INTRODUCTION TO MACHINE LEARNING

Leila GHARSALLI (leila.gharsalli@ipsa.fr)

IPSA, AERO 4

2022-2023

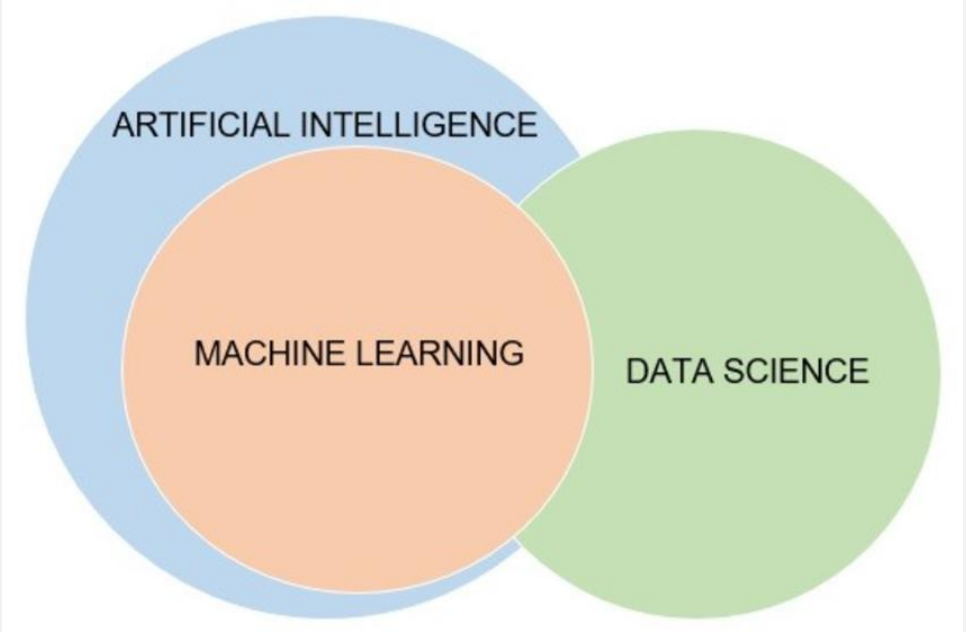
IPSA
ÉCOLE D'INGÉNIEURS
DE L'AIR, DE L'ESPACE ET DE LA MOBILITÉ DURABLE

GOALS OF THE COURSE

- *Instructor : Leila GHARSALLI*
- *mail to : leila.gharsalli@ipsa.fr*
- Provide a general introduction to data science, data mining and statistical learning,
- Understand the kind of problems each architecture of a ML technique is useful for,
- Grading : Participation in class as well as a final project will be graded,
- Main background needed: basic notions in probabilities and statistics, programming skill.

GOALS OF THE COURSE

- Understand different tools behind data science,
- Distinguish between different approaches to data modeling,
- Understand what are the different problems encountered in machine learning.



MAIN REFERENCES

Classical Machine Learning Textbooks:

- **Elements of statistical learning** (ESL), Hastie et al.,Springer
- **An introduction to statistical learning** (ISLR), Hastie et al.,Springer

Both books are availale online.

- **Pattern recognition and Machine Learning**, Bishop, Springer

A lot of courses and tutorials on the web:

- **Online courses** Coursera, Andrew Ng (CS229: Machine Learning (stanford.edu)), DataCamp

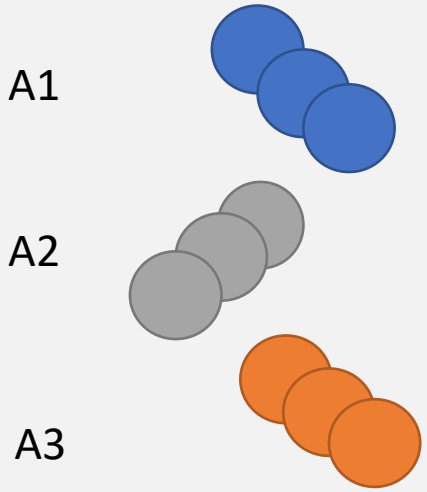
CONTENT

INTRODUCTION

MACHINE LEARNING

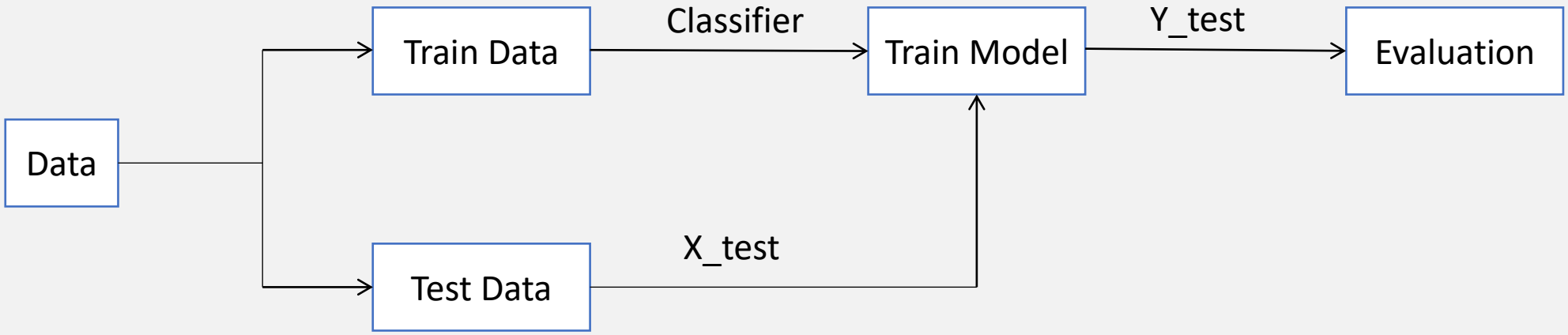
Supervised learning : knowledge of output : learning with the presence with an expert / teacher,

- Data Is labelled with a class or value,
- Goal: predict a class or value label (Neural Network, Support Vector Machine, Decision Trees, Bayesian classifiers…),
- Examples: face recognition, spam detection.



MACHINE LEARNING TERMS -1

- **Supervised - Classifiers**



INTRODUCTION

- Classification:

  -Use an object characteristics to identify which class/category it belongs to.

- Example:

  -A new email is ‘spam’ or ‘non-spam’;

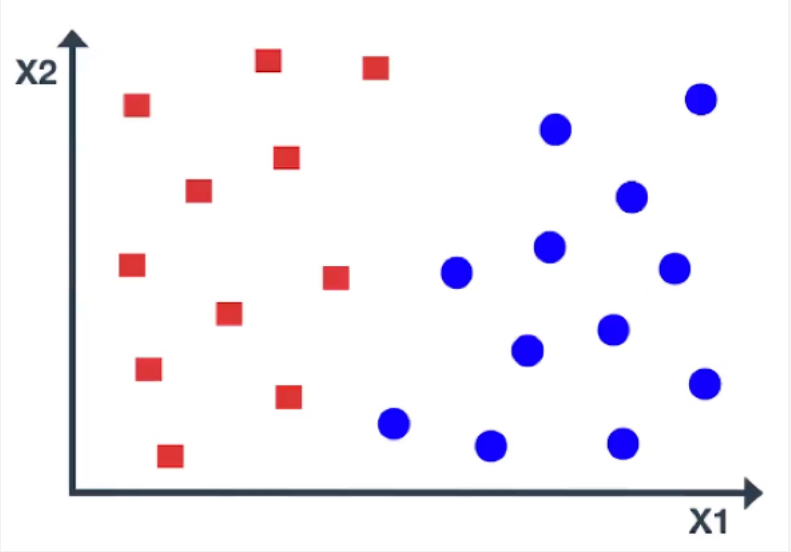
  -A patient diagnosed with a disease or not;

- Classification is an example of pattern recognition.

SUPPORT VECTOR MACHINE

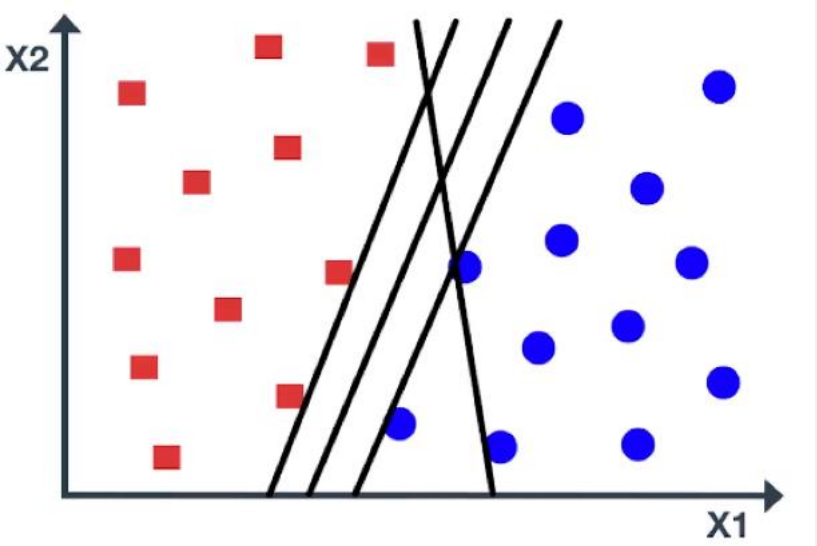
SUPPORT VECTOR MACHINE

Consider a data set of two different classes, shown in blue and red.



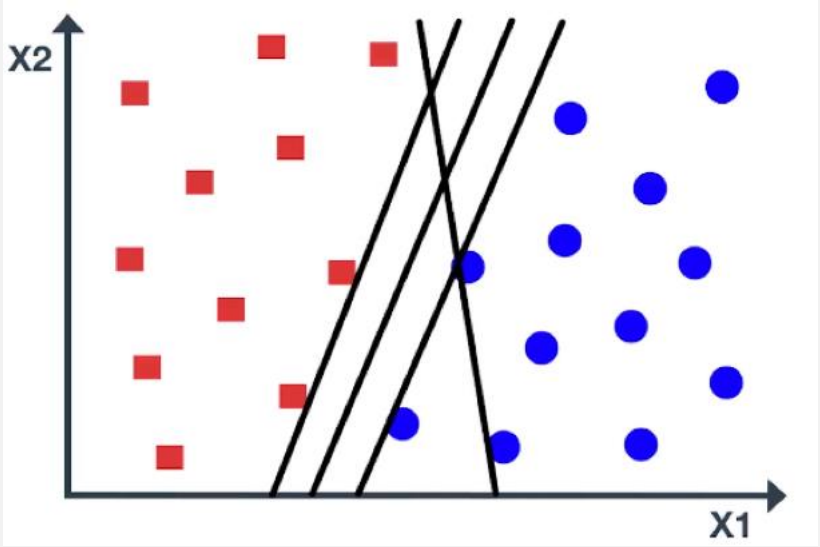
SUPPORT VECTOR MACHINE

- The data is linearly separable, and there exist multiple separating lines, which are shown in black. All these lines offer a solution, but only one line is optimal and can separate the classes accurately.



SUPPORT VECTOR MACHINE

- If the line is very close to the data points, even a small noise would lead to misclassification. Here is another line that separates the classes but doesn't look like a very natural one. So, the question is which is the best line?



SUPPORT VECTOR MACHINE

- SVMs were introduced by Boser, Guyon, Vapnik in 1992.
- SVMs have become popular because of their success in handwritten digit recognition, object recognition, speaker identification, face detections in images and target detection.
- SVMs are now important and active field of all Machine Learning research and are regarded as a main example of “kernel methods”.

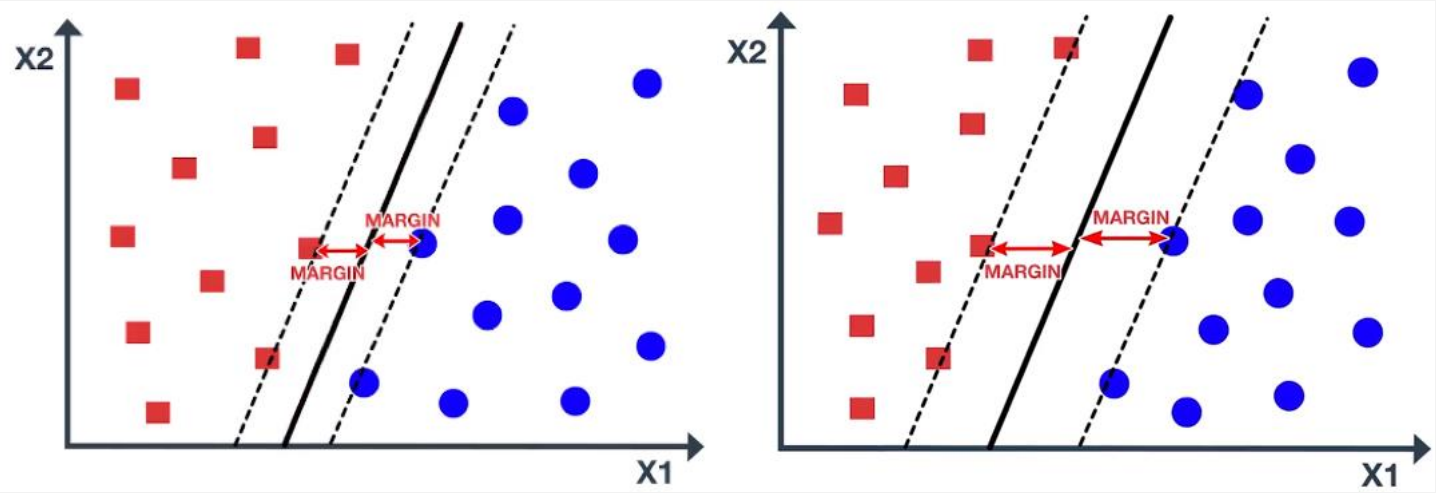
SUPPORT VECTOR MACHINE

- **Task:** given a set $S = \{x_i \in \mathbb{R}^n\}, i = 1,2,..,N$. Each point x_i belongs to either of two classes and thus given a lablel $y_i \in \{-1, 1\}$. The goal is to establish the equation of a hyperplane that divides S leaving all the points of the same class on the same side.

- SVM performs **classification** by constructing N-dimensional hyperplane that optimally separates the data into two categories.

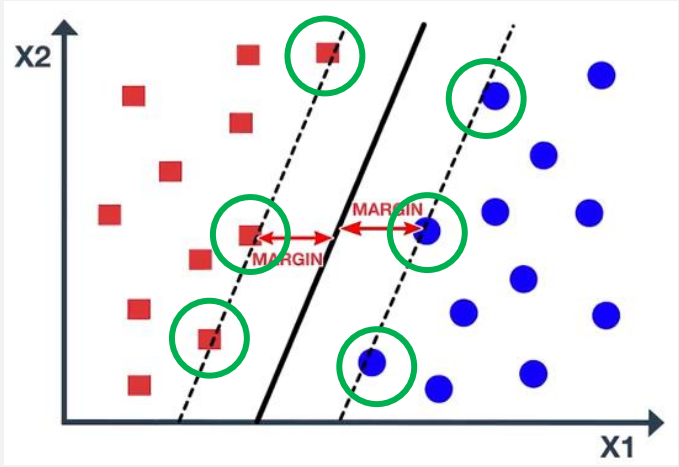
SUPPORT VECTOR MACHINE

- **Margins** is the perpendicular distance between the closest data points and Hyperplane.
- We select the hyperplane where the distance of the hyperplane from the closest data points is as large as possible (So, the Support Vector Machine is sometimes called Maximum Margin Classifier).



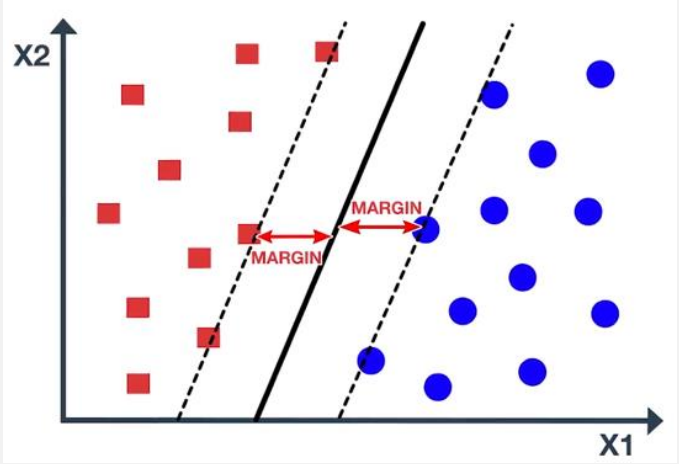
SUPPORT VECTOR MACHINE

- Let’s notice that not all the training points are important when choosing the hyperplane.
- **Support vectors** are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane (blue and red data points on the dashed lines).



SUPPORT VECTOR MACHINE

- Using these support vectors, we **maximize the margin of the classifier**. Deleting the support vectors will change the position of the hyperplane. These are the points that help us build our SVM.



SUPPORT VECTOR MACHINE

- Learning can be regarded as finding the maximum margin separating hyperplane between two classes of points. Suppose that a pair $(\boldsymbol{w}, b)$ defines a **hyperplane** which has the following equation:

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{w}.x + b$$

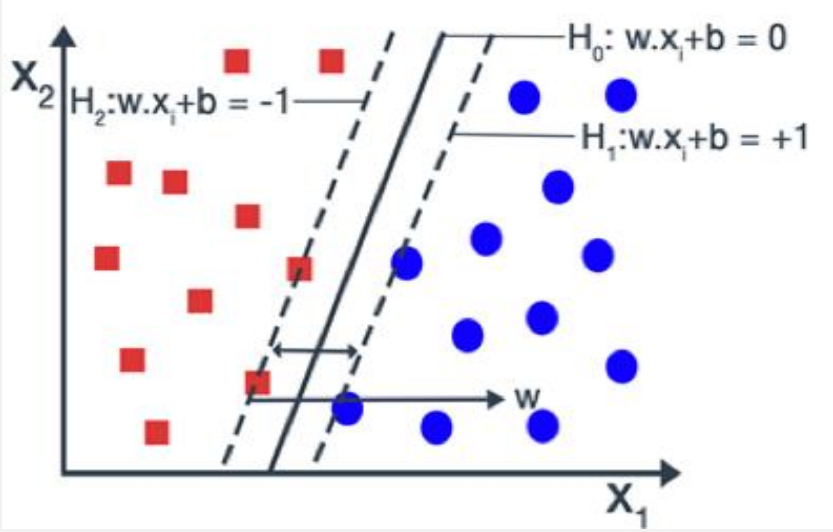
- Let $\{x_1, \ldots, x_m\}$ be our data set and let $y_i \in \{-1,1\}$ be the class label of x_i. The **decision boundary** should classify all points correctly i.e., the following equations must be satisfied:

$$y = \begin{cases} 1 & if \quad \boldsymbol{w}.x_i + b \geq 1 \\ -1 & if \quad \boldsymbol{w}.x_i + b \leq -1 \end{cases}$$

SUPPORT VECTOR MACHINE

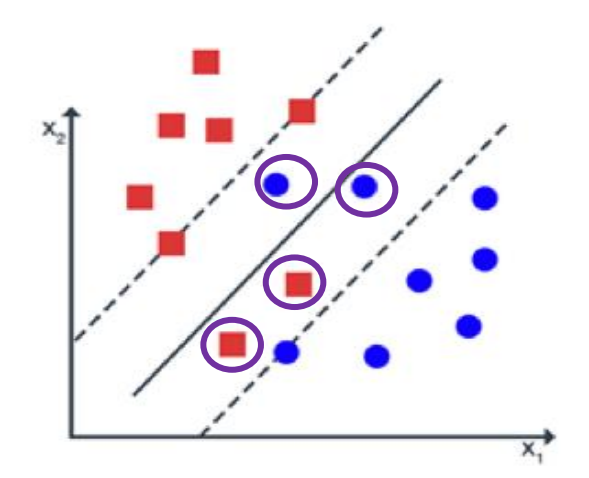
- Among all hyperplanes separating the data, there exists **a unique one yielding the maximum margin** of separation between the classes which can be determined in the following way;

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



SUPPORT VECTOR MACHINE

- Sometimes high noise in the data causes overlap of the classes as shown in the figure (there are points between the margin).



- In such cases, we can do the classification task by using **Soft Margin SVM**.

SOFT MARGIN SVM

- **A soft-margin SVM** provides freedom to the model to misclassify some data points by minimizing the number of such samples. Soft-margin SVM allows for the possibility of violating the constraints:

$$y_i\,(\boldsymbol{w}.x_i + b) \geq 1$$

- by introducing **slack variable** $\boldsymbol{\xi_i}$:

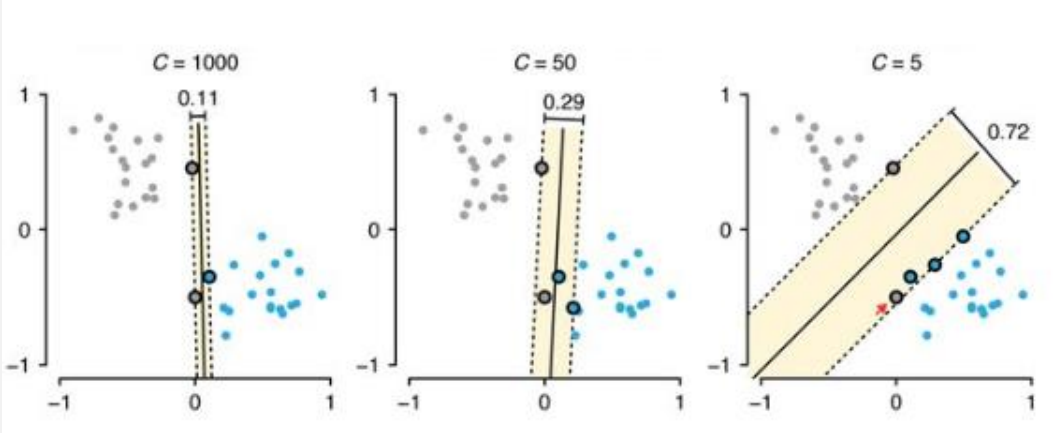
$$\frac{1}{2}||\boldsymbol{w}||^{\,2} + \boldsymbol{C}\sum_{i=1}^{n}\boldsymbol{\xi_i}$$; under the constraints: $y_i\,(\boldsymbol{w}.x_i + b) \geq 1 - \boldsymbol{\xi_i}$ and $\boldsymbol{\xi_i} \geq 0$

where $\boldsymbol{C}$ is a regularization parameter that balances the data fitting (left hand term) and the regularization (right hand term).

- The goal then is to maximize the margin by keeping the $\boldsymbol{\xi_i}$ as small as possible.

RELATION BETWEEN REGULARIZATION PARAMETER (C) AND SVM

- As the value of C increases the margin decreases thus Hard SVM.
- If the values of C are very small the margin increases thus Soft SVM.
- Large value of C can cause overfitting therefore we need to select the correct value using Hyperparameter Tuning.



HYPERPARAMETER TUNING

- **GridSearchCV** method is responsible to fit() models for different combinations of the parameters and give the best combination based on the accuracies.

- **Example:**

```
from sklearn.model_selection import GridSearchCV

param_grid = { 'C':[0.1,1,100,1000],'kernel':['rbf','poly','sigmoid','linear'],'degree':[1,2,3,4,5,
6]}
grid = GridSearchCV(SVC(),param_grid)
grid.fit(X_train,y_train)
```

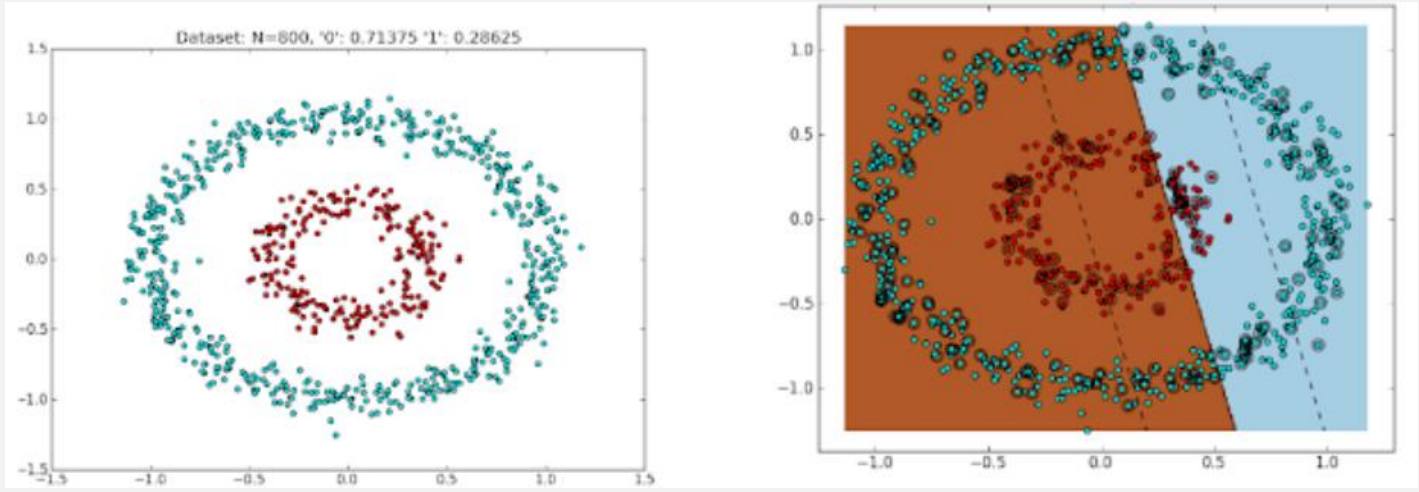
BIAS-VARIANCE TRADE-OFF

- The goal of any supervised machine learning algorithm is to achieve low bias (the error between model prediction and the ground truth) and low variance (the changes in the model when using different portions of the training data set). Hence, the algorithm should achieve good prediction performance.

- The SVM algorithm has low bias and high variance, but the trade-off can be changed by increasing the **C** parameter that influences the number of violations of the margin allowed in the training data which increases the bias but decreases the variance.

KERNELS AND NON-LINEAR SVM

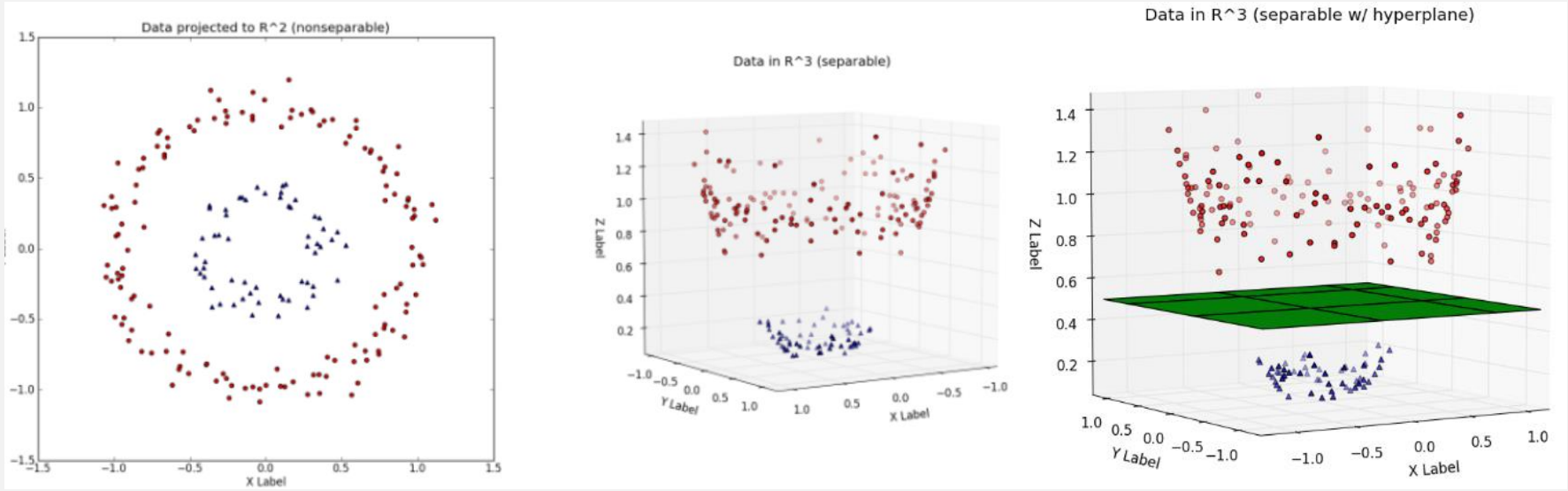
- What if our data is not linearly separable? Running a SVM on this data would yield horrible results.



- Can SVM only be used to separate linearly separable data?

KERNELS AND NON-LINEAR SVM

- We can modify our data and project it into higher dimensions to make it linearly separable.



- 2-D data projected onto 3-D using a transformation $[x1, x2] = [x1, x2, x_1^2 + x_2^2]$ thus making the data linearly separable.

WHAT ARE KERNELS?

- Maps data into a new space, then take the inner product of the new vectors.
- **Kernels functions** transform nonlinear spaces into linear ones.
- Kernel functions can be viewed as a similarity measure; the more similar the points x and y are the larger the value of $K(x, y)$ should be.
- When we run a linear SVM on such transformed data the probability of getting on accuracy of classification is nearly 100%.

POPULAR KERNELS

Some popular kernels are:

- **Linear kernels**: it is one of the simplest kernels and is just the inner product of x and y.

- **Polynomial kernel**: It represents the similarity of vectors in the training set of data in a feature space over polynomials of the original variables used in the kernel:

$$K(x, y) = (x^t y + c)^d, x \text{ and } y \text{ are vectors in the input space.}$$

- **RBF kernels**: (adding radial basis method to improve the transformation):

$$K(x, y) = \exp(-\gamma \left\| x_i - x_j \right\|^2)$$

SUPPORT VECTOR MACHINE

Python (sckit-learn)

```
>>> from sklearn.svm import SVC
```

https://scikit-learn.org/stable/modules/svm.html

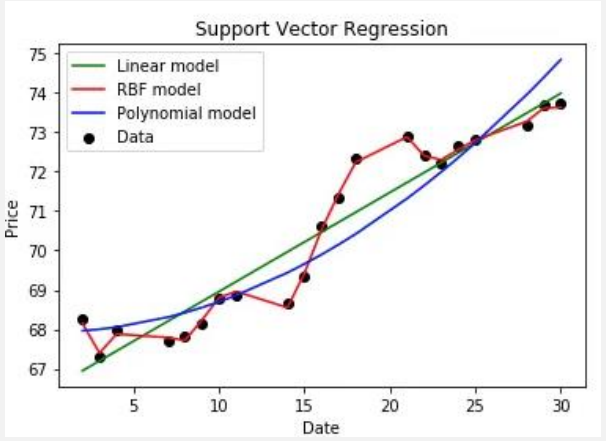
https://scikit-learn.org/stable/modules/svm.html#kernel-functions

SUPPORT VECTOR REGRESSOR

SUPPORT VECTOR REGRESSOR

- Support Vector Regression is used rarely it carries certain advantages:

1. It is robust to outliers.
2. Decision model can be easily updated.
3. It has excellent generalization capability, with high prediction accuracy.
4. Its implementation is easy.




```
>>> from sklearn.svm import SVR
```

https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

THANK YOU!