MATHEMATICAL TOOLS FOR DATA SCIENCE

Leila GHARSALLI (leila.gharsalli@ipsa.fr)

IPSA, AERO 4

2023-2024

GOALS OF THE COURSE

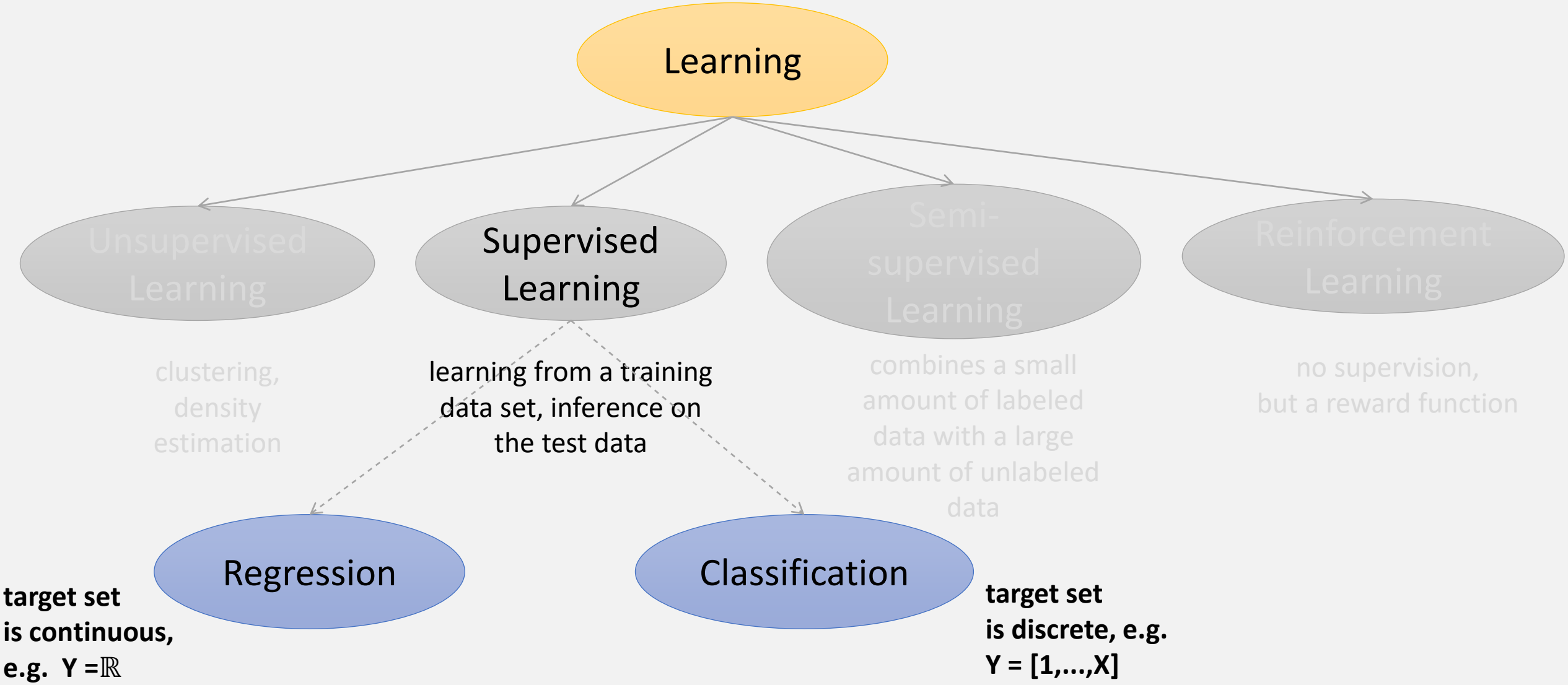
- *Instructor : Leila GHARSALLI*
- *mail to : leila.gharsalli@ipsa.fr*
- Provide a general introduction to data mining and data science.
- Introduce some important tools for solving problems in data science.
- Grading : Participation in class (practical work) as well as a final exam will be graded.
- Main background needed: basic notions in probabilities and statistics, programming skill.

CONTENT OF THE COURSE

1. Introduction
2. Linear regression
3. Sparse regression
4. Linear classification
5. Using trees for predictive analysis
6. Principal Component Analysis
7. Clustering
8. Density estimation

INTRODUCTION

INTRODUCTION



INTRODUCTION

- Classification:

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

- Example:

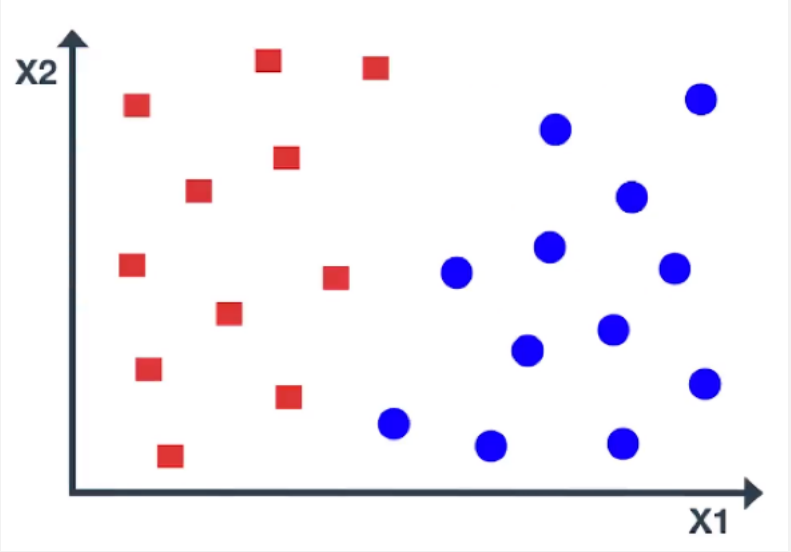
  -A new email is ‘spam’ or ‘non-spam’

  -A patient diagnosed with a disease or not

- Classification is an example of pattern recognition

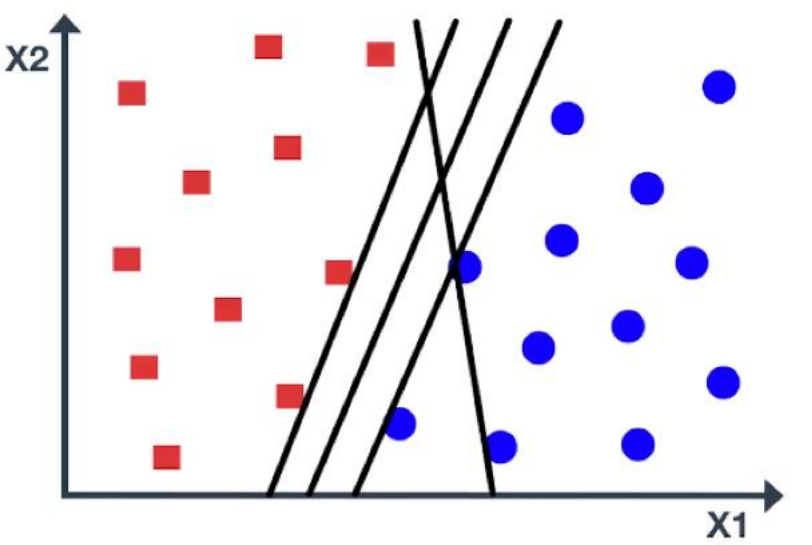
INTRODUCTION

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



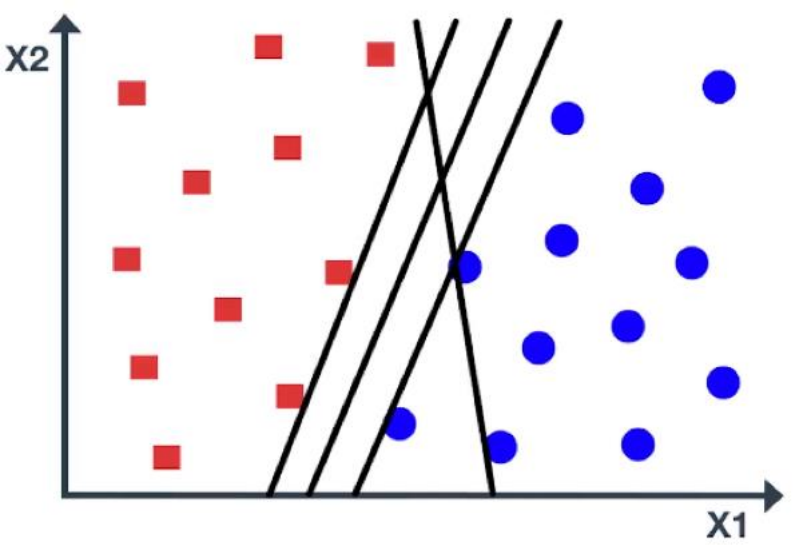
INTRODUCTION

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.



INTRODUCTION

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?



INTRODUCTION

- 3 major algorithms in linear binary classification:

1. **Logistic Regression**
2. **Perceptron**
3. **Support Vector Machine**

INTRODUCTION

- 3 major algorithms in linear binary classification:

1. **Logistic Regression** (we take weighted linear combination of input features and pass it through a sigmoid function which outputs a number between 1 and 0)

2. **Perceptron** (we take weighted linear combination of input features and pass it through a thresholding function which outputs 1 or 0)

3. **Support Vector Machine** (there can be multiple hyperplanes that separate linearly separable data. **SVM** calculates the optimal separating hyperplane using concepts of geometry)

SUPPORT VECTOR MACHINE

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”.

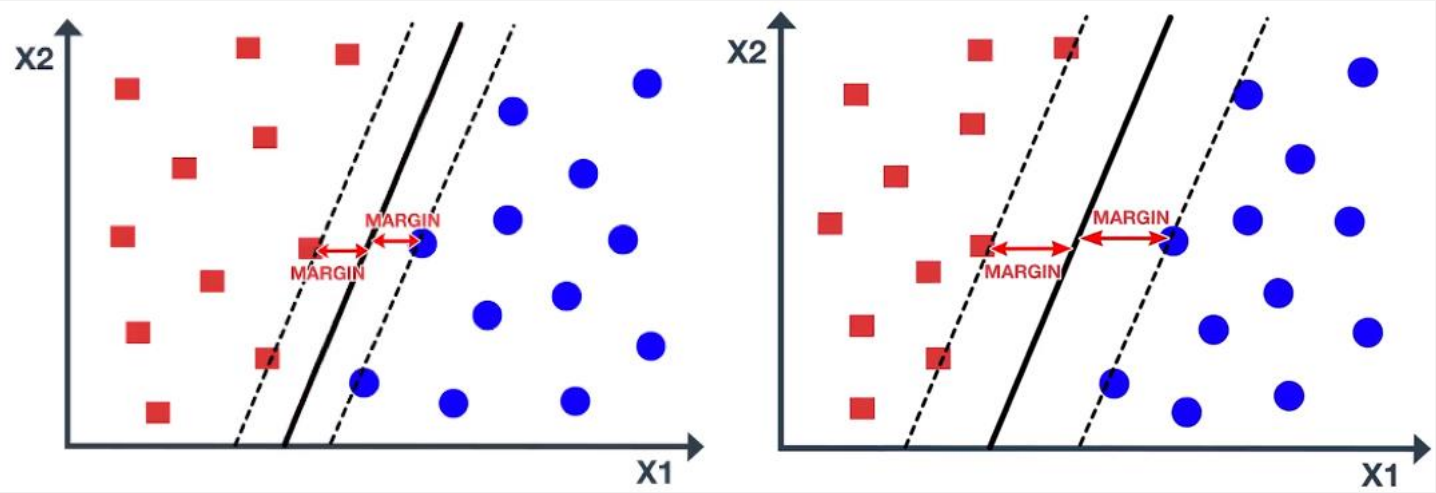
SUPPORT VECTOR MACHINE

- **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.

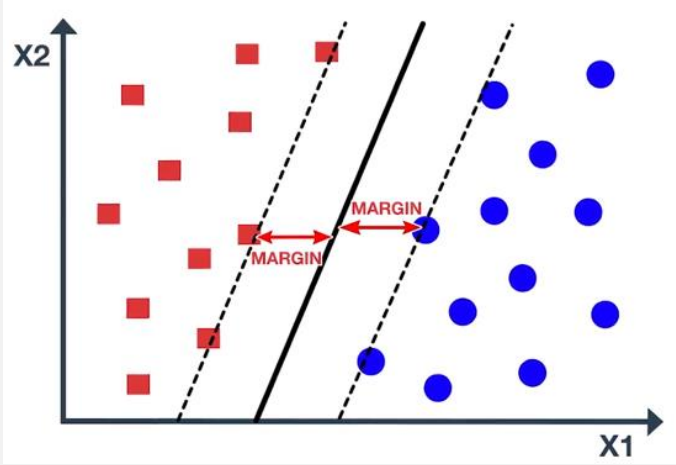
SUPPORT VECTOR MACHINE

- **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

- **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).



- 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:

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

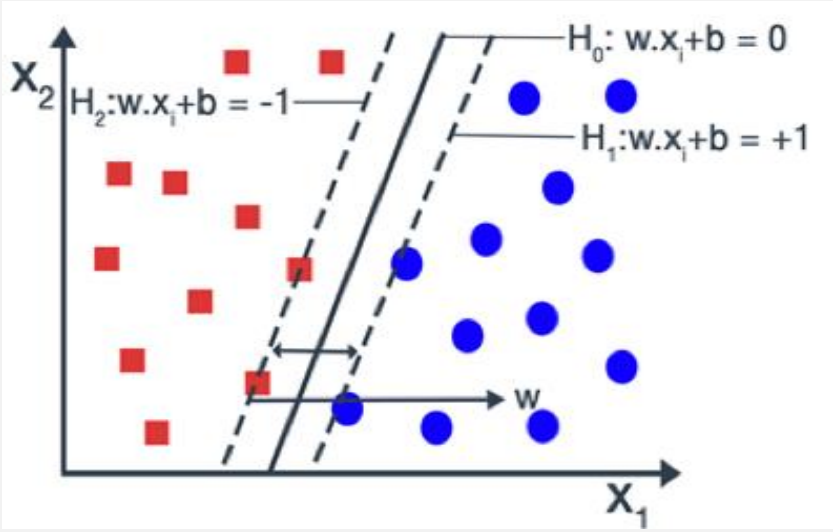
- Let $\{x_1, \ldots, x_m\}$ be our data set and let $y_i \in \{-1,1\}$ be the class label of x_i. The **decision boundary** should classify all points correctly i.e., the following equations must be satisfied:

$$\begin{aligned} &\boldsymbol{w}.x_i + b \geq 1 \; if \; y_i = 1 \\ &\boldsymbol{w}.x_i + b \leq -1 \; if \; y_i = -1 \end{aligned} \qquad \rightarrow \quad y_i\,(\boldsymbol{w}.x_i + b) \geq 1$$

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;

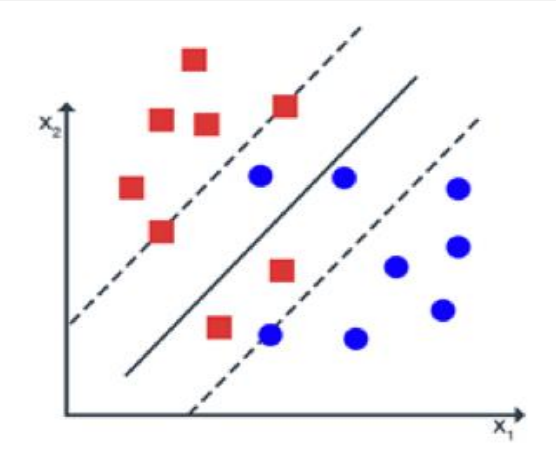
$$\boldsymbol{max_{w,b} min}\{\|\boldsymbol{x} - \boldsymbol{x_i}\| : \boldsymbol{x} \in \mathbb{R}^{\boldsymbol{N}}, (\boldsymbol{w}.\boldsymbol{x}) + \boldsymbol{b} = \boldsymbol{0}, \boldsymbol{i} = \boldsymbol{1}, \dots, \boldsymbol{m}\}$$



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

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 ξ_i**:

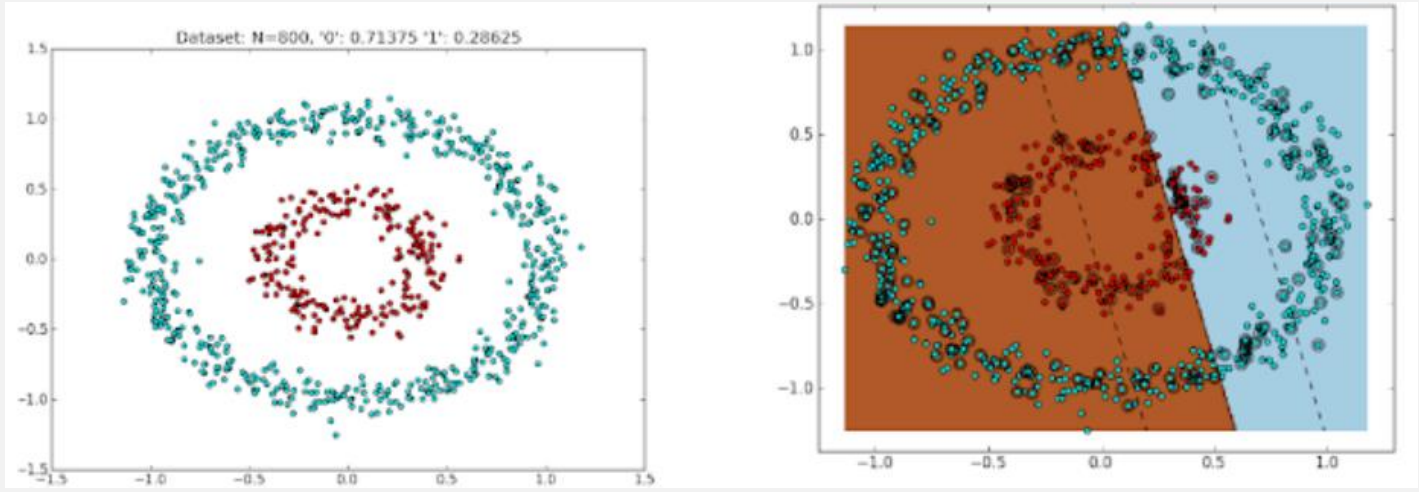
$$y_i\,(\boldsymbol{w}.x_i + b) \geq 1 - \xi_i\;;\; \xi_i \geq 0$$

- The goal then is to maximize the margin by keeping the ξ_i as small as possible.

KERNELS AND NON-LINEAR SVM

KERNELS AND NON-LINEAR SVM

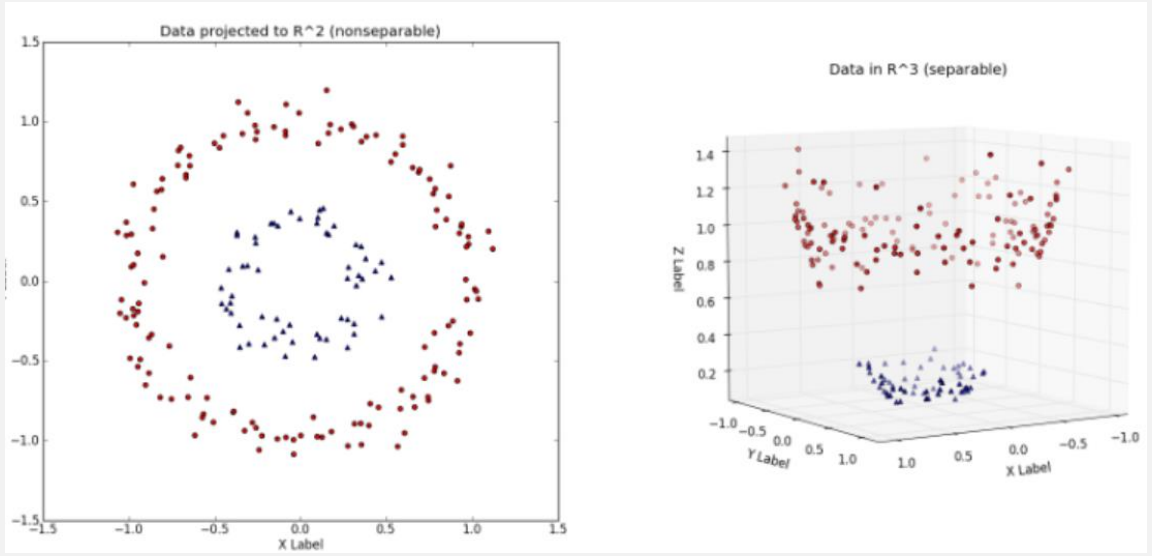
- What if out data is not linearly separable? Running a SVM on this data would yield horrible results.



- 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 $[x_1,x_2] = [x_1, x_2, |x_{12} + x_{22}]$ 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

- **Linear kernels**: it is one of the simplest kernel 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):

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

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

CROSS VALIDATION

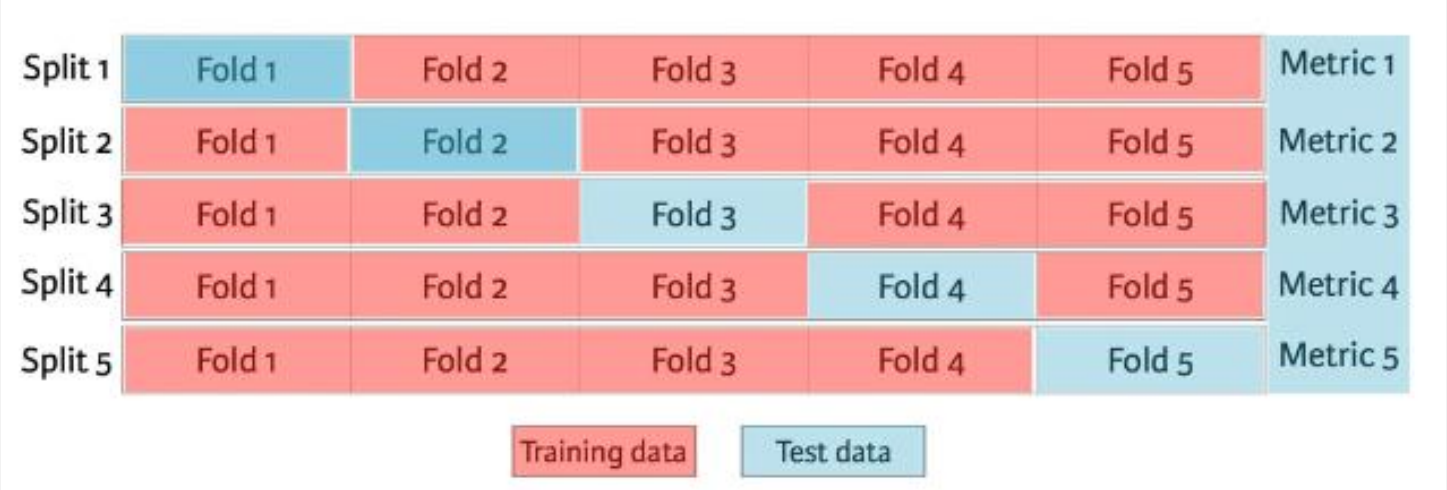
- A model validation technique for accessing how the result of statistical analysis will generalize to an independent dataset.
- Finding or estimating expected error.
- Selecting the best fit model.
- Avoiding overfitting.

CROSS VALIDATION METHODS

- Hold out sample validation (split into training and testing sets).
- K-folds cross validation (randomly split up into ‘k’ groups).
- Leave one out cross validation (split a into a training set and a testing set, using all but one observation as part of the training set).
- Bootstraps methods (statistical techniqueq for estimating quantities about a population by averaging estimates from multiple small data samples).

K-FOLDS CROSS VALIDATION

- For 5-fold cross validation, the dataset would be split into 5 groups, and the model would be trained and tested 5 separate times so each group would get a chance to be the test set.



EXERCISE 1

In this exercise, we will use the SVM to predict whether a bank currency note is authentic or not based upon four attributes of the note i.e., skewness of the wavelet transformed image, variance of the image, entropy of the image, and curtosis of the image.

1. Import the python libraries necessary for reading, writing and viewing data.
2. Load the data then display it for analysis.
3. Preprocess the data by dividing it firstly into attributes and labels, then into training and test sets.
4. Train now a liner SVM into training set.
5. Make predictions with the obtained model on the testing set.
6. Evaluate the model (the *classification_report* and *confusion_matrix* methods can be readily used to find out the values for these important metrics).

EXERCISE 2

In this exercise, we will use the iris dataset to predict the category to which a plant belongs based on four attributes: sepal-width, sepal-length, petal-width and petal-length using Kernel SVM.

1. Import the python libraries necessary for reading, writing and viewing data.
2. Load the data then preprocess it by dividing it firstly into attributes and labels, then into training and test sets.
3. Train now a polynomial SVM into training set (degree = 8).
4. Make predictions with the obtained model on the testing set.
5. Make the evaluation for the polynomial kernel. What do you obtain?
6. Repeat the same steps for Gaussian and sigmoid kernels.
7. Compare the different kernel performances. Conclude.