frac{1}{n^+}}$ and $H_{w, -\frac{1}{n^-}}$ separate the points. We also have $d(H_{w, \frac{1}{n^+}}, H_{w, -\frac{1}{n^-}}) = \sqrt{\frac{1}{n^-} + \frac{1}{n^+}}$ which means that the maximum margin is reached for $H_{w, \frac{1}{n^+}}$ and $H_{w, -\frac{1}{n^-}}$ which therefore constitute the borders of the maximum margin hyperplane classifier.

Shattering Orthogonal Vectors

Remarks:

If we note $k+1 = n^+ + n^-$, $\{z_i\}_{i \in [\![1,1+k]\!]} = \{x_i\}_{i \in [\![1,n^+]\!]} \cup \{y_j\}_{j \in [\![1,n^-]\!]}$ and $z = \frac{1}{d} \sum_{i=1}^{i=1+k} z_i$ then:

- $\forall i \in [\![1, 1+k]\!]$, $d(z, z_i) = \sqrt{\frac{k}{k+1}}$ so z and the $k+1$ points z_i are in an affine space of dimension k and the z_i are on the sphere of center z and radius $\sqrt{\frac{k}{k+1}}$ of this affine space
- $\min_{i \in [\![0,k+1]\!]} \sqrt{\frac{1}{i} + \frac{1}{k+1-i}} = \sqrt{\frac{1}{[\frac{k+1}{2}]} + \frac{1}{k+1-[\frac{k+1}{2}]}}$
- the $k+1$ points $p_i = \sqrt{\frac{k+1}{k}}$ are on a sphere of radius 1 and according to the previous proposition can always be classified (whatever there label is) with a margin equal to $\sqrt{\frac{k+1}{k}} \sqrt{\frac{1}{[\frac{k+1}{2}]} + \frac{1}{k+1-[\frac{k+1}{2}]}}$

Shattering Orthogonal Vectors

Corollary: Maximum Margin on $S^1_{\mathbb{N}}$

The maximum margin for a Gap tolerant classifier of radius 1 for $k+1$ points in is attained by taking $k+1$ points p_i forming an orthogonal family with norms $\sqrt{\frac{k+1}{k}}$. Seen from the affine space of dimension k they generate these points lay on a sphere of radius 1 and can be separated with the maximum possible margin.

Demonstration:

According to the previous remarks the points z_i can be classified with Gap classifiers reaching the maximum margin according the admitted theorem in the section on Gap classifiers

Remarks:

The $k+1$ points p_i form a simplex in the affine space of dimension k that they generate as $\forall i \neq j,\ d(p_i, p_j) = \sqrt{2}$

Shattering Orthogonal Vectors

Example:

Let z_1, z_2, z_3 be three orthogonal vectors of norms 1. We note $w = \frac{1}{3}(z_1 + z_2 + z_3)$ and $H_{w,-1}$ the hyperplane of the vector space $Vect(z_1, z_2, z_3)$ defined by $H_{w,-1} = \{x \in Vect(z_1, z_2, z_3), \langle w, x \rangle = \frac{1}{3}\}$ then:

- z_1, z_2, z_3 and w belongs to $H_{w,-1}$ as they all verify $\langle w, x \rangle = \frac{1}{3}$
- z_1, z_2, z_3 lay on a circle of center w and radius $\sqrt{\frac{2}{3}}$ as they all verify $d(z_i, w) = \sqrt{(\frac{2}{3})^2 + (\frac{1}{3})^2 + (\frac{1}{3})^2}$
- z_1, z_2, z_3 form a simplex / equilateral triangle as for all $i \neq j$, $d(z_i, z_j) = \sqrt{2}$
- the distance between the segment (convex envelope) formed by any 2 points z_i, z_j and the third one z_k, which is also the maximum margin of an hyperplane classifier separating the points, equal: $d(\frac{1}{2}(z_i + z_j), z_k) = \sqrt{\frac{1}{4} + \frac{1}{4} + 1} = \sqrt{\frac{3}{2}}$

ν-SVM

ν-SVM: Schölkopf, Smola, Williamson, Bartlett

Definition: ν-SVM

For any learning sample $\{(x_i, y_i)\}_{i\epsilon[\![1,n]\!]}$ and $\nu > 0$ we call ν-SVM the solution of:

$$(P_\nu) \begin{cases} \min\limits_{w,b,\rho,\xi_i} \frac{1}{2}\|w\|^2 - 2\rho + \frac{\nu}{n}\sum\limits_{i=1}^{i=n}\xi_i \\ y_i(\langle w, x_i\rangle + b) \geq \rho - \xi_i \\ \xi_i \geq 0 \end{cases}$$

Remark 1:
The definition is similar to the definition of a $C - SVM$ but the new parameter ρ is introduced to enable a better geometric interpretation of the problem and to have an upper bound on the fraction of misclassified points ($\xi_i > 0$) and a lower bound on the fraction of support vectors. We did not put the condition $\rho \geq 0$ which is automatically verified for a solution of this problem.

ν-SVM

Remark 2:
In (P_ν) the two hyperplanes which classify the points -1 and 1 are $H_{w,b-\rho}$ and $H_{-w,-b+\rho}$ and the distance between them (which represents the margin of the classifier) is $\frac{2\rho}{\|w\|_d}$. In the minimization the quantity $\frac{\|w\|_d}{\rho}$ does not appear but instead the quantity $\|w\|_d - \rho$ which leads to simpler numerical implementations and geometric interpretations of the results.

Proposition: Dual Problem for ν-SVM

$$(P_\nu) \Leftrightarrow (D_\nu) \text{ where } (D_\nu) \begin{cases} -\frac{1}{2} \min\limits_{\alpha_i} \| \sum\limits_{i=1}^{i=n} \alpha_i y_i x_i \|_d^2 \\ \sum\limits_{i=1}^{i=n} \alpha_i y_i = 0 \\ \sum\limits_{i=1}^{i=n} \alpha_i = 2 \\ 0 \leq \alpha_i \leq \nu \end{cases} \quad \text{and } w^* = \sum_{i=1}^{i=n} \alpha_i^* y_i x_i$$

Demonstration (hint):

The Lagrangian $L(w, b, \rho, \xi, \alpha, \beta)$ of P_ν is:

$$\frac{1}{2}\|w\|^2 - 2\rho + \nu \sum_{i=1}^{i=n} \xi_i - \sum_{i=1}^{i=n} \alpha_i [y_i(\langle w, x_i \rangle + b) - \rho + \xi_i] - \sum_{i=1}^{i=n} \beta_i \xi_i.$$

so we get: $\frac{\partial L}{\partial w} = w' - \sum_{i=1}^{i=n} \alpha_i y_i x_i' = 0 \ (C_\nu 1)$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{i=n} \alpha_i y_i = 0 \ (C_\nu 2)$$

$$\frac{\partial L}{\partial \rho} = -2 + \sum_{i=1}^{i=n} \alpha_i = 0 \ (C_\nu 3)$$

$$\frac{\partial L}{\partial \xi_i} = \frac{\nu}{n} - \alpha_i - \beta_i = 0 \ (C_\nu 4)$$

From these equations we see that (D_ν) is the dual of (P_ν) and that consequently (due to the form of the problem) the solutions will be the same each time (D_ν) has a finite solution. We note also that for (D_ν) to have a finite solution we need $\nu \geq 2$ otherwise the last two constraints of (D_ν) cannot be satisfied simultaneously.

Theorem and Definition : Reduced Convex Envelope

Let $\{(x_i, y_i)\}_{i\epsilon[\![1,n]\!]}$ be a sample. We assume that the two classes $\{-1, 1\}$ are represented in this sample (i.e $\mathcal{X}_{-1} \neq \emptyset$ and $\mathcal{X}_1 \neq \emptyset$).

$$\text{Let } \mathcal{E}_\nu(\mathcal{X}_1) = \{ \sum_{\{i, y_i=1\}} \alpha_i x_i / \sum_{\{i, y_i=1\}} \alpha_i = 1 \text{ and } 0 \leq \alpha_i \leq \frac{\nu}{n} \} \text{ and}$$

$$\mathcal{E}_\nu(\mathcal{X}_{-1}) = \{ \sum_{\{i, y_i=-1\}} \alpha_i x_i / \sum_{\{i, y_i=-1\}} \alpha_i = 1 \text{ and } 0 \leq \alpha_i \leq \frac{\nu}{n} \} \text{ then:}$$

- $\mathcal{E}_\nu(\mathcal{X}_1)$ and $\mathcal{E}_\nu(\mathcal{X}_{-1})$ are convex sets and are called reduced convex envelopes of $\mathcal{X}_{-1}$ and $\mathcal{X}_1$
- finding $d(\mathcal{E}_\nu(\mathcal{X}_1), \mathcal{E}_\nu(\mathcal{X}_{-1}))$ and solving (D_ν) is the same problem

Demonstration (hint):

Demonstrating the convexity of $\mathcal{E}_\nu(\mathcal{X}_1)$ and $\mathcal{E}_\nu(\mathcal{X}_{-1})$ is straightforward.

The points on which $d(\mathcal{E}_\nu(\mathcal{X}_1), \mathcal{E}_\nu(\mathcal{X}_{-1}))$ is attained are the solutions of:

$$\begin{cases} \min_{\alpha_i \geq 0} \| \sum_{\{i, y_i = 1\}} \alpha_i x_i - \sum_{\{i, y_i = -1\}} \alpha_i x_i \|_d^2 \\ \sum_{\{i, y_i = 1\}} \alpha_i = 1 \\ \sum_{\{i, y_i = -1\}} \alpha_i = 1 \\ 0 \leq \alpha_i \leq \frac{\nu}{n} \end{cases} \Leftrightarrow \begin{cases} \min_{\alpha_i \geq 0} \| \sum_{i=1}^{i=n} \alpha_i x_i y_i \|_d^2 \\ \sum_{i=1}^{i=n} \alpha_i y_i = 0 \\ \sum_{i=1}^{i=n} \alpha_i = 2 \\ 0 \leq \alpha_i \leq \frac{\nu}{n} \end{cases}$$

Q.E.D

Corollary: Geometric Interpretation

If $z_1 \in \mathcal{E}_\nu(\mathcal{X}_1)$ and $z_2 \in \mathcal{E}_\nu(\mathcal{X}_{-1})$ verify $\|z_1 - z_2\|_d = d(\mathcal{E}_\nu(\mathcal{X}_1), \mathcal{E}_\nu(\mathcal{X}_{-1}))$ then $H_{w^*,b^*} \perp (z_1 - z_2)$ i.e the hyperplanes $H_{w^*,b^*-\rho^*}$ and $H_{-w^*,-b^*-\rho^*}$ derived from (P_ν) are both orthogonal to the segment of minimum distance between the two reduced convex envelopes of $\mathcal{X}_1$ and $\mathcal{X}_{-1}$.

Demonstration :

z_1 and z_2 are equal to $\sum_{\{i,y_i=1\}} \alpha_i^* x_i$ and $\sum_{\{i,y_i=-1\}} \alpha_i^* x_i$ so their difference equals $\sum_{i=1}^{i=n} \alpha_i^* y_i x_i$ which is also the expression of $w^*$for (P_ν). Q.E.D

Corollary: Number of Support Vectors, Number of Errors for (P_ν)

For the classification problem (P_ν):

- $\frac{1}{n}\#\{i, \xi_i \neq 0\} \leq \frac{2}{\nu}$ (majoration of the proportion of points from the sample misclassified)
- $\frac{1}{n}\#\{i, \alpha_i \neq 0\} \geq \frac{2}{\nu}$ (minoration of the proportion of points from the sample used as support vectors)

Demonstration : The KKT conditions for (P_ν) are:
$(KKT_\nu 1) : \alpha_i[y_i(\langle w, x_i\rangle + b) - \rho + \xi_i] = 0$
$(KKT_\nu 2) : \beta_i \xi_i = 0$
$(KKT_\nu 2)$ and $(C_\nu 4) \Rightarrow (\nu - \alpha_i)\xi_i = 0$ so $\xi_i \neq 0 \Rightarrow \alpha_i = \frac{\nu}{n}$
using $(C_\nu 3) : \sum_{i=1}^{i=n} \alpha_i = 2 \Rightarrow \sum_{i,\xi_i \neq 0} \alpha_i \leq 2 \Rightarrow \#\{i, \xi_i \neq 0\}\frac{\nu}{n} \leq 2$
so $\frac{1}{n}\#\{i, \xi_i \neq 0\} \leq \frac{2}{\nu}$ Q.E.D

ν-SVM

according to $(C_\nu 4)$ $0 \leq \alpha_i \leq \frac{\nu}{n}$

using $(C_\nu 3)$: $\sum_{i=1}^{n} \alpha_i = 2 \Rightarrow \sum_{i,\alpha_i \neq 0} \frac{\nu}{n} \geq 2 \Rightarrow \frac{1}{n}\#\{i, \alpha_i \neq 0\} \geq \frac{2}{\nu}$ Q.E.D

Remark: The ν-SVM enables to control the number of errors committed by the classifier through the parameter ν.

Theorem (admitted): B Schoelkopf, A Smola, R Williamson, P Bartlett

Under certain conditions of continuity on $P_{(X,Y)}$

- $\frac{1}{n}\#\{i, \xi_i \neq 0\} \longrightarrow \frac{2}{\nu}$ (convergence in probability)
- $\frac{1}{n}\#\{i, \alpha_i \neq 0\} \longrightarrow \frac{2}{\nu}$ (convergence in probability)

ν-SVM

Proposition: Relationship between $C-SVM$ and $\nu-SVM$

Let $w(\nu)$, $b(\nu)$,$\rho(\nu)$,$\xi(\nu)$ be the solutions of the ν-SVM (P_ν) with $\rho(\nu) \neq 0$, then:
$\frac{w(\nu)}{\rho(\nu)}$, $\frac{b(\nu)}{\rho(\nu)}$,$\frac{\xi(\nu)}{\rho(\nu)}$ are the solutions of the C-SVM (P_C) with $C = \frac{2\nu}{n\rho(\nu)}$.
As a consequence these two classifiers have the same decision boundaries.

Demonstration :

$$(P_\nu) \Leftrightarrow \begin{cases} \min\limits_{w,b,\rho,\xi_i} \frac{1}{2}\|w\|^2 - 2\rho + \frac{\nu}{n}\sum\limits_{i=1}^{i=n}\xi_i \\ y_i(\langle w, x_i\rangle + b) \geq \rho - \xi_i \\ \xi_i \geq 0 \end{cases}$$

First note that $\rho(\nu) = 0$ would correspond to a trivial solution for (P_ν) because in this case the function to minimize would always be positive and would then reach its minimum value of zero for the trivial solution $w = 0, b = 0, \xi = 0$. So we consider here $\nu - SVM$ for which $\rho(\nu) \neq 0$.

ν-SVM

If we assume now that (P_ν) has a non trivial solution (i.e $\rho(\nu) \neq 0$) then $w(\rho(\nu)), b(\rho(\nu)), \xi(\rho(\nu))$ are solutions of

$$\begin{cases} \min\limits_{w,b,\xi_i} \frac{1}{2}\|w\|^2 - 2\rho(\nu) + \frac{\nu}{n}\sum\limits_{i=1}^{i=n}\xi_i \\ y_i(\langle w, x_i\rangle + b) \geq \rho(\nu) - \xi_i \\ \xi_i \geq 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \min\limits_{w,b,\xi_i} \frac{1}{2}\|\frac{w}{\rho(\nu)}\|^2 - \frac{2}{\rho(\nu)} + \frac{\nu}{n\rho(\nu)}\sum\limits_{i=1}^{i=n}\frac{\xi_i}{\rho(\nu)} \\ y_i(\langle \frac{w}{\rho(\nu)}, x_i\rangle + \frac{b}{\rho(\nu)}) \geq 1 - \frac{\xi_i}{\rho(\nu)} \\ \frac{\xi_i}{\rho(\nu)} \geq 0 \end{cases}$$

So the arguments are the solutions of:

ν-SVM

$$\begin{cases} \min_{w,b,\xi_i} \| \frac{w}{\rho(\nu)} \|^2 + \frac{2\nu}{n\rho(\nu)} \sum_{i=1}^{i=n} \frac{\xi_i}{\rho(\nu)} \\ y_i(\langle \frac{w}{\rho(\nu)}, x_i \rangle + \frac{b}{\rho(\nu)}) \geq 1 - \frac{\xi_i}{\rho(\nu)} \\ \frac{\xi_i}{\rho(\nu)} \geq 0 \end{cases} \text{ which is a } C-\text{SVM with } C = \frac{2\nu}{n\rho(\nu)}$$

Q.E.D.

The decision boundaries for the ν- classifier are the hyperplanes: $H_{w(\nu),b(\nu)-\rho(\nu)}$ and $H_{-w(\nu),-b(\nu)-\rho(\nu)}$ and the decision boundaries for the C- classifier are : $H_{\frac{w(\nu)}{\rho(\nu)},\frac{b(\nu)}{\rho(\nu)}-1}$ and $H_{-\frac{w(\nu)}{\rho(\nu)},-\frac{b(\nu)}{\rho(\nu)}-1}$ which are the same hyperplanes. Q.E.D

Single Class SVM, Unsupervised Learning

Single Class SVM, Unsupervised Learning

Background :

- For a learning sample $(x_i)_{i\in[\![1,n]\!]}$ issued from a probability P_X we search a subset of $\mathbb{R}^d$ as "simple" and "small" as possible containing the $(x_i)_{i\in[\![1,n]\!]}$.
- The embedding is done after an immersion into $\mathbb{R}^{\mathbb{N}}$ via a function ϕ based on a Kernel K. Some points may be allowed to be misclassified (i.e left outside the domain) in $\mathbb{R}^{\mathbb{N}}$ but at a cost. In this case a trade-off is made between the size of the domain $\mathcal{D}_K$ chosen to embed the $\phi(x_i)$ and the measure of the errors of classification made.
- For a new observation z in $\mathbb{R}^d$ the hypothesis that z is issued from the probability distribution P_X will be accepted (with a certain confidence level) if z is in $\mathcal{D}$.

Single Class SVM, Unsupervised Learning

Remarks:

- $\mathcal{D}_K$ viewed from $\mathbb{R}^d$ as $\{x \in \mathbb{R}^d, \phi(x) \in \mathcal{D}_K\} = \mathcal{D}$ may appear as a single or several clusters of $\mathbb{R}^d$.
- As the sample $(x_i)_{i \in [\![1,n]\!]}$ consists here of unlabelled data, the problem of determining $\mathcal{D}$ is called unsupervised learning
- From now on we will use the Kernel $K_\sigma(x,y) = exp\left(-\frac{\|x-y\|_d^2}{2\sigma^2}\right)$ we note ϕ_σ the associated immersion and note $\mathcal{D}_\sigma$ the domain of $\mathbb{R}^d$ associated to $\mathcal{D}_{K_\sigma}$.

Single Class SVM: Clusterization without errors

We consider first for the sample $(x_i)_{i\in[\![1,n]\!]}$ of $\mathbb{R}^d$ the problem:

$$(U_\sigma) \Leftrightarrow \begin{cases} \min\limits_{w\in\mathbb{R}^{\mathbb{N}}} \|w\|^2 \\ \\ \forall i \in [\![1,n]\!], \langle w, \phi_\sigma(x_i)\rangle \geq 1 \end{cases}$$

Remarks:

- As mentioned previously, $\forall x \in R^d$, $\phi_\sigma(x) \in S^1_{\mathbb{N}}$ (the sphere of center 0 and radius 1 of $\mathbb{R}^{\mathbb{N}}$)
- (U_σ) has a domain of definition which is not empty because $w = \sum\limits_{j=1}^{j=n} \phi_\sigma(x_j)$ verifies

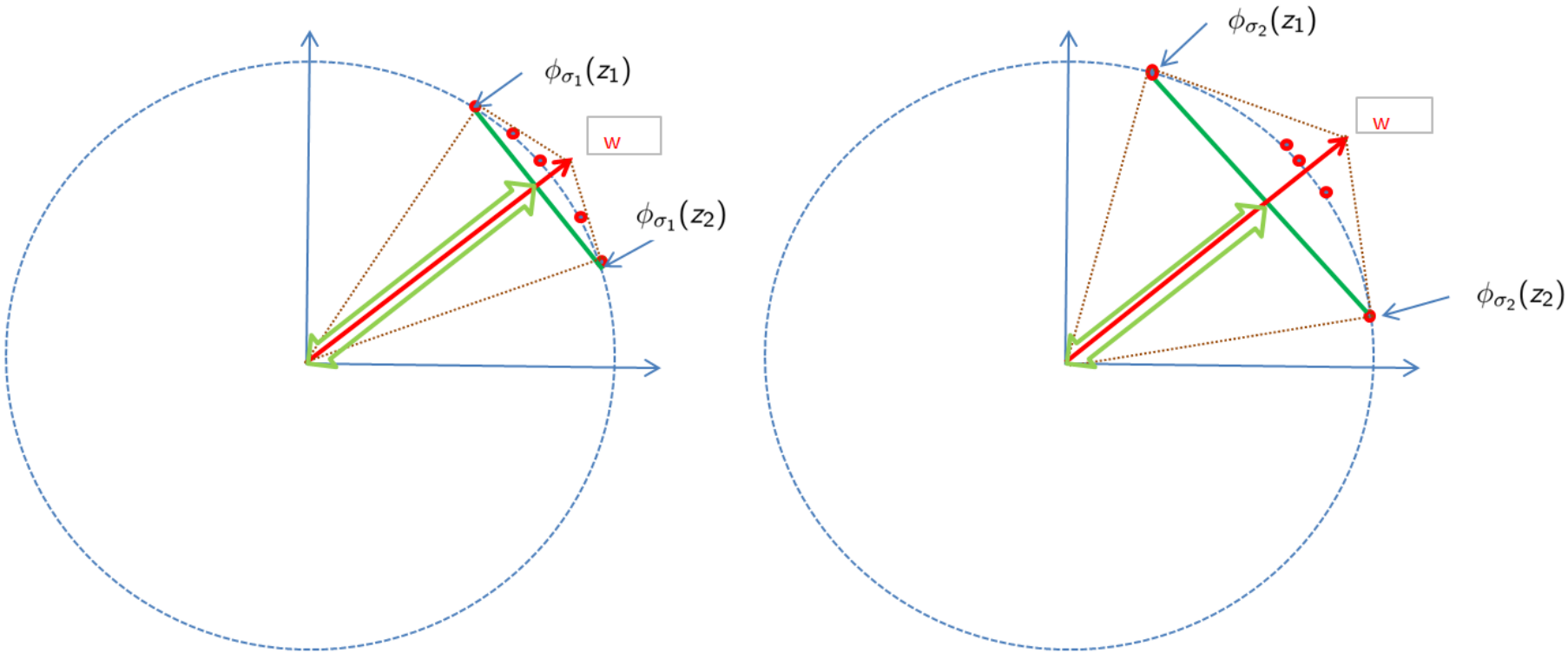
$$\langle w, \phi_\sigma(x_i)\rangle = \sum_{j=1}^{j=n} K_\sigma(x_j, x_i) = 1 + \sum_{j\neq i} K_\sigma(x_j, x_i) \geq 1$$

- Because of compacity and continuity the min for (U_σ) is attained. We note w_σ such a solution for (U_σ)

Single Class SVM: Clusterization without errors

- $\|w_\sigma\|_{\mathbb{N}} \geq 1$ because $1 \leq \langle w_\sigma, \phi_\sigma(x_i)\rangle \leq \|w_\sigma\|_{\mathbb{N}}\|\phi_\sigma(x_i)\|_{\mathbb{N}} = \|w_\sigma\|_{\mathbb{N}}$
- $\{\phi_\sigma(x), x \in \mathbb{R}^d, \langle w, \phi_\sigma(x_i)\rangle \geq 1\}$ are the points in the portion of the sphere delimited by $H_{w,-1}$
- The distance between the center of the sphere $S^1_{\mathbb{N}}$ and $H_{w_\sigma,-1}$ is $\frac{1}{\|w_\sigma\|_{\mathbb{N}}}$. By minimizing $\|w_\sigma\|_{\mathbb{N}}$ we minimize the portion of $S^1_{\mathbb{N}}$ delimited by $H_{w_\sigma,-1}$ which defines $\mathcal{D}_{K_\sigma}$
- σ defines the complexity of the model used and thus the complexity of the separation domain $\mathcal{D}_{K_\sigma}$. At σ fixed $\|w_\sigma\|_{\mathbb{N}}$ defines the size of the domain
- in the graph below we see w_σ and D_{K_σ} for various values of σ. Note that despite the fact that D_{K_σ} seems to increase when σ decreases, in fact D_σ decreases as σ decreases.

Single Class SVM: Clusterization without errors

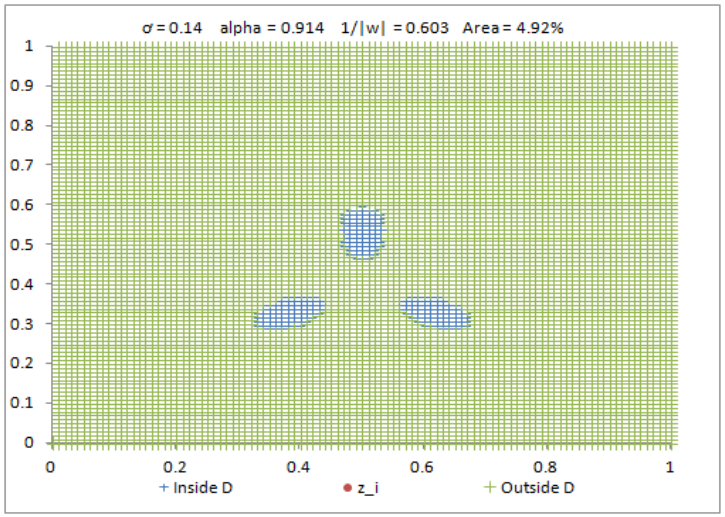


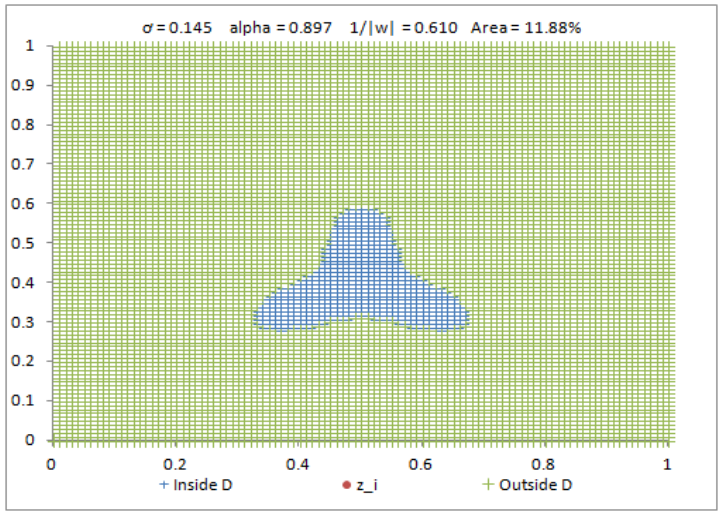
Hyperplane separating the points with maximum distance to the origin (delimiting the smallest portion of the sphere)

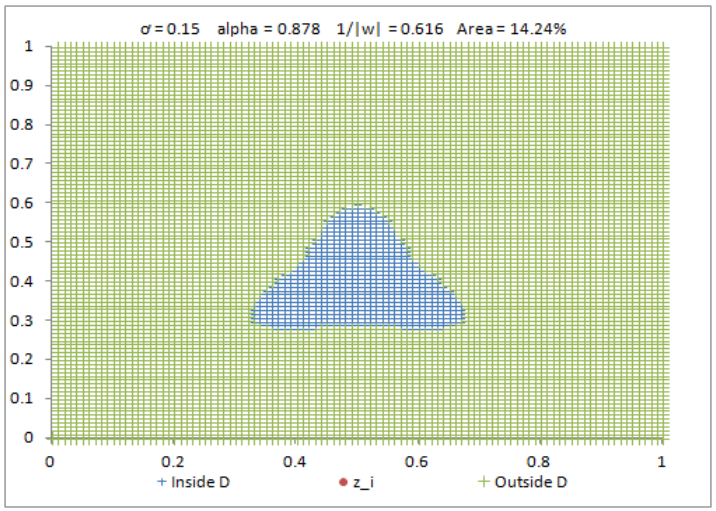
Single Class SVM: Clusterization without errors

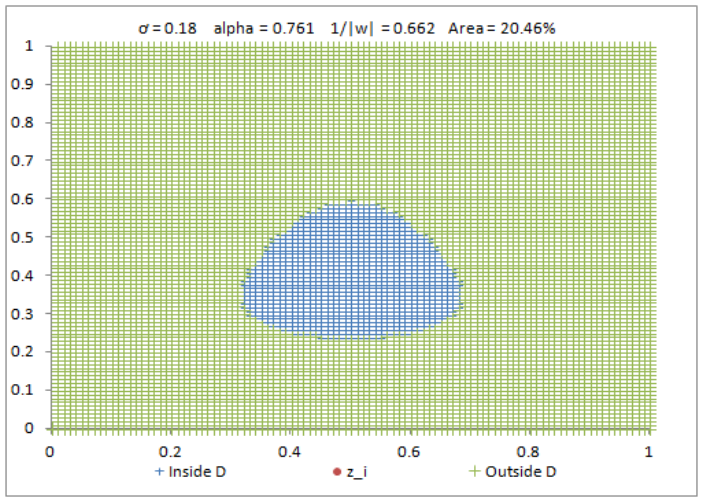
Example: we consider the points:
$x_1 = \begin{pmatrix} 0.327 \\ 0.3 \end{pmatrix}$ $x_2 = \begin{pmatrix} 0.673 \\ 0.3 \end{pmatrix}$ $x_3 = \begin{pmatrix} 0.5 \\ 0.6 \end{pmatrix}$ which form an equilateral triangle in $\mathbb{R}^2$ (with sides of lengths $d = 0.346$).
The problem is symmetric in $\mathbb{R}^{\mathbb{N}}$ as $\forall i \neq j$, $\langle \phi_\sigma(x_i), \phi_\sigma(x_j) \rangle = exp(\frac{-d}{2\sigma^2})$ and the solution w_σ of (U_σ) will be a linear combination of the $\phi_\sigma(x_i)$ with equal coefficients $\alpha(\sigma)$ and the three $\phi_\sigma(x_i)$ will be on the hyperplane.
So, $\sum_{j=1}^{j=3} \alpha(\sigma) K_\sigma(x_j, x_i) = 1$ and $\alpha(\sigma) = \left[1 + 2exp(\frac{-d}{2\sigma^2})\right]^{-1}$
and $\mathcal{D}_{K_\sigma} = \{x \in \mathbb{R}^2, \sum_{i=1}^{i=3} \alpha(\sigma) K_\sigma(x_i, x) \geq 1\}$
We plot below $\mathcal{D}_{K_\sigma}$ for various values of σ. The parameter σ defines the complexity of a class of domains and $\alpha(\sigma)$ defines within this class the domain of minimum size that contains the $\{x_i\}_{i \in [\![1,3]\!]}$.

Single Class SVM: Clusterization without errors









Single Class SVM: Clusterization without errors

Table: Size of the domain $\mathcal{D}_{K_\sigma}$ for various levels of complexity

σ (complexity)	$\alpha(\sigma)$	$\frac{1}{\|w\|_{\mathbb{N}}}$	$\lambda(\mathcal{D}_{K_\sigma})$ (size of the domain)
0.140	0.914	0.603	4.92%
0.145	0.897	0.610	11.88%
0.150	0.878	0.616	14.24%
0.180	0.761	0.662	20.46%

Remarks:

- as $\sigma \to 0$, the domain $\mathcal{D}_{K_\sigma}$ in $\mathbb{R}^d$ "converges" to the set formed by the sample points only, while the $\phi_\sigma(x)$ for all points of $\mathbb{R}^d$ get "orthogonalized" i.e verify $\forall x \neq y \langle \phi_\sigma(x) \phi_\sigma(y) \rangle \longrightarrow 0$

Single Class SVM: Clusterization without errors

- as $\sigma \to 0$, $w_\sigma \sim \sum_{i=1}^{i=n} \phi_\sigma(x_i)$ as the problem becomes symmetric and all the z_j from the sample verify $\langle w_\sigma, \phi_\sigma(x_j) \rangle \sim 1$ while any other point x in $\mathbb{R}^d$ satisfies $\langle w_\sigma, \phi_\sigma(x) \rangle \sim 0$
- as $\sigma \to 0$, $d(H_{w_\sigma,-1}, 0) \sim \frac{1}{\sqrt{n}}$ so it is not an achievement to be able to separate the $\phi_\sigma(x_i)$ by an hyperplane of distance only $\frac{1}{\sqrt{n}}$ because any random set of n points "sufficiently orthogonalized" could have been separated with the same distance to the origin
- generally the adequation of the model chosen will be tested by cross validation.

Single Class SVM: Clusterization with errors

We consider now the problem:

$$(U_{\sigma,\nu}) \begin{cases} \min\limits_{w\in\mathbb{R}^{\mathbb{N}},\rho,\xi_i} \frac{1}{2}\|w\|^2 - 2\rho + \frac{\nu}{n}\sum\limits_{i=1}^{i=n}\xi_i \\ \langle w, \phi_\sigma(x_i)\rangle \geq \rho - \xi_i \\ \xi_i \geq 0 \end{cases}$$

which is the extension of the previous clustering problem but this time with some errors ξ_i allowed in the classification. A ν formulation has been chosen instead of a C-formulation in order to have a better-interpretability of the parameters.

Single Class SVM: Clusterization with errors

The dual problem can be calculated as in the ν-SVM section and we obtain

Proposition: Dual Problem for ν-SVM

$$(U_{\sigma,\nu}) \Leftrightarrow (U^*_{\sigma,\nu}) \text{ where } (U^*_{\sigma,\nu}) \begin{cases} -\frac{1}{2} \min_{\alpha_i} \sum_{i=1}^{i=n} \sum_{j=1}^{j=n} \alpha_i \alpha_j k(x_i, x_j) \\ \sum_{i=1}^{i=n} \alpha_i = 1 \\ 0 \leq \alpha_i \leq \nu \end{cases}$$

with: $w^* = \sum_{i=1}^{i=n} \alpha_i^* \phi_\sigma(x_i)$

Single Class SVM: Alternative Geometric Approach

Consider the problem

$$(B_{\sigma,\nu}) \begin{cases} \min\limits_{R\in\mathbb{R},c\in\mathbb{R}^{\mathbb{N}},\xi\in\mathbb{R}^n} R^2 + \frac{\nu}{n}\sum\limits_{i=1}^{i=n}\xi_i \\ \|\phi(x_i) - c\|_{\mathbb{N}}^2 \leq R^2 + \xi_i \\ \xi_i \geq 0 \end{cases}$$

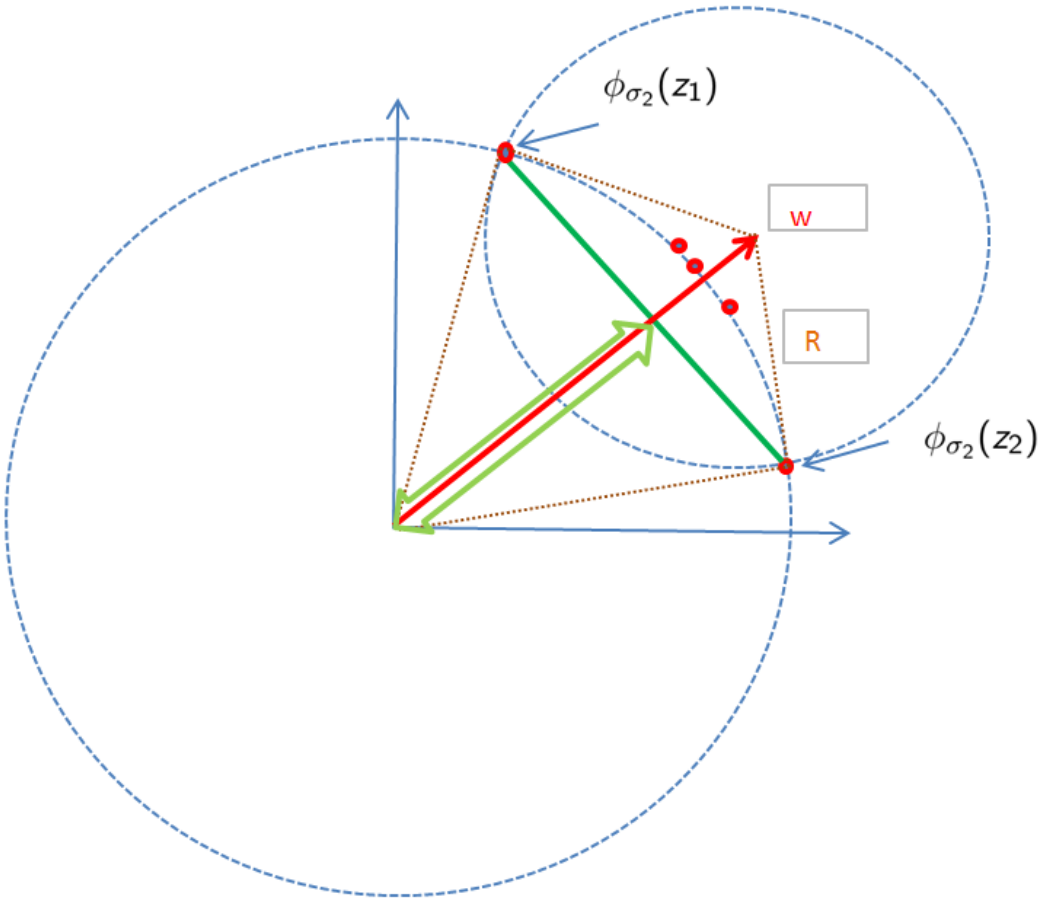
where we search the ball in $R^{\mathbb{N}}$ of minimum radius which contains the $\phi(x_i)$. The cluster in $\mathbb{R}^d$ will be defined as the points whose images $\phi(x)$ belongs to this ball. Some errors are permitted (some points from the sample are left outside the domain) in order to minimize the radius.

Proposition

$(B_{\sigma,\nu})$ and $(U_{\sigma,\nu})$ have the same dual (so they are two different formulations of the same approach) and $c^* = w^*$

Demonstration: calculate the dual of $(B_{\sigma,\nu})$ derived from the Lagrangien

Single Class SVM: Alternative Geometric Approach



Equivalent geometric approaches for clusterization

References

Bernhard Schölkopf, Alex J.Smola, Robert C.Williamson, Peter L.Bartlett

New Support Vector Algorithms

Neural Computation 12, 2000, pp.1207-1245

David J Crisp, Christopher J.C Burges

A Geometric Interpretation of ν-SVM Classifiers

NIPS Conference, 1999

Christopher J.C Burges

A Tutorial on Support Vector Machines for Pattern Recognition

Data Mining and Knowledge Discovery 2, 1998 pp. 121-167

Don Hush, Clint Scovel

On the VC Dimension of Bounded Margin Classifiers

Machine Learning Volume 45 Issue 1, October 1 2001 pp. 33 - 44

References

Vladimir N.Vapnik

An Overview of Statistical Learning Theory

IEEE Transactions on Neural Networks, vol 10, No 5 September 1999

P H Chen, C J Lin, B Schölkopf

A Tutorial on ν-Support Vector Machines

Applied Stochastic Models in Business and Industry, No 21, 2005 pp. 111 - 136

Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola

Estimating the Support of a High-Dimensional Distribution

Neural Computation, No 13, 2001 pp. 1443 - 1471