Classical Data Analysis

Master in Big Data Solutions 2017-2018

Francisco Gutierres
francisco.gutierres@bts.tech

Sara Hajian
sara.hajian@bts.tech

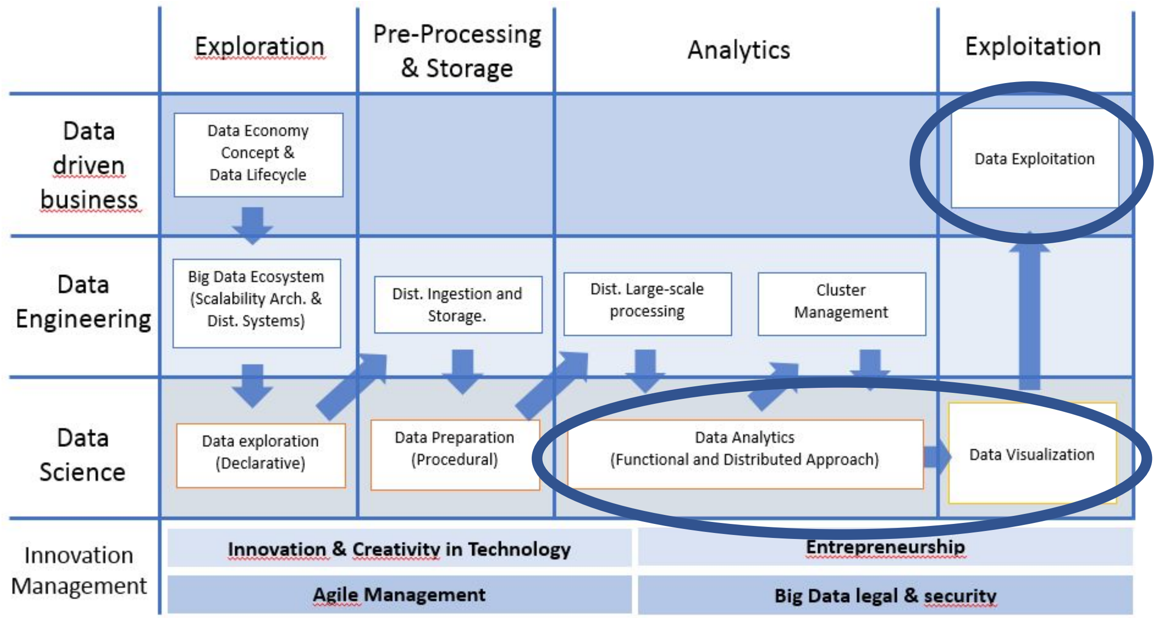
Session 2 - Classification
Support Vector Machine (SVM)

Sara Hajian

What we will learn

Session1: Classification

Support Vector Machine (SVM)

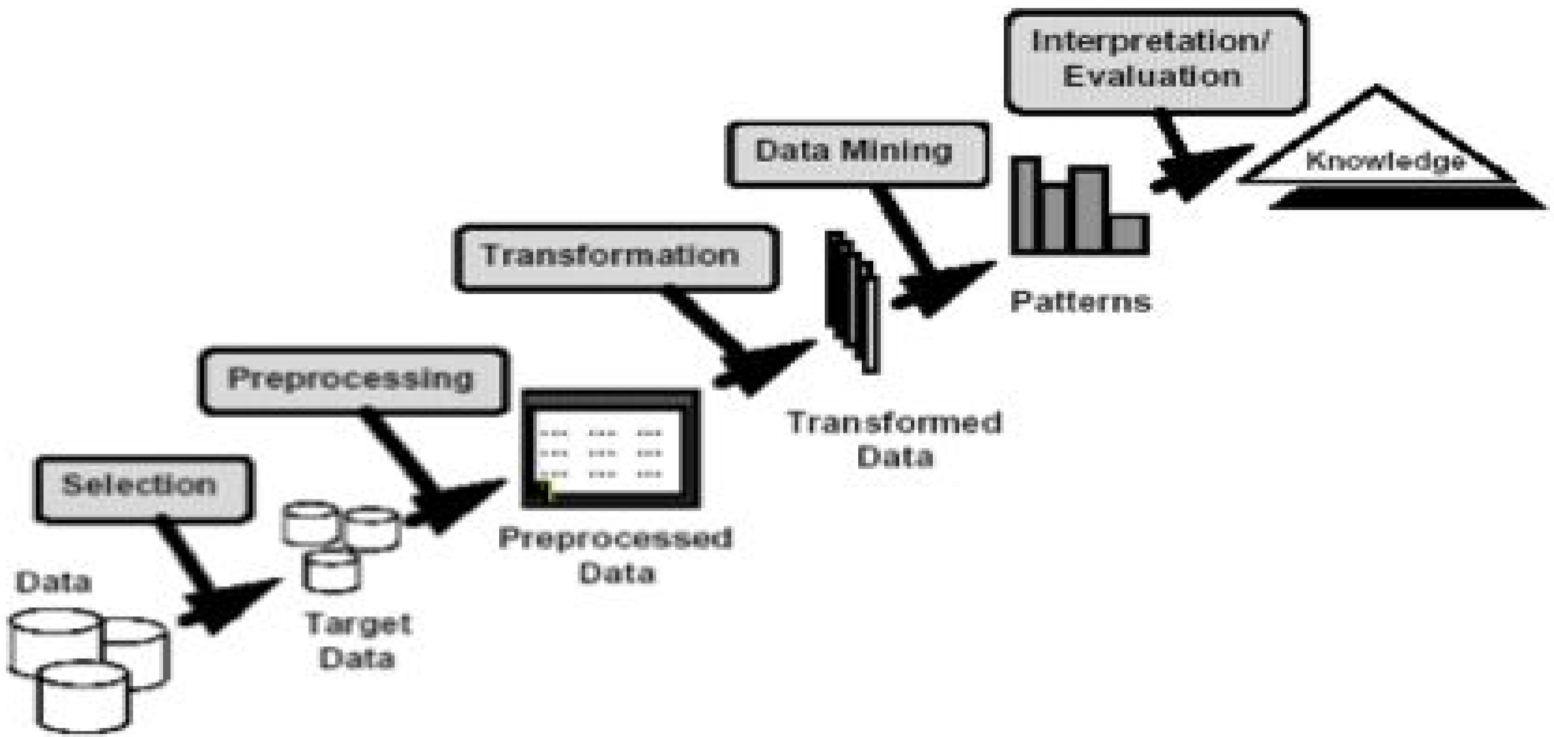


We will learn how to:

- What is data mining, its origin, different types of data mining algorithms
- What is classification, and know about its practical applications
- What is Support Vector Machines and how it works

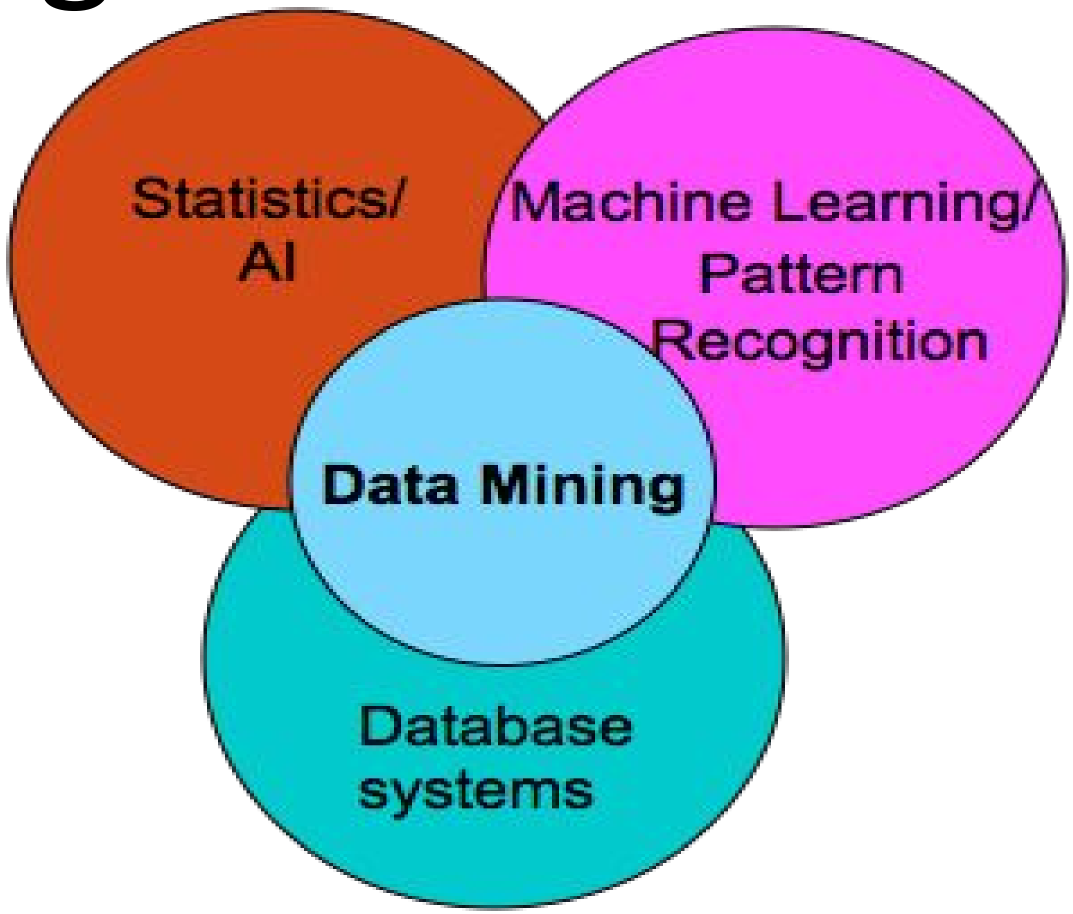
What is data mining

Non-trivial extraction of implicit, previously unknown and potentially useful information from data (i.e., **discovery of meaningful patterns**)



Tan, Pang-Ning. *Introduction to data mining*. Pearson Education India, 2006.

Origins of data mining



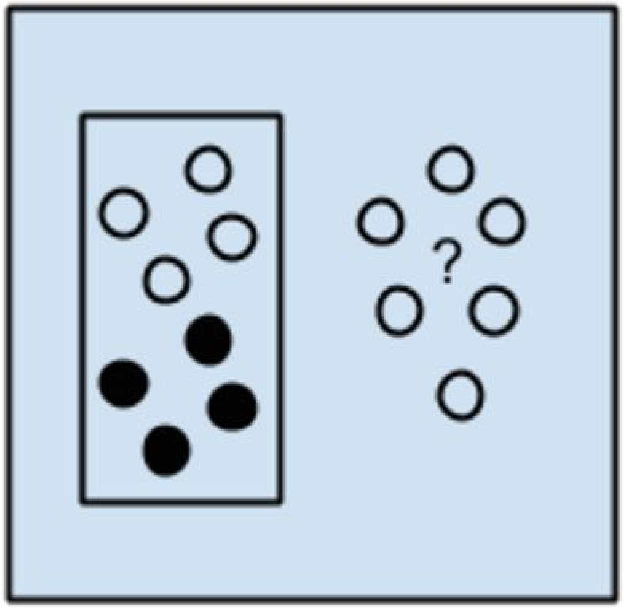
Tan, Pang-Ning. *Introduction to data mining*. Pearson Education India, 2006.

Data mining algorithms

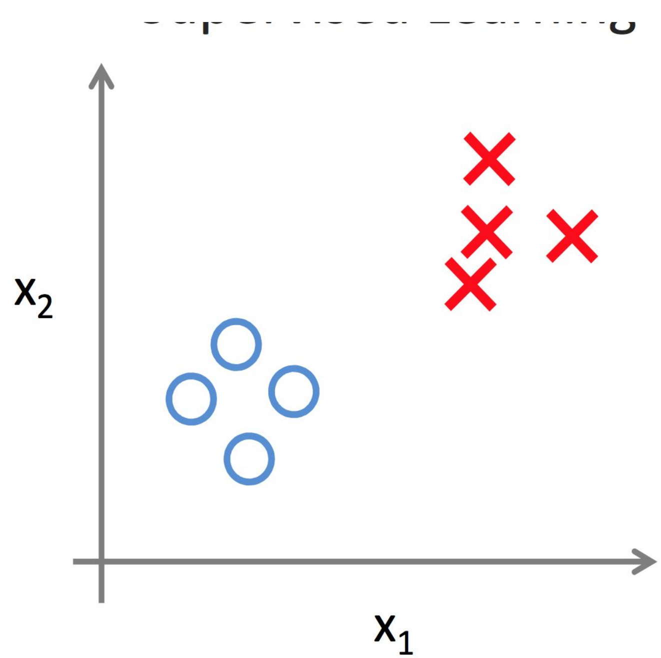
Supervised: Classification and regression

Unsupervised: Clustering

Data mining algorithms

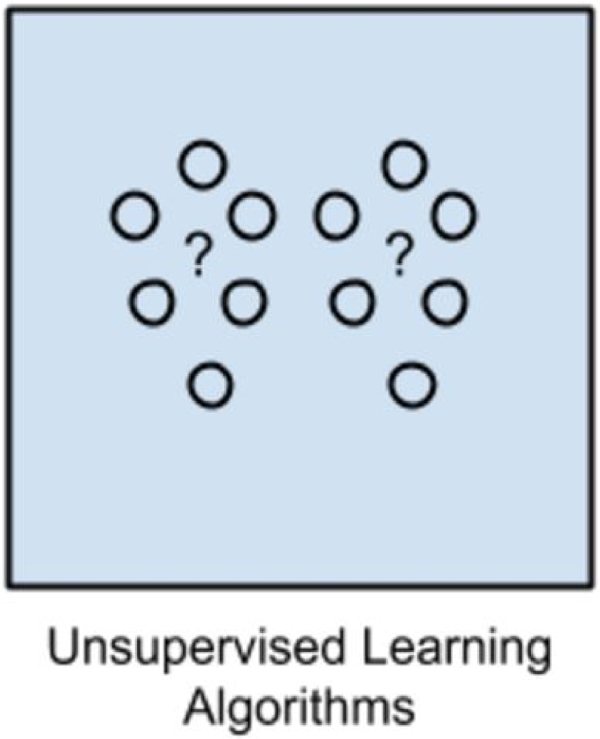


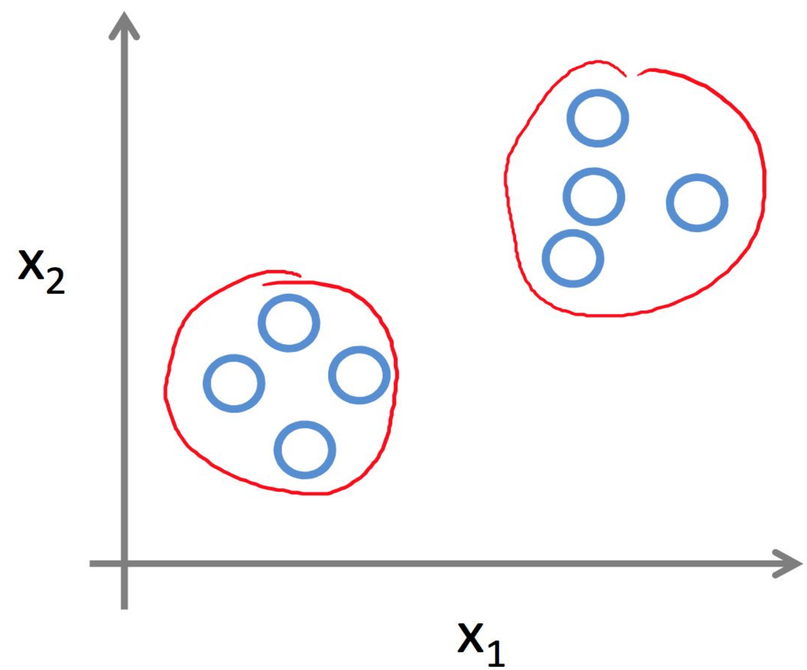
Supervised Learning Algorithms



Examples: regression and classification

Data mining algorithms





Examples: clustering

Classical Data analysis: Curriculum

Part I: Regression

Supervised learning algorithms

Linear regression

Logistic regression

Part II: Classification

Support vector machines (SVM)

Decision trees

Ensemble methods

Clustering

Classical Data analysis: Curriculum

Part I: Regression

Linear regression

Logistic regression

Part II: Classification

Supervised learning algorithms

Support vector machines (SVM)

Decision trees

Ensemble methods

Clustering

Classical Data analysis: Curriculum

Part I: Regression

Linear regression

Logistic regression

Part II: Classification

Support vector machines (SVM)

Decision trees

Ensemble methods

Clustering

Unsupervised learning algorithms

Classification: definition

Given a collection of records (*training set*)

- Each record contains a set of *attributes*, one of the attributes is the *class*.

Find a *model* for class attribute as a function of the values of other attributes.

Goal: <u>previously unseen</u> records should be assigned a class as accurately as possible.

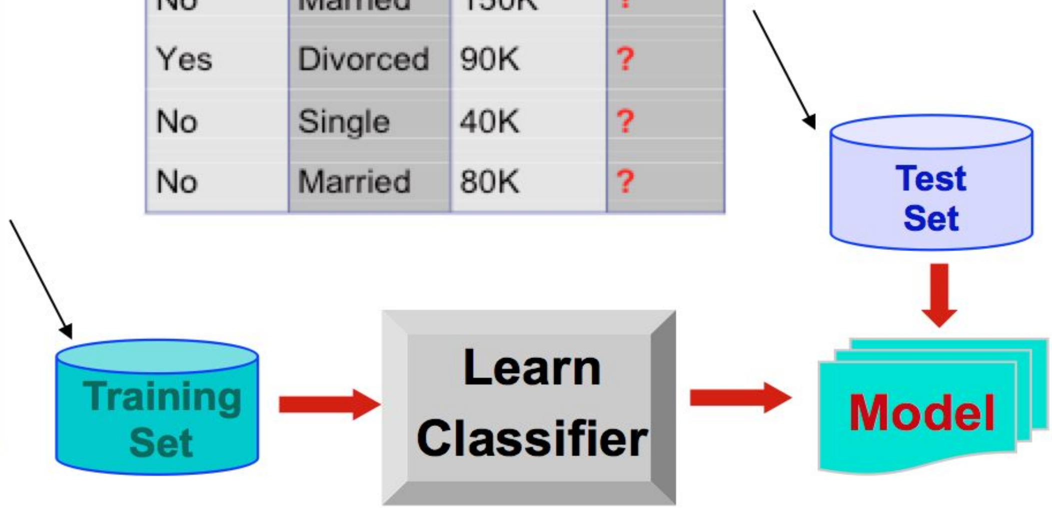
- A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification: example

categorical categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Tan, Pang-Ning. *Introduction to data mining*. Pearson Education India, 2006.

Classification: applications

Direct Marketing

– Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.

Fraud Detection

– Goal: Predict fraudulent cases in credit card transactions.

Customer Attrition/Churn:

– Goal: To predict whether a customer is likely to be lost to a competitor.

Classification: algorithms

Support vector machines (SVM)

Decision trees

Ensemble methods

Rule-based Methods

Neural Networks

Bayesian algorithms such as Naïve Bayes

Instance-based algorithms such kNN

Deep learning

Classification: algorithms

Support vector machines (SVM)

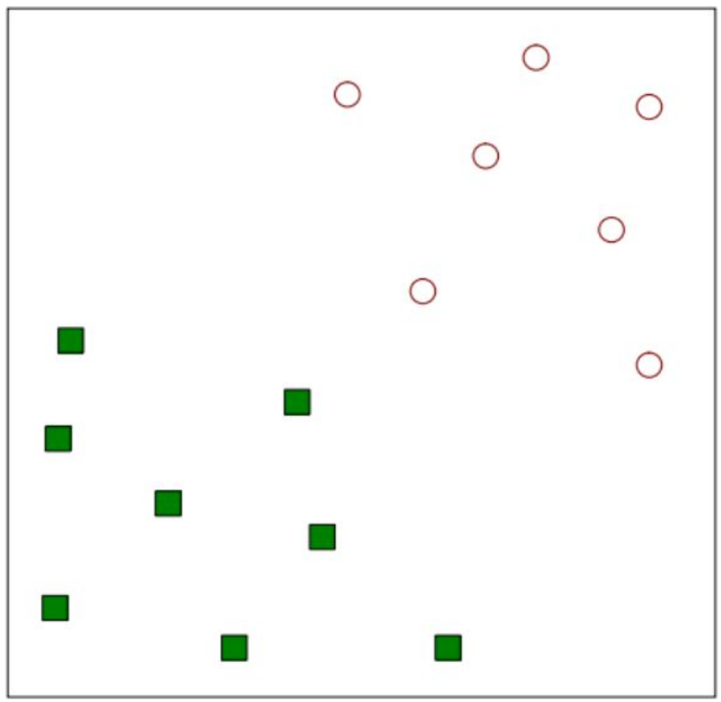
Applications: handwritten digit recognition, text categorization, image classification

Advantages: works well with high-dimensional data

works both with numerical and categorical data

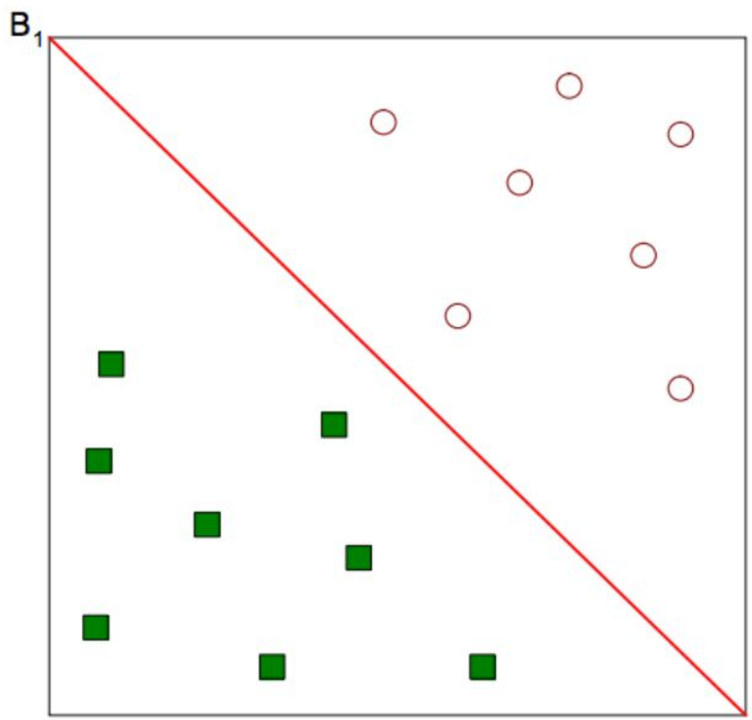
Support Vector Machines (SVM)

Find a linear hyperplane (**decision boundary**) that will separate the data

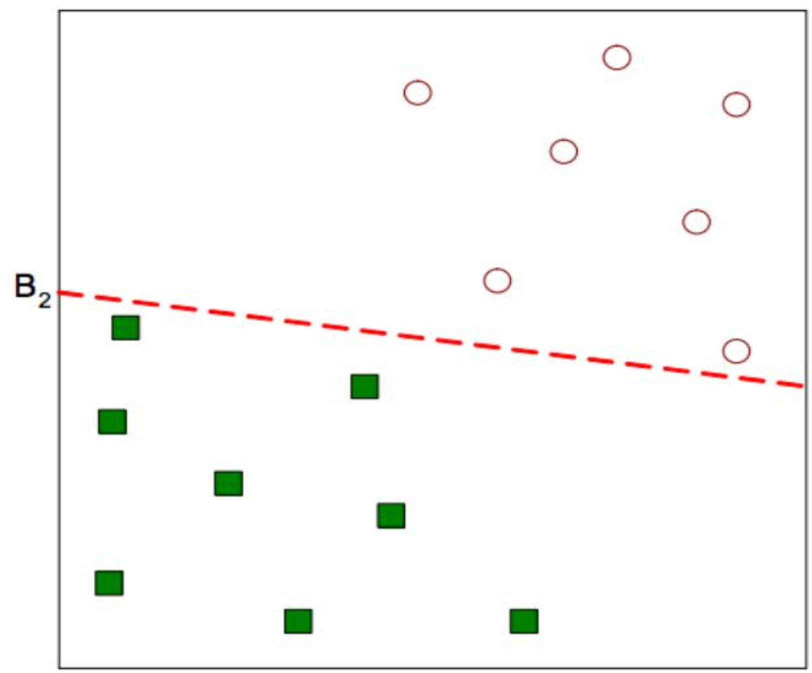


Support Vector Machines (SVM)

One possible solution

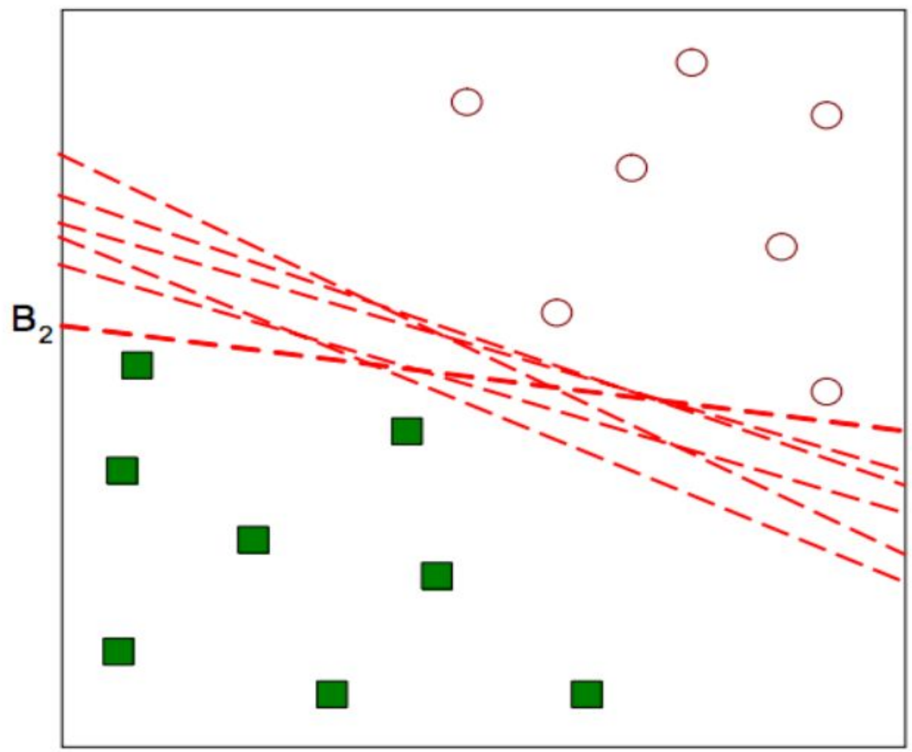


Another possible solution



Support Vector Machines (SVM)

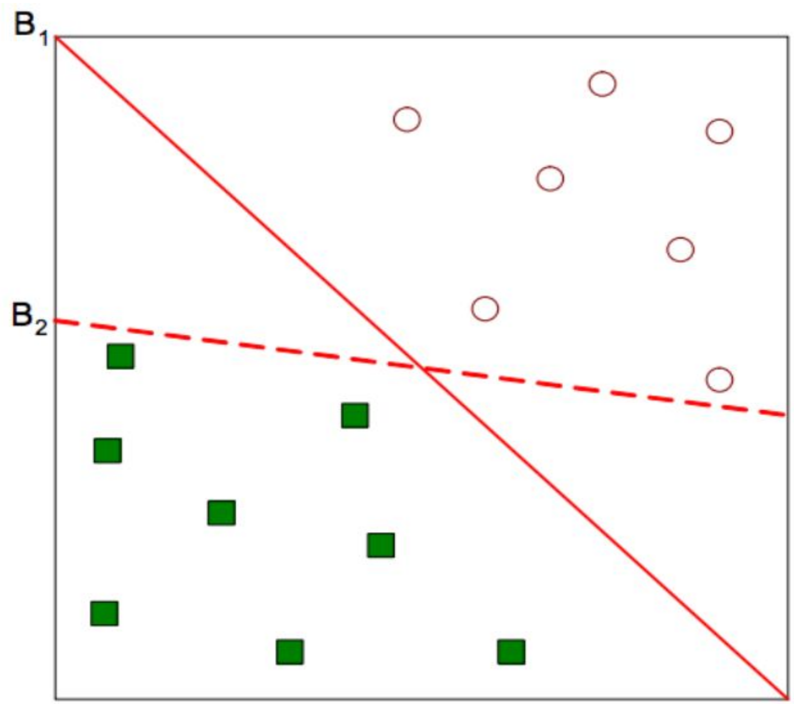
Other possible solutions



Support Vector Machines (SVM)

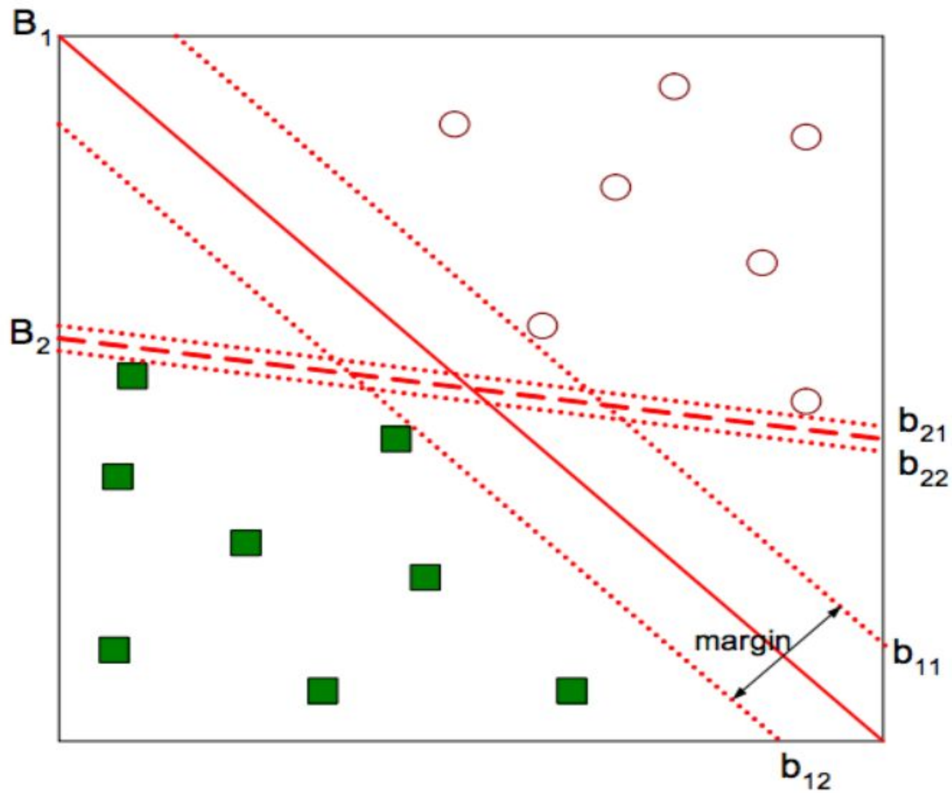
Which one is better? B1 or B2?

How do you define better?



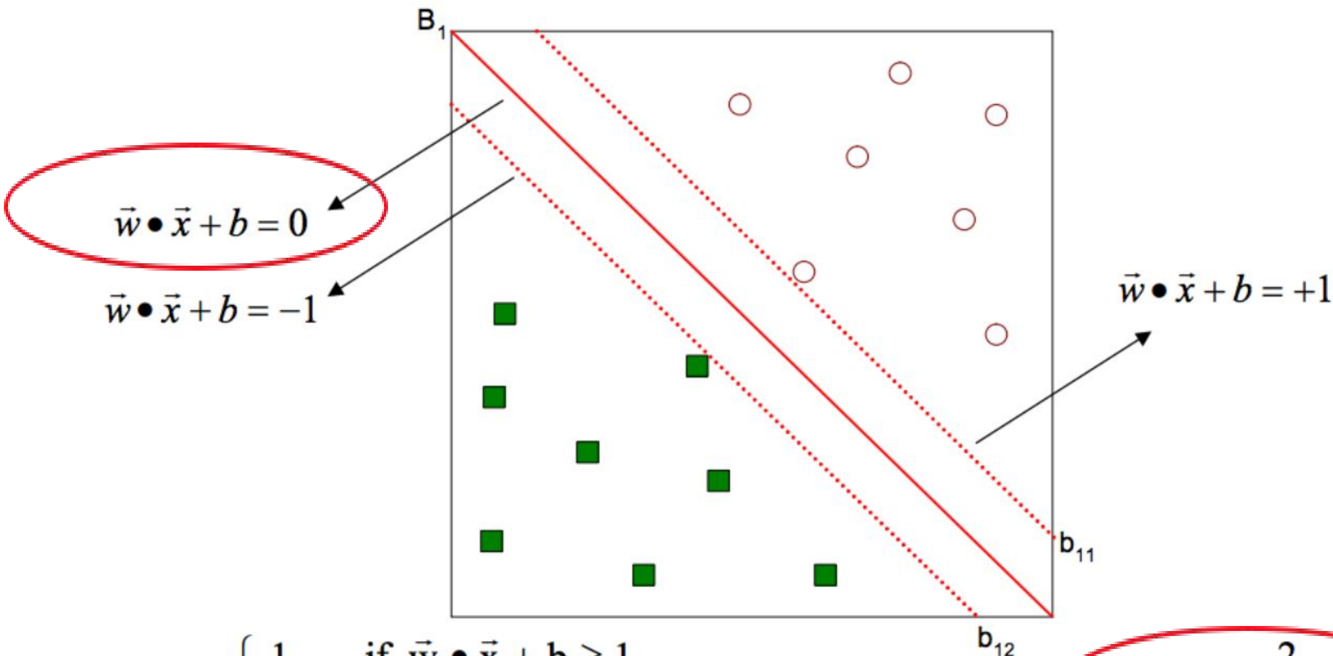
Support Vector Machines (SVM)

Find hyperplane **maximizes the margin** => B1 is better than B2



Support Vector Machines (SVM)

linear decision boundary and margin of a linear classifier



$$f(\vec{x}) = \begin{cases} 1 & \text{if } \vec{w} \bullet \vec{x} + b \geq 1 \\ -1 & \text{if } \vec{w} \bullet \vec{x} + b \leq -1 \end{cases}$$

$$\text{Margin} = \frac{2}{\|\vec{w}\|^2}$$

Support Vector Machines (SVM)

Learning the parameters of the decision boundary

We want to maximize: $\text{Margin} = \frac{2}{\|\vec{w}\|^2}$

- Which is equivalent to minimizing: $L(w) = \frac{\|\vec{w}\|^2}{2}$
- But subjected to the following constraints:

$$f(\vec{x}_i) = \begin{cases} 1 & \text{if } \vec{w} \bullet \vec{x}_i + b \geq 1 \\ -1 & \text{if } \vec{w} \bullet \vec{x}_i + b \leq -1 \end{cases}$$

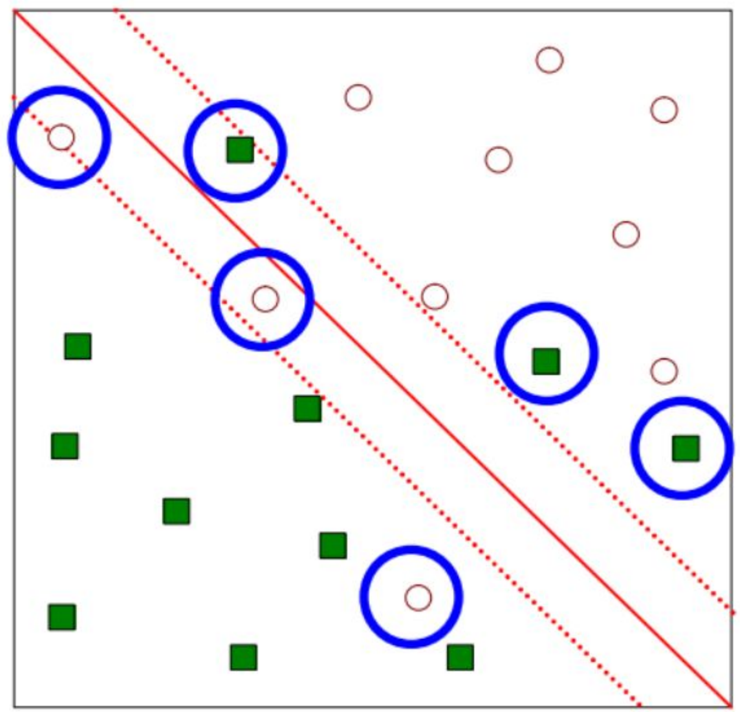
- This is a constrained optimization problem
  - Numerical approaches to solve it (e.g., quadratic programming)

Barcelona Technology School

eurecat Centre Tecnològic de Catalunya

Support Vector Machines (SVM)

What is the problem is not linearly separable?



Support Vector Machines (SVM)

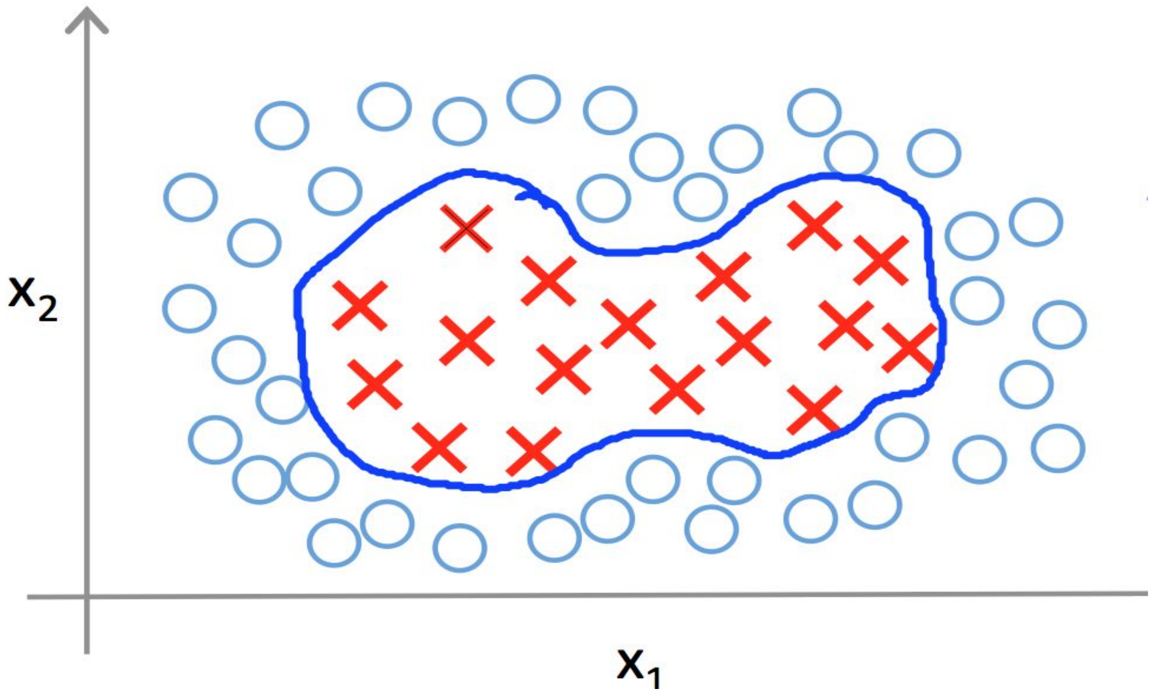
The learning algorithm must consider the **trade-off between the width of margin and the number of training errors**

- Introduce slack variables
  - Need to minimize: $$L(w) = \frac{\| \vec{w} \|^2}{2} + C\left(\sum_{i=1}^{N} \xi_i^k\right)$$
  - Subject to:

$$f(\vec{x}_i) = \begin{cases} 1 & \text{if } \vec{\mathrm{w}} \bullet \vec{\mathrm{x}}_{\mathrm{i}} + \mathrm{b} \geq 1 - \xi_{\mathrm{i}} \\ -1 & \text{if } \vec{\mathrm{w}} \bullet \vec{\mathrm{x}}_{\mathrm{i}} + \mathrm{b} \leq -1 + \xi_{\mathrm{i}} \end{cases}$$

Support Vector Machines (SVM)

What if the decision boundary is not linear?



Non-linear SVM

Support Vector Machines (SVM)

Key points and characteristics

Individual assignment

Choose one of the following problem scenarios and describe you approach to solve them:

Scenario1 : Direct Marketing

- Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.

Scenario 2: Fraud Detection

- Goal: Predict fraudulent cases in credit card transactions.

Scenario 3: Customer Attrition/Churn

- Goal: To predict whether a customer is likely to be lost to a competitor.

Next class

Support vector machines with Python:

- load the breast cancer dataset from Scikit Learn
- split the data into train and test set
- train a support vector classier, use the trained model for prediction of test set
- evaluation of the SVM model and predictions

Resources

- Tan, Pang-Ning. *Introduction to data mining*. Pearson Education India, 2006.

- Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. New York: Springer series in statistics, 2001.

Thank you

Barcelona, 2017