Data Mining - SVM

Dr. Jean-Michel RICHER

université angers

FACULTÉ DES SCIENCES
Unité de formation et de recherche
DÉPARTEMENT INFORMATIQUE

2018

jean-michel.richer@univ-angers.fr

Outline

1. Introduction

What we will cover

- remainder of linear regression
- Support Vector Machine
  - principle
  - primal and dual formulation
  - soft margin and slack variables
  - kernel
  - iris dataset in python

2. Linear regression - a remainder

Principle

Given a set of individuals $X = \{x_1, x_2, \ldots, x_n\}$ where each $x_i = (x_i^1, \ldots, x_i^d)$ and a vector y of size n:

- find a function $f(x) = w.x + w_0 = y + \epsilon$ such that the y_i are close to $f(x)$

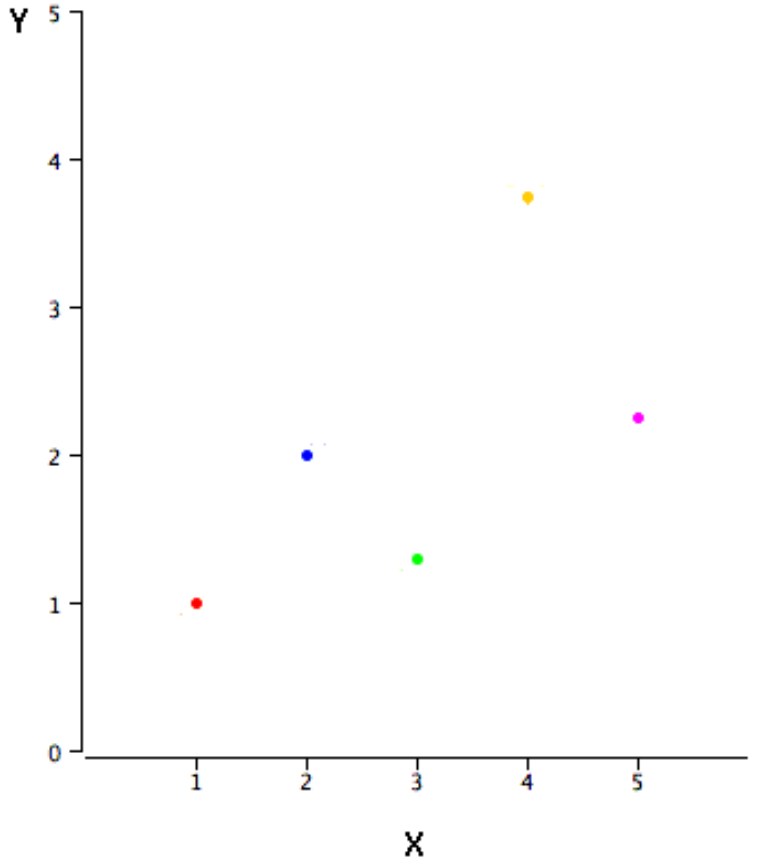
X			y	$f(x_i)$
x_1^1	$\ldots$	x_1^d	y_1	y'_i
$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
x_n^1	$\ldots$	x_n^d	y_n	y'_n

input data

Linear regression principle - Example

Consider the following example where $n = 5$ and $d = 1$:

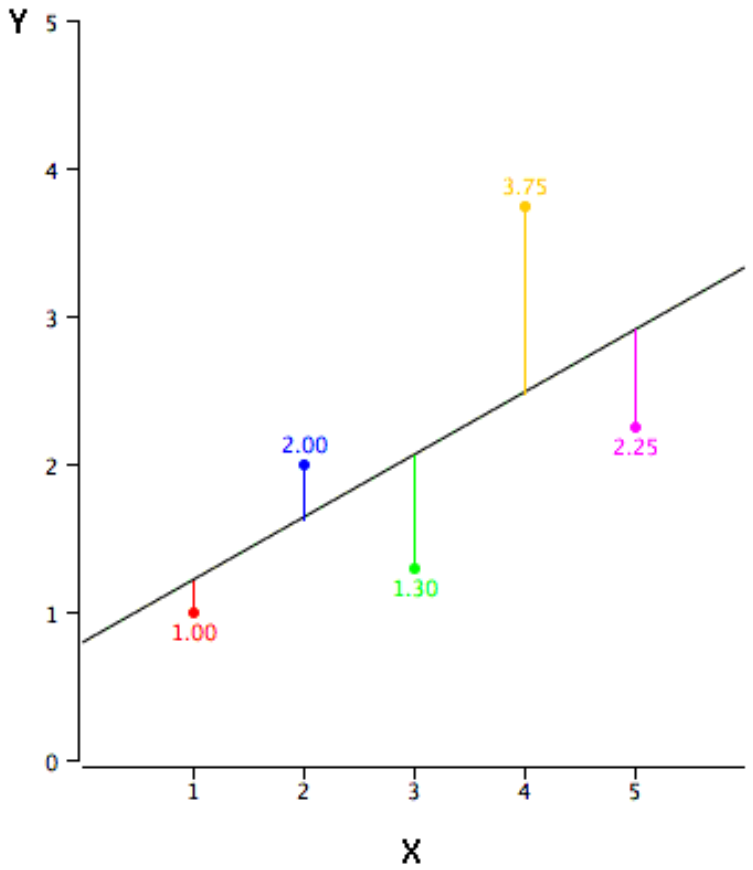
X	y
1.00	1.00
2.00	2.00
3.00	1.30
4.00	3.75
5.00	2.25



Linear regression principle - Example

Consider the following example where $n = 5$ and $d = 1$:

In linear regression, the observations (color points) are assumed to be the result of random deviations (green) from an underlying relationship (black)



Formula

Ordinary least squares (OLS) is the simplest and most common estimator that minimizes the sum of squared residuals

$$RSS = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

we can obtain w:

$$w = (X^T.X)^{-1})X^T y = (\sum (x_i x_i^T))^{-1} (\sum x_i y_i)$$

Linear regression principle

Formula

If $d = 1$, given

- $\overline{X}$ the mean of X, $\overline{y}$ the mean of y,
- σ_X, σ_Y the standard deviation of X and y
- R the correlation between X and Y

we can compute

- the **slope** $w_1 = R \times \frac{\sigma_Y}{\sigma_X}$
- the **intercept** $w_0 = \overline{y} - w_1 \times \overline{X}$

Linear regression - slope and intercept

With our example:

$\overline{X}$	$\bar{y}$	σ_X	σ_Y	R
3	2.06	1.581	1.072	0.627

- $w_1 = (0.627)(1.072)/1.581 = 0.425$
- $w_0 = 2.06 - (0.425)(3) = 0.785$

Linear regression - prediction

$$y' = f(x_i) = 0.425 \times x_i + 0.785$$

X	y	y'	$y - y'$	$(y - y')^2$
1.00	1.00	1.210	-0.210	0.044
2.00	2.00	1.635	0.365	0.133
3.00	1.30	2.060	-0.760	0.578
4.00	3.75	2.485	1.265	1.600
5.00	2.25	2.910	-0.660	0.436

So the total sum of errors is 2.791

Linear regression - but

So linear regression is good, but, see Anscombe's quartet

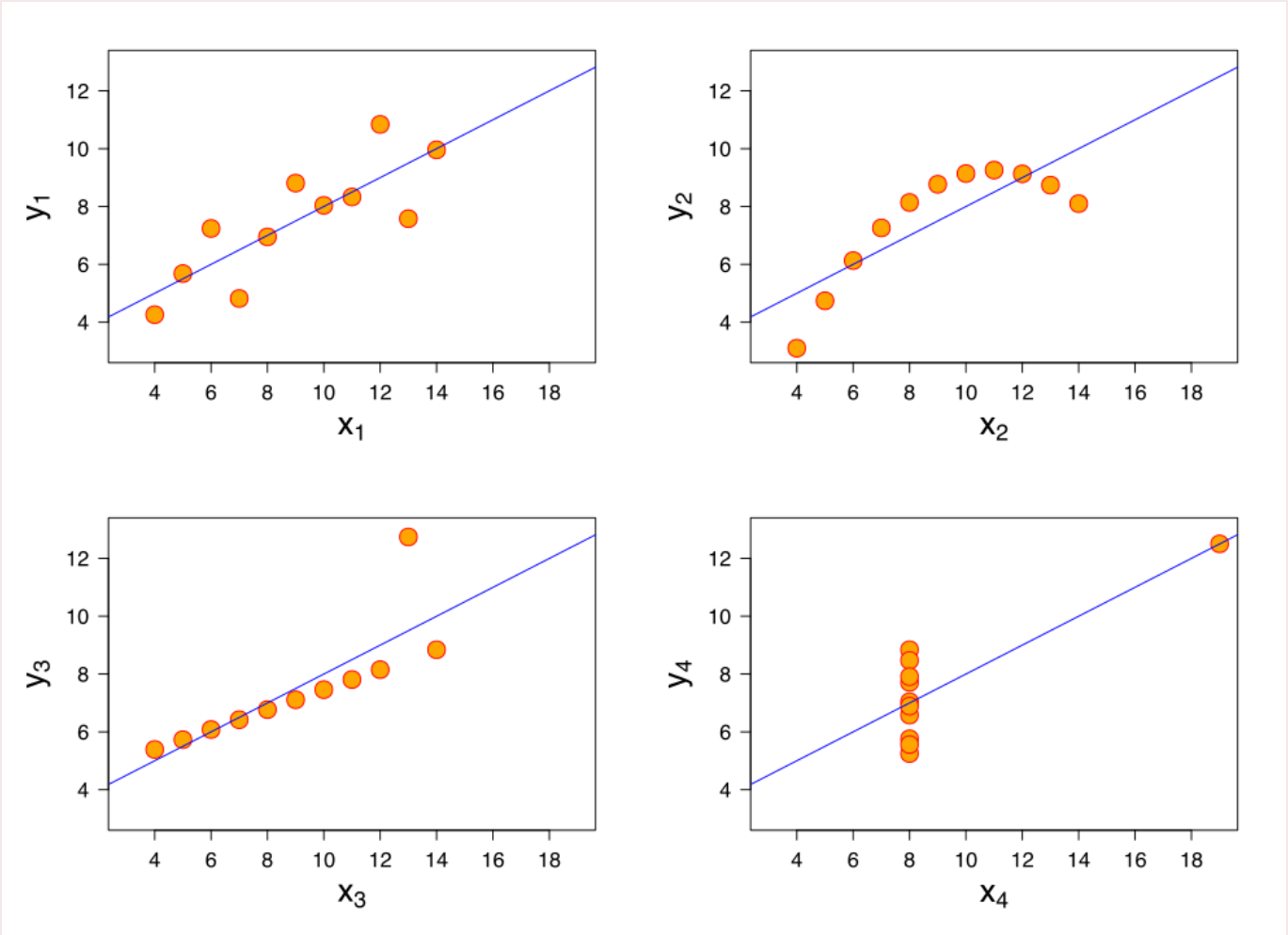
Anscombe's quartet (Francis Anscombe, 1973)

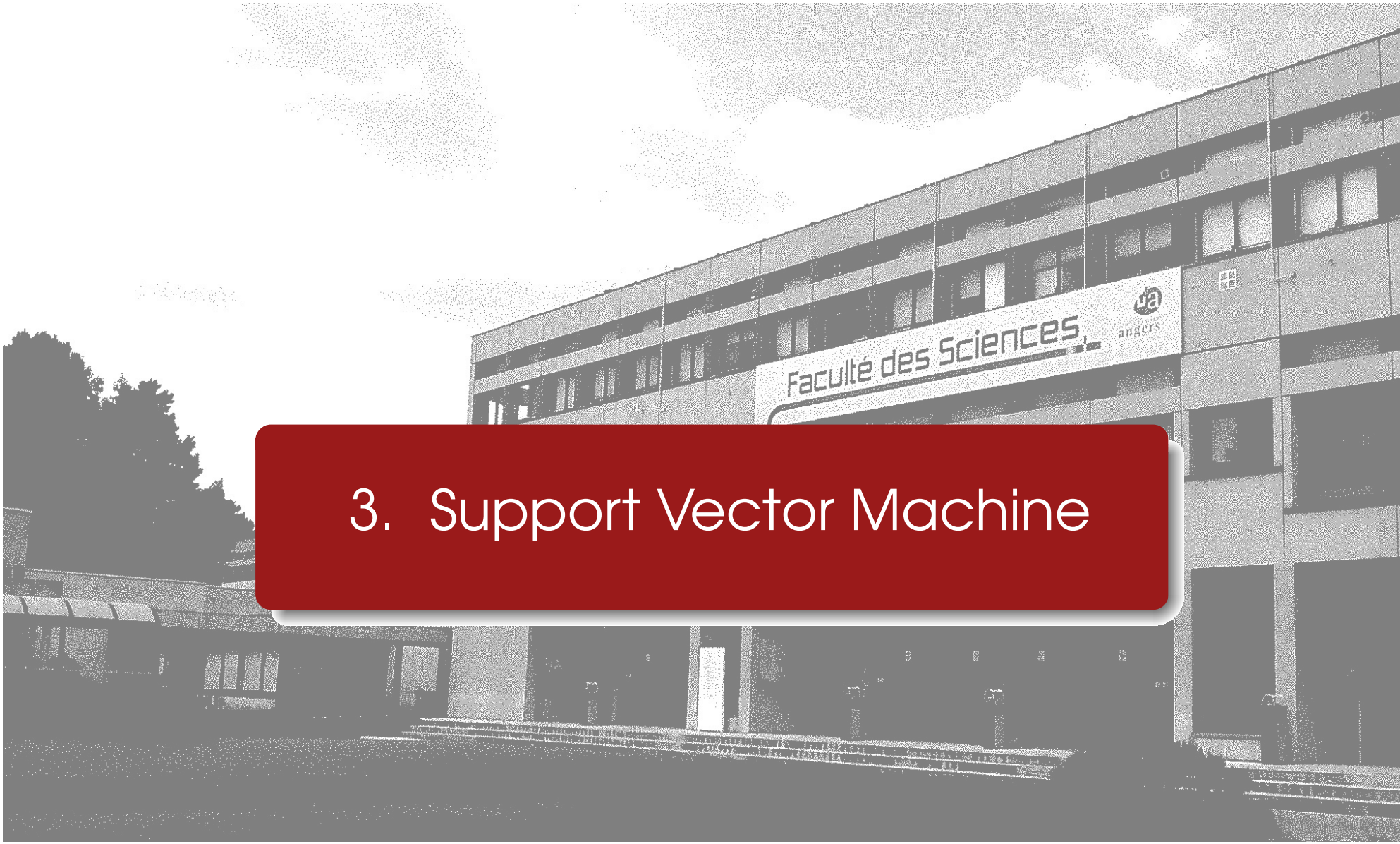
four datasets that have nearly identical simple descriptive statistics, yet **appear very different when graphed**

Mean of x	9
variance of x	11
Mean of y	7.50
variance of y	4.125
Correlation	0.816
Linear regression line	$y = 3.00 + 0.500x$
Coefficient of determination	0.67

often used to illustrate the importance of looking at a set of data graphically before starting to analyze it

Anscombe's quartet





3. Support Vector Machine

SVM Principle

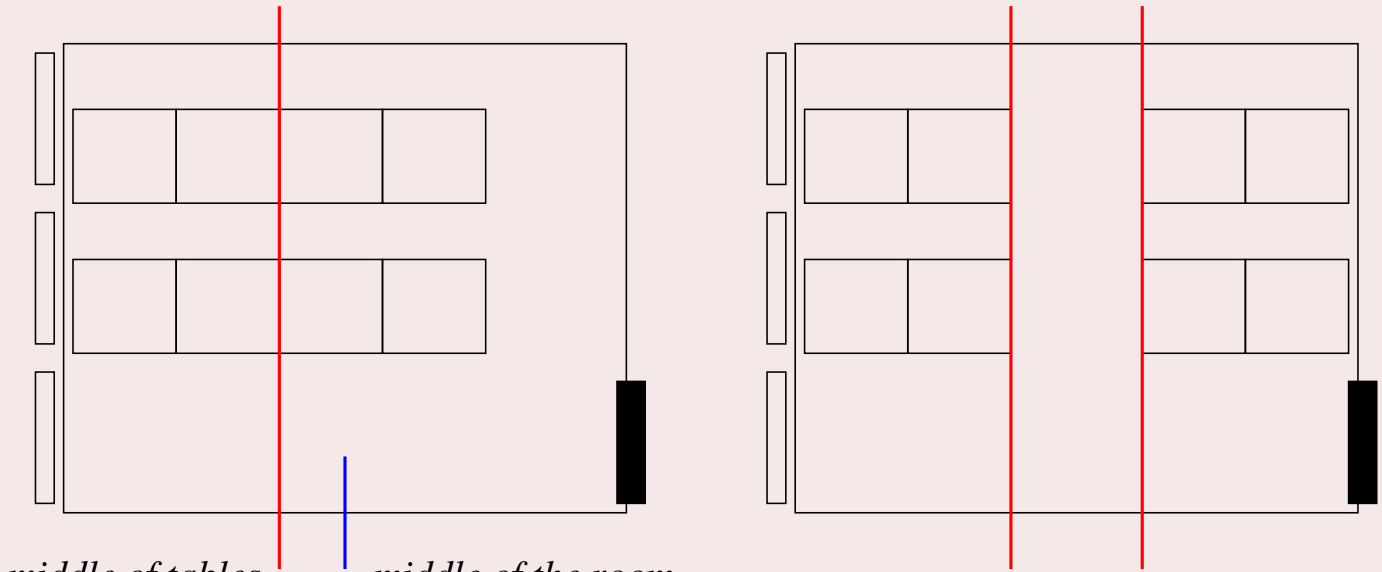
SVM are called **large-margin** classifier:

- separate two classes or more
- they look for an hyperplane with a large (wide) margin in order to make it:
  - less sensitive to perturbations
  - more stable for prediction
- the initial primal formulation is transformed into a dual formulation easier to solve
- **slack variables** help for misclassified elements
- non linear examples can be transformed into linear examples using a **kernel** function

Large margin principle

Separate students in a classroom:

- class 1: students close to the window
- class 2: students close to the door



Principle

Given a set of individuals $X = \{x_1, x_2, \ldots, x_n\}$ where each $x_i = (x_i^1, \ldots, x_i^d)$ and a vector y of size n of classes $\{-1, +1\}$:

- find a function $f(x) = w.x + w_0$ such that

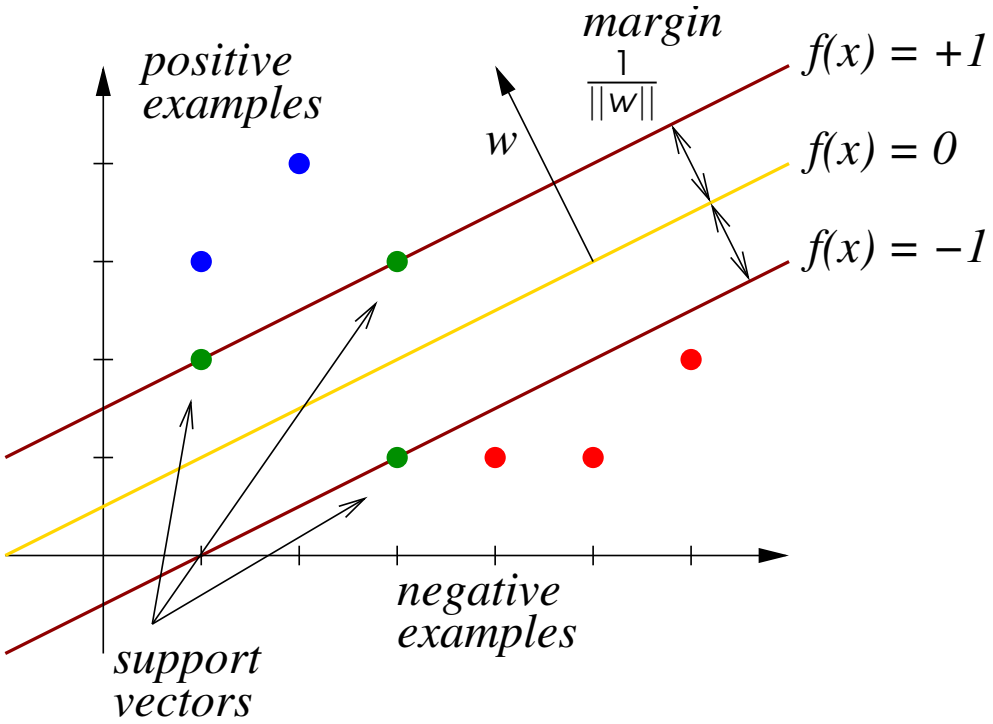
$$\begin{cases} \text{if } y_i = +1, f(x_i) \geq +1 \\ \text{if } y_i = -1, f(x_i) \leq -1 \end{cases}$$

- or simply:

$$y_i \times f(x_i) \geq 1$$

$f(x) = 0$ is the hyperplane (or separation line)

SVM - Graphically explained



$$d(x_i) = \frac{w\,x_i + w_0}{||w||}$$

Important !

Support Vectors

- individuals (objects) closest to the separating hyperplane
- very few compared to n

Aim of SVM

orientate the hyperplane in order it is as far as possible from the **SV** of both classes

Formulation

$$\begin{cases} Max & \frac{2}{||w||} \\ \text{subject to} & y_i(w\,x_i + w_0) \geq 1,\ \forall i \end{cases} \tag{1}$$

or

$$\begin{cases} Min & \frac{1}{2}||w||^2 \\ \text{subject to} & y_i(w\,x_i + w_0) - 1 \geq 0,\ \forall i \end{cases} \tag{2}$$

Minimizing $||w||$ is equivalent to minimizing $1/2||w||^2$ but in this case we can perform **Quadratic Programming** (QP) optimization

SVM on a mathematical point of view

Quadratic programming (QP)

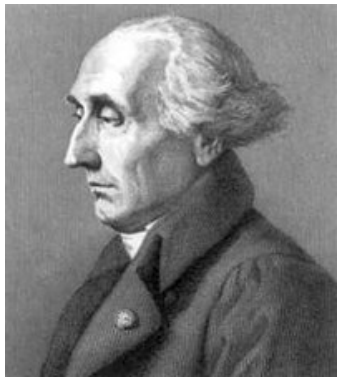
- is the process of solving a special type of mathematical optimization problem
- optimize (minimize or maximize) a quadratic function (x^2)
- subject to linear constraints

SVM on a mathematical point of view

how to solve the problem ?

- the problem of finding the optimal hyper plane is an optimization problem and can be solved by optimization techniques
- the problem is difficult to solve this way but we can rewrite it
- we use Lagrange multipliers to get this problem into a form that can be solved analytically.

Joseph-Louis Lagrange



- born **Guiseppe Lodovico Lagrangia** (1736 - 1813) was a franco-italian mathematician and astronomer
- made significant contributions to the fields of analysis, number theory, and both classical and celestial mechanics
- in **1787**, at age 51, moved from Berlin to Paris and became a member of the **French Academy of Sciences**
- remained in France until the end of his life

Lagrangian multipliers

Principle

- you want to minimize or maximize $f(x)$ subject to $g(x) = 0$
- under certain conditions (quadratic, linear constraints)
- define the function

$$\mathcal{L}(x, \alpha) = f(x) + \alpha g(x)$$

where $\alpha \geq 0 \ \in \mathbb{R}$ is called the **lagrangian multiplier**

Resolution

- a solution of $\mathcal{L}(x, \alpha)$ is a point of gradient 0
- so compute and solve

$$\frac{\partial \mathcal{L}(x,\alpha)}{\partial x} = 0$$

$$\frac{\partial \mathcal{L}(x,\alpha)}{\partial \alpha} = g(x) = 0$$

- or reuse in $\mathcal{L}(x, \alpha)$

Statement of the example

Suppose you want to put a fence around some field which as a form of a rectangle (x, y) and you want to maximize the area knowing that you have P meters of fence:

$$\begin{cases} Max & x \times y \\ \text{such that} & P = 2x + 2y \end{cases}$$

then $f(x, y) = xy$ and $g(x) = P - 2x - 2y = 0$

$$\begin{cases} Max & xy \\ \text{such that} & P - 2x - 2y = 0 \end{cases}$$

Lagrange formulation

$$\mathcal{L}(x, y, \alpha) = xy + \alpha(P - 2x - 2y)$$

the derivatives give us

$$\frac{\partial \mathcal{L}(x, y, \alpha)}{\partial x} = y - 2\alpha = 0$$

$$\frac{\partial \mathcal{L}(x, y, \alpha)}{\partial y} = x - 2\alpha = 0$$

$$\frac{\partial \mathcal{L}(x, y, \alpha)}{\partial \alpha} = P - 2x - 2y = 0$$

Method of Lagrange - example

Resolution

- the first two constraints give us $y = 2\alpha = x$, so $x = y$
- in other words, the area is a square
- and the last one that $P = 4x = 4y$
- consequently $\alpha = P/8$ because $\alpha = x/2 = y/2$

Back to the SVM formulation

Primal formulation

$$\begin{array}{rl} \textit{Min} & \frac{1}{2}||w||^2 \\ \text{such that} & y_i(wx_i + w_0) - 1 \geq 0, \ \forall i \end{array}$$

Lagrangian of the primal

$$\mathcal{L}_P(w, w_0, \alpha) = \frac{1}{2}||w||^2 - \alpha \left[\sum_i y_i(wx_i + w_0) - 1 \right]$$

with $\alpha \geq 0$

Lagrangian of the primal

$$\mathcal{L}_P(w, w_0, \alpha) = \frac{1}{2}||w||^2 - \alpha \left[\sum_i y_i(wx_i + w_0) - 1 \right]$$

$$\mathcal{L}_P(w, w_0, \alpha) = \frac{1}{2}||w||^2 - \sum_j \alpha_j \left[\sum_i y_i(wx_i + w_0) - 1 \right]$$

$$\mathcal{L}_P(w, w_0, \alpha) = \frac{1}{2}||w||^2 - \sum_i \sum_j \alpha_i y_i(wx_i + w_0) + \sum_j^n \alpha_j \quad (3)$$

Lagrangian of the primal

we want to find w and w_0 which minimizes and α that maximizes (3):

$$\frac{\partial \mathcal{L}_P(w, w_0, \alpha)}{\partial w} = w - \sum_i \alpha_i y_i x_i = 0 \tag{4}$$

$$\frac{\partial \mathcal{L}_P(w, w_0, \alpha)}{\partial w_0} = \sum_i \alpha_i y_i = 0 \tag{5}$$

Dual formulation

By substituting (4) and (5) into (3), we get:

$$\begin{cases} \underset{\alpha}{Max} & \mathcal{L}_D(\alpha) = \sum_j \alpha_j - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i x_j \\ \text{such that} & \alpha_i \geq 0, \ \forall i \\ & \sum_i \alpha_i y_i = 0 \end{cases}$$

or

$$\begin{cases} \underset{\alpha}{Max} & \mathcal{L}_D(\alpha) = \sum_j \alpha_j - \frac{1}{2} \sum_{i,j} \alpha_i H_{ij} \alpha_j \\ \text{such that} & \alpha_i \geq 0, \ \forall i \\ & \sum_i \alpha_i y_i = 0 \\ \text{with} & H_{ij} = y_i y_j x_i x_j \end{cases} \tag{6}$$

Need only the dot product of x

Dual resolution

- from (6) with a QP solver we can find α and thus w
- w_0 can be obtained from the Support Vectors

$$w_0 = y_s - \sum_{m \in S} \alpha_m y_m x_m x_s$$

where S is the set of indices of SV, such that $\alpha_i > 0$

Extension of SVMs

What if

- the data are not classified to be linearly separable ?
  - we use **slack variables**
- we don't have a linear support ?
  - we use a **kernel** function

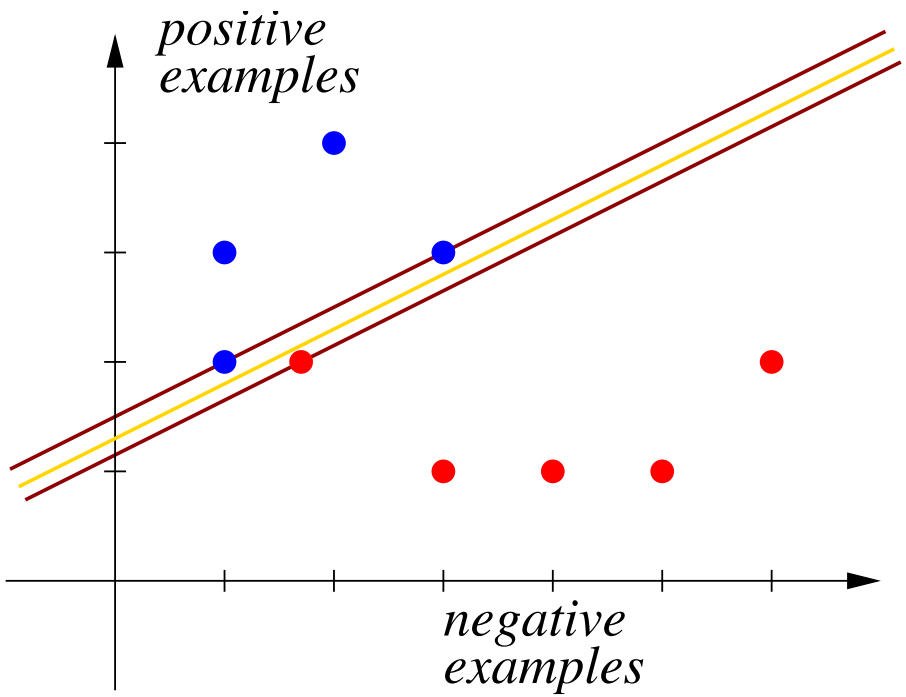
Slack variable

Not fully linearly separable

- we need to relax the constraints
- introduce slack variables to keep a wide margin

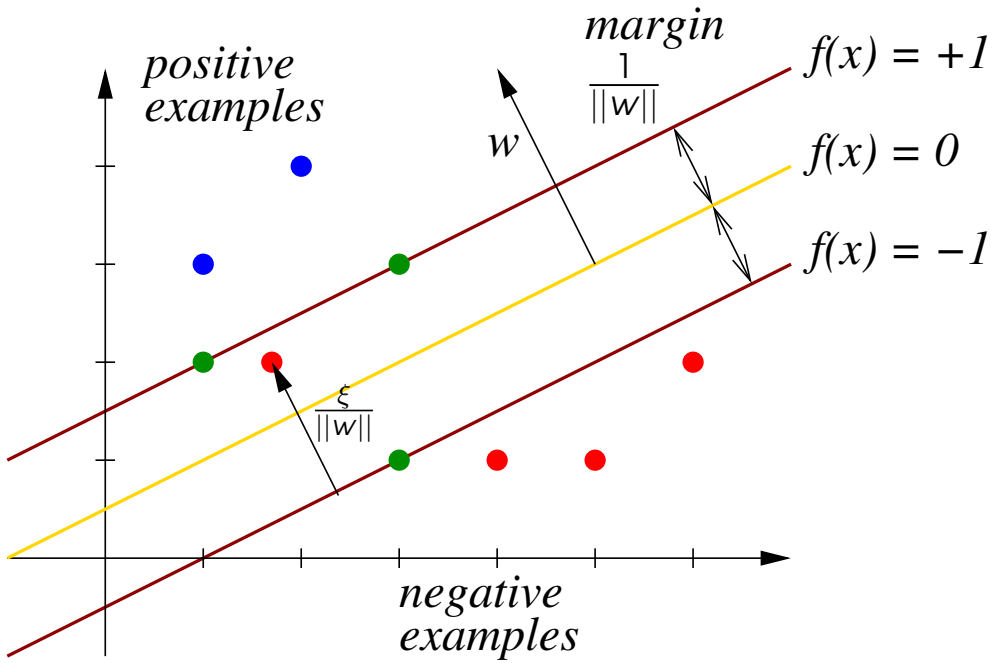
$$\begin{cases} w\,x_i + w_0 \geq +1 - \xi_i & \text{for } y_i = +1 \\ w\,x_i + w_0 \leq -1 + \xi_i & \text{for } y_i = -1 \\ \xi_i \geq 0,\ \forall i \end{cases}$$

Slack variable



Violates the large margin principle !

Slack variable



Formulation with slack variables

We have a **soft margin**:

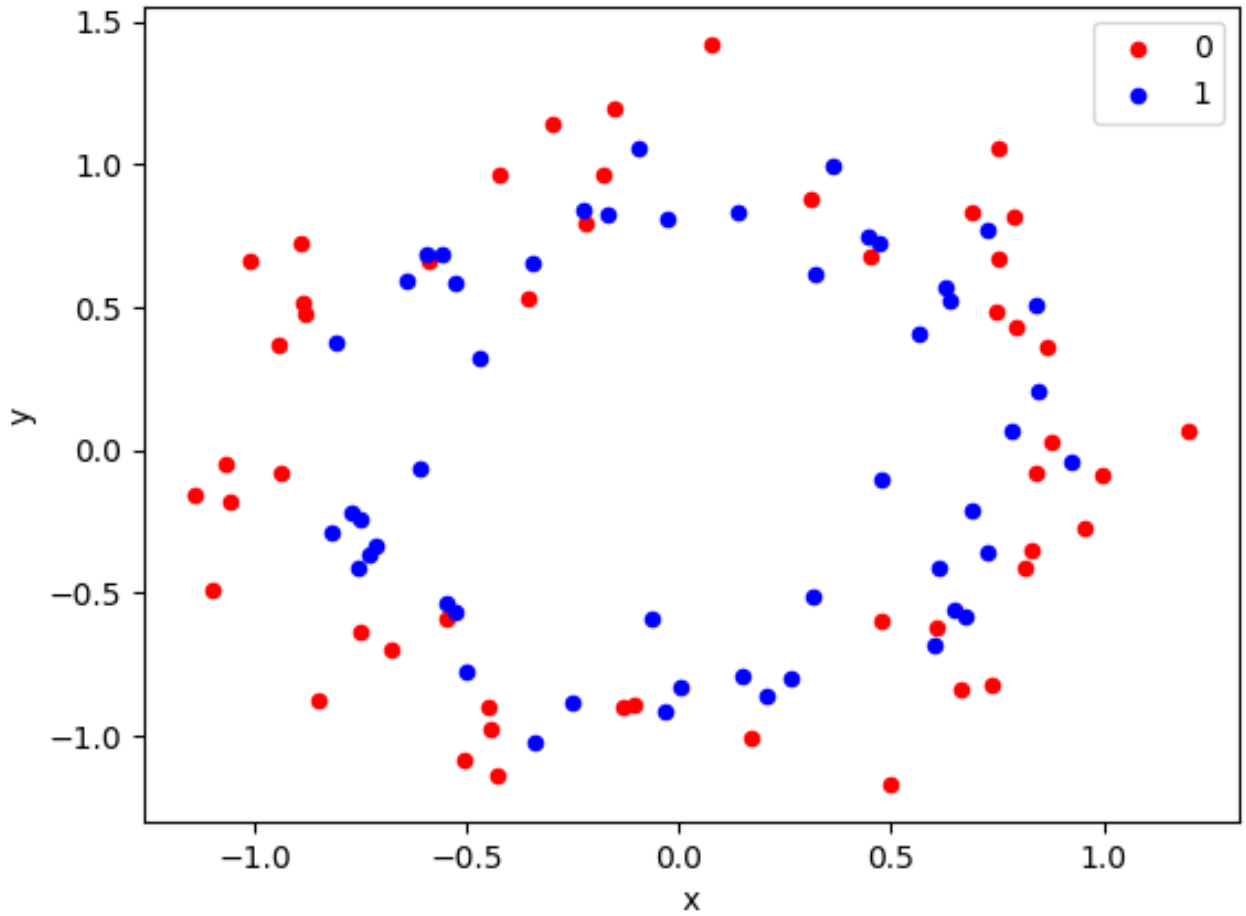
$$\begin{cases} Min \quad \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad y_i(w\,x_i + w_0) - 1 + \xi_i \geq 0, \ \forall i \end{cases} \tag{7}$$

The parameter C controls the trade-off between the slack variable penalty and the size of the margin

Why a kernel ?

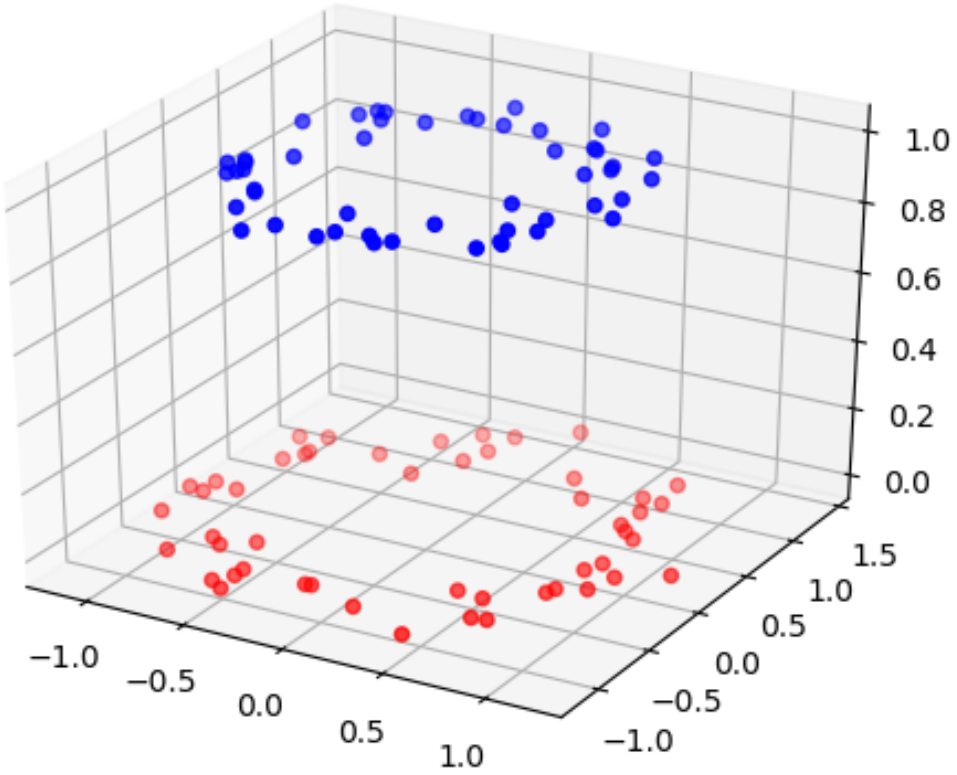
- Many problems are not linearly separable in the space of the input X
- $k(x_i, x_j) = \phi(x_i)\phi(x_j)$

A problem where individuals are not linearly separable in X with $d = 2$



Kernel

But individuals are linearly separable in $X + y$ (consider that 0 on z represents the class of the negative individuals)



or use polar coordinates (r, θ)

SVM in Python

Import svm from sklearn package:

```
from sklearn import svm

svc = svm.SVC(kernel='linear', C=1,
      gamma='auto').fit(X, y)
```

You can choose different kernels : linear, poly, rbf, sigmoid

Kernel functions

- linear: $\langle x, x' \rangle$
- polynomial: $(\gamma \langle x, x' \rangle + r)^d$
  - d is specified by keyword `degree`
  - and r by `coef0`
- rbf (Radial Basis Function): $exp(-\gamma ||x - x'||^2)$
  - γ is specified by keyword `gamma`, must be greater than 0.
- sigmoid: $tanh(\gamma \langle x, x' \rangle + r))$
  - where r is specified by `coef0`

7. Example

Example in python

IRIS

- we use the IRIS example of sklearn
- 150 individuals (50 Setosa, 50 Versicolour, 50 Virginica)
- use the SVC or SVR implementation

Example in python

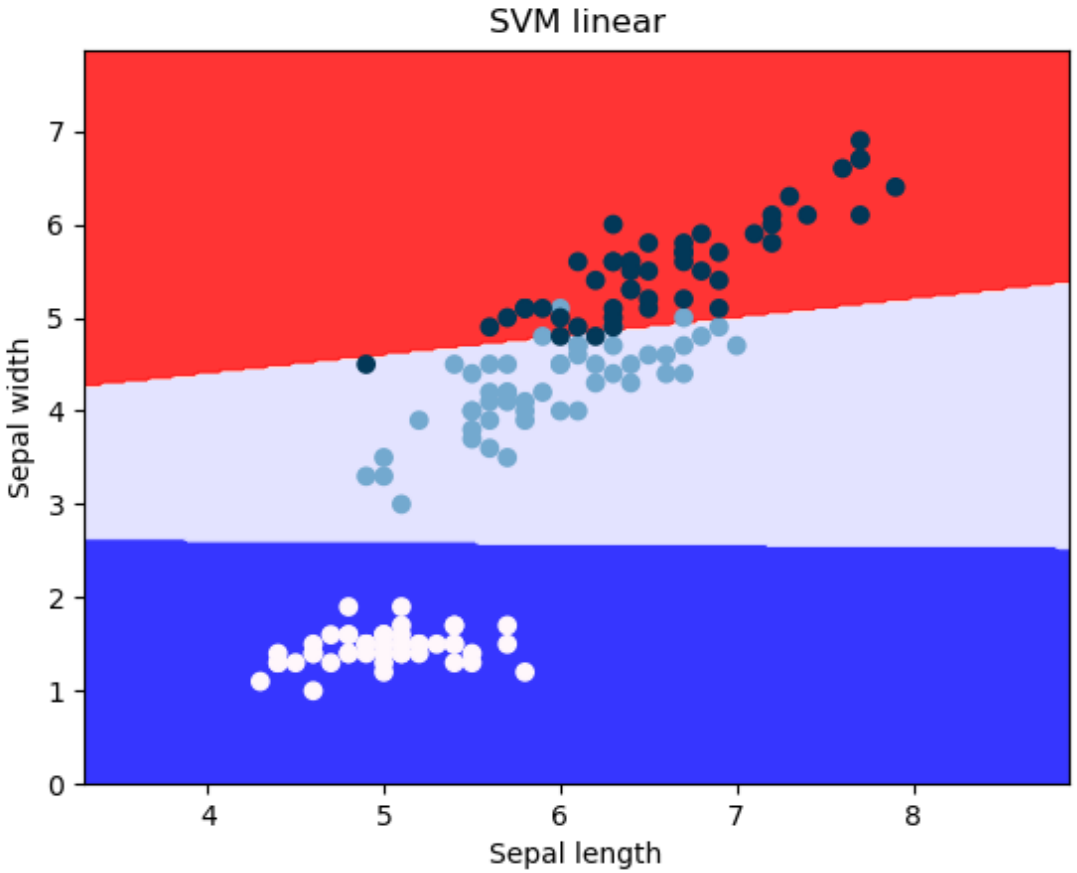
IRIS

```
from sklearn import svm

svc = svm.SVC(kernel='linear', C=1, gamma='auto')
# svc = svm.SVC(kernel='rbf', C=100, gamma=100)
# svc = svm.SVR(kernel='rbf', C=1, gamma=1)
svc.fit(X, y)
y_predict = svc.predict(X)
```

Example in python - Linear SVM

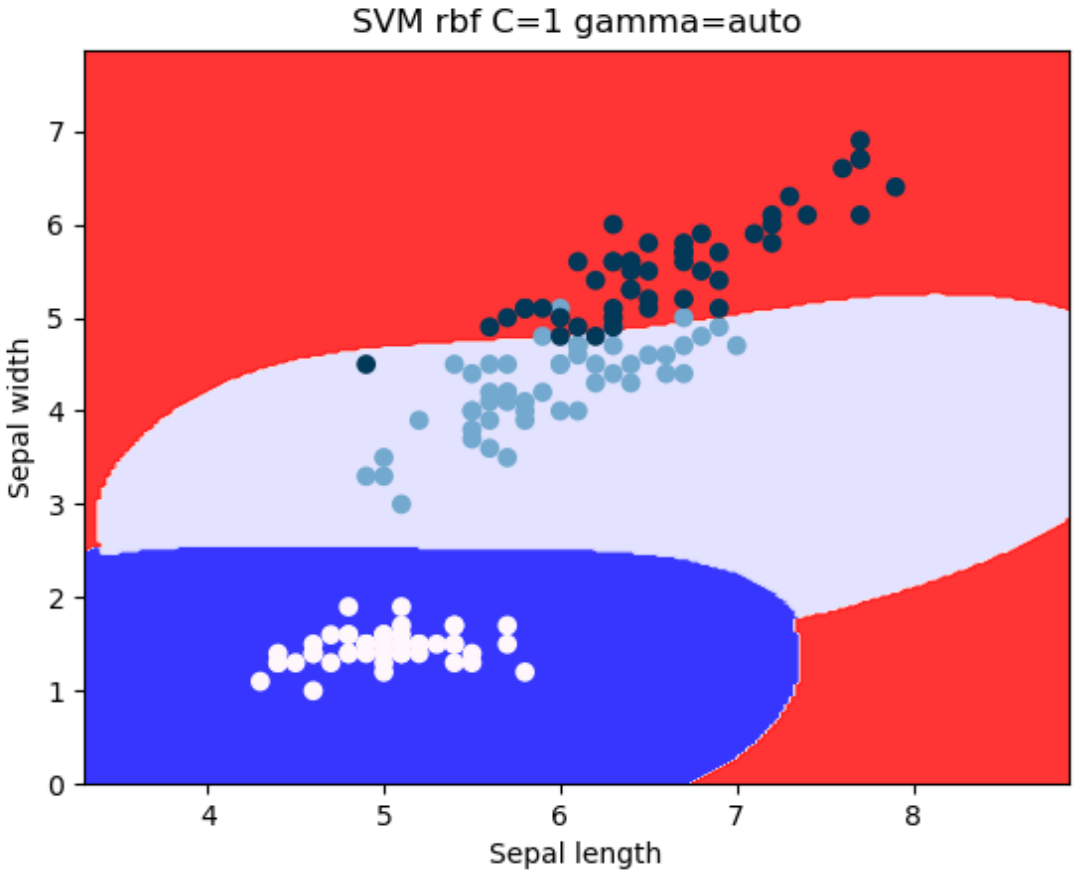
Linear svm, $C = 1$



7 mispredicted

Example in python - RBF SVM

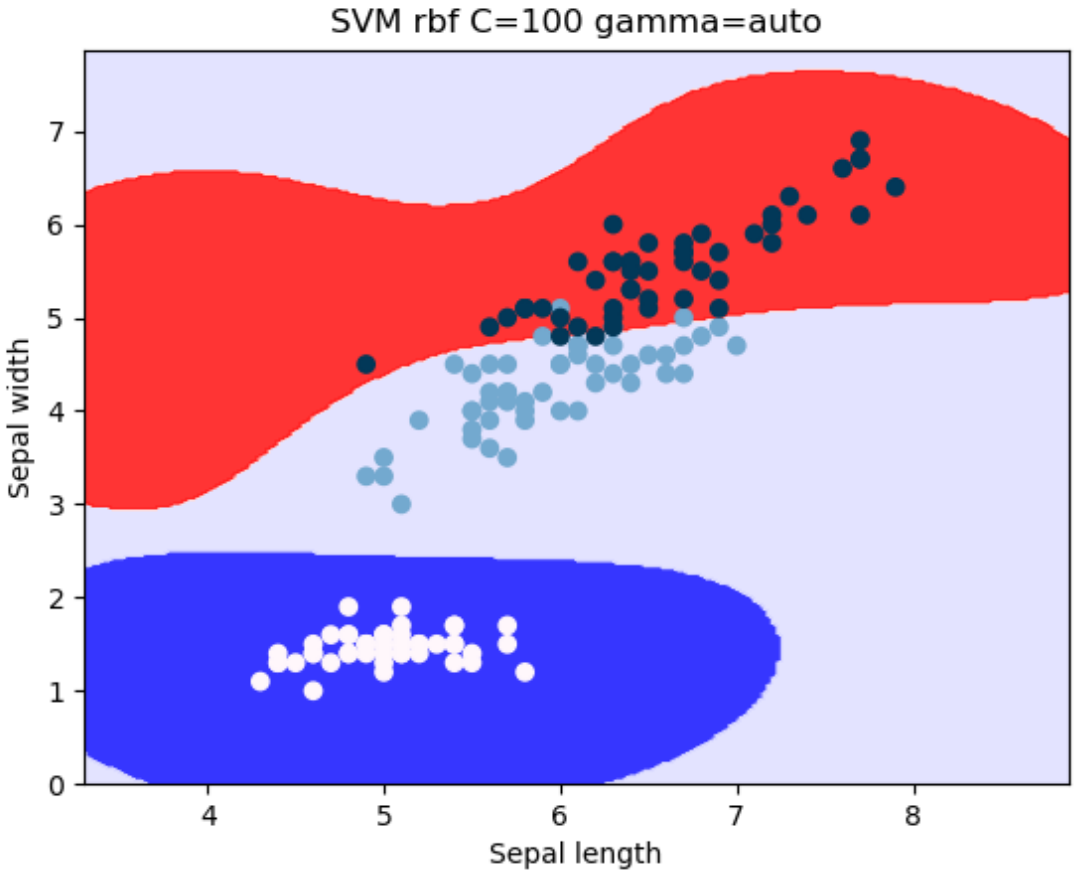
RBF $C = 1$



6 mispredicted

Example in python - RBF SVM

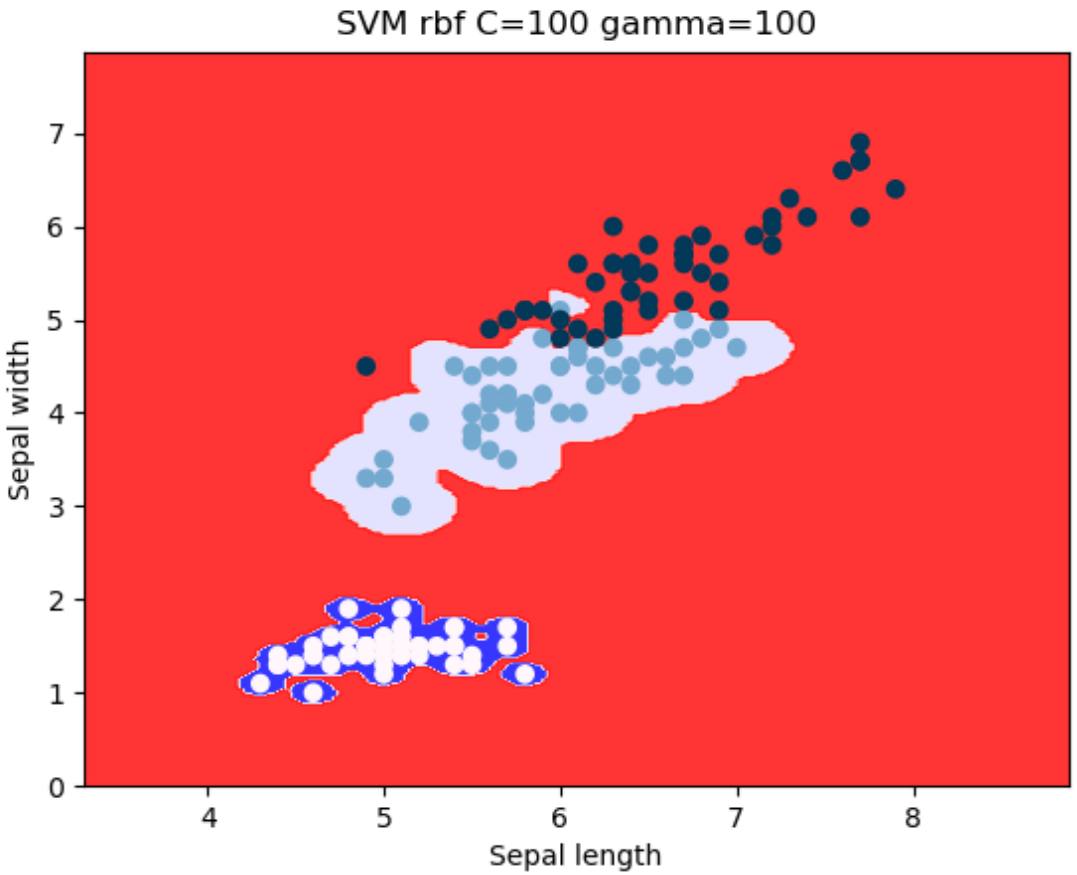
RBF $C = 100$



6 mispredicted

Example in python - RBF SVM

RBF $C = 100, \gamma = 100$



1 mispredicted

Example in python - WEKA

WEKA

- use LibSVM (install it using package manager in Tools)
- use default parameters (C-SVC, rbf)
- set C (cost) to 10 and gamma to 10
- in order to get 150 correctly classified instances

7. End

FACULTÉ
DES SCIENCES
Unité de formation
et de recherche
DÉPARTEMENT
INFORMATIQUE

UA - Angers
2 Boulevard Lavoisier 49045 Angers
Cedex 01
Tel: (+33) (0)2-41-73-50-72

université
angers
FACULTÉ
DES SCIENCES
Unité de formation
et de recherche
DÉPARTEMENT
INFORMATIQUE