TRANSFORM 4EUROPE

Co-funded by the
Erasmus+ Programme
of the European Union

Machine learning (I)

David Tomás Díaz
@d_tomas

Universidad de Alicante

lsi Departamento de Lenguajes y Sistemas Informáticos

Contents

- What is machine learning?
- Components
  - Data
  - Features
  - Algorithms
- Classification
- Regression
- Evaluation

Contents

- **What is machine learning?**
- Components
  - Data
  - Features
  - Algorithms
- Classification
- Regression
- Evaluation

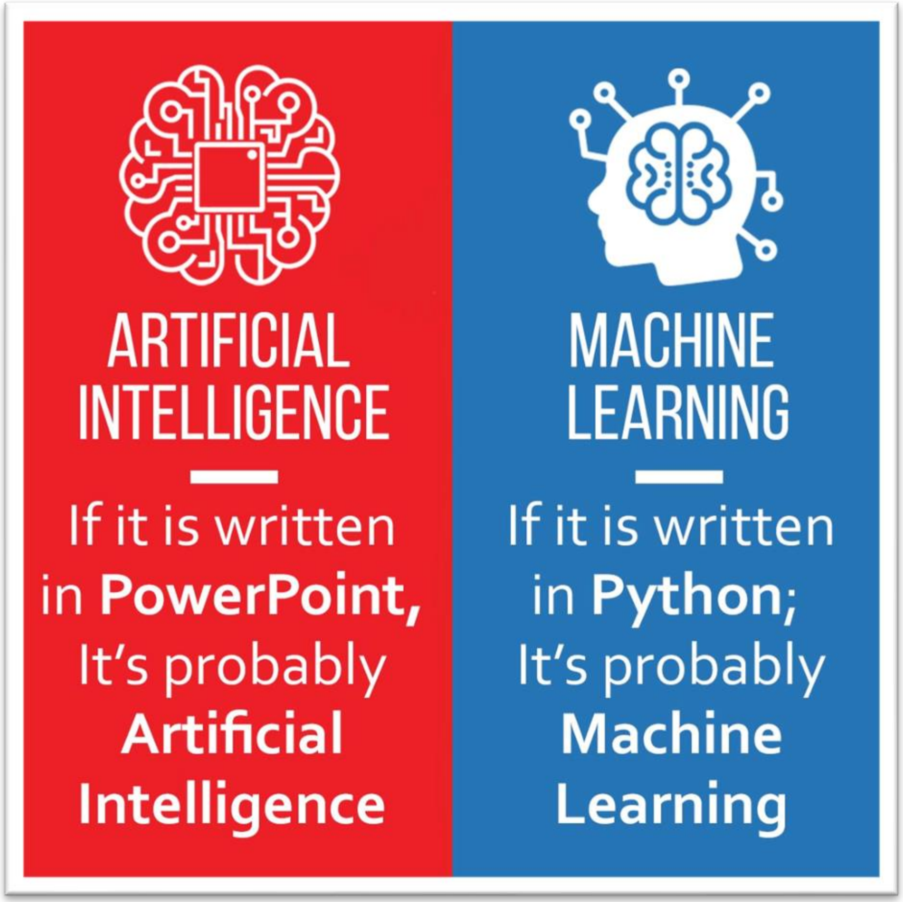
What is machine learning?

- Branch of Artificial Intelligence (AI)
  - Develop techniques that allow computers to learn from data
  - Generalize from experience (induction) and build a model
  - Gives support to data mining and text mining tasks

> A computer program is said to learn from experience **E** with respect to some class of tasks **T** and performance measure **P**, if its performance at tasks in **T**, as measured by **P**, improves with experience **E**

Tom M. Mitchell

What is machine learning?



ARTIFICIAL INTELLIGENCE

If it is written in **PowerPoint,** It's probably **Artificial Intelligence**

MACHINE LEARNING

If it is written in **Python;** It's probably **Machine Learning**

Contents

- What is machine learning?
- **Components**
  - Data
  - Features
  - Algorithms
- Classification
- Regression
- Evaluation

Components

- Components of a machine learning system
  - Data
    - Set of samples that are used to train/evaluate the system
  - Features
    - Attributes that represent each of the samples in the dataset
  - Algorithms
    - Operations that allow learning from the features obtained from the training data to generate a model

Contenidos

- What is machine learning?
- Components
  - **Data**
  - Features
  - Algorithms
- Classification
- Regression
- Evaluation

Data

- Data refers to collections of objects
  - Patients in a hospital
  - Customers of a telephone operator
  - Travels by train from Barcelona to Madrid
  - Access to a web server
  - Animals from a zoo
  - Houses sold in an area
  - …
- It is the fuel of machine learning systems
- The data available in an application is called a *dataset* or *corpus* (in the case of texts)

Data

- It is necessary to dedicate a lot of effort to be sure that the data is of quality
  - Get a broad set
  - Make it representative
  - Remove false observations
  - Clean
  - Format
  - …
- It doesn't matter how sophisticated the algorithms are if the data is not adequate

Data

We don’t have better algorithms.
We just have more data.

Peter Norving (Google Inc.)

Contents

- What is machine learning?
- Components
  - Data
  - **Features**
  - Algorithms
- Classification
- Regression
- Evaluation

Features

- Each data (object) is described by a number of attributes (features) that represent its properties
  - E.g., for a person: eye color, height, weight, age, ...
- Two fundamental types of features
  - Discrete
    - Contains labels that represent categories
    - E.g., color of an object, zip code, pass/fail, ...
  - Continuous
    - Takes numerical values
    - E.g., number of children, height, age, weight, ...

Features

- A dataset is typically represented as a table or a series of feature vectors
  - Each column is a feature
  - Each row is an instance (object)

Features

Instances

Id	Refund	Marital status	Salary	Fraud
1	Yes	Single	125,000€	No
2	No	Married	100,000€	No
3	No	Single	70,000€	No
4	No	Divorced	95,000€	Yes
…	…	…	…	…

Features

- Labelled data
  - There is a special feature for each instance called *class*

class

Id	Refund	Marital status	Salary	Fraud
1	Yes	Single	125,000€	No
2	No	Married	100,000€	No
3	No	Single	70,000€	No
…	…	…	…	…

- Unlabelled data
  - No *class* is defined

Contents

- What is machine learning?
- Components
  - Data
  - Features
  - **Algorithms**
- Classification
- Regression
- Evaluation

Algorithms

- Supervised learning (predictive methods)
  - Training instances are labelled
  - Use some variables to predict future or unknown values of other variables
  - Approaches
    - Classification
    - Numeric regression
- Unsupervised learning (descriptive methods)
  - Training instances are unlabelled
  - Find human-interpretable patterns that describe the data
  - Approaches
    - Clustering
    - Association rules

Contents

- What is machine learning?
- Components
  - Data
  - Features
  - Algorithms
- **Classification**
- Regression
- Evaluation

Classification

- Most common application in data mining and text mining
- Works with labelled data (*supervised learning*)
- Assign a label (*class*) to a new unlabelled input instance based on the knowledge acquired in the training process
  - E.g., `{positive, negative, neutral}`, `{man, woman}`, ...
- We must find a model for the class attribute based on the values of the other attributes

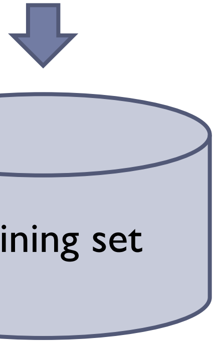
Id	Refund	Marital status	Salary	Fraud
1	Yes	Single	125,000€	No
2	No	Married	100,000€	No
3	No	Single	70,000€	No
…	…	…	…	…

Classification

- Training set
  - Instances with assigned labels
  - Used to train and build the model
- Validation set
  - Instances with assigned labels
  - Used to adjust parameters of the model and select the best configuration
  - Not always necessary
- Test set
  - Instances with assigned labels
  - Used to validate the model, comparing the assigned labels with the labels produced by the model

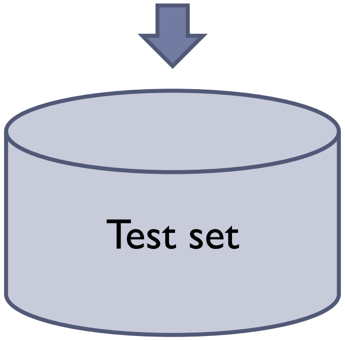
Classification

Re.	Status	Salary	Fraud
Yes	Single	125,000€	No
No	Married	100,000€	No
No	Single	70,000€	No
Yes	Married	120,000€	No
No	Divorced	95,000€	Yes
No	Married	60,000€	No
…	…	…	…



Training set

Re.	Status	Salary	Fraud
No	Single	75,000€	?
Yes	Married	50,000€	?
No	Married	150,000€	?
Yes	Divorced	90,000€	?
…	…	…	…



Test set

Classification

Re.	Status	Salary	Fraud
Yes	Single	125,000€	No
No	Married	100,000€	No
No	Single	70,000€	No
Yes	Married	120,000€	No
No	Divorced	95,000€	Yes
No	Married	60,000€	No
…	…	…	…

Training set

Class to predict

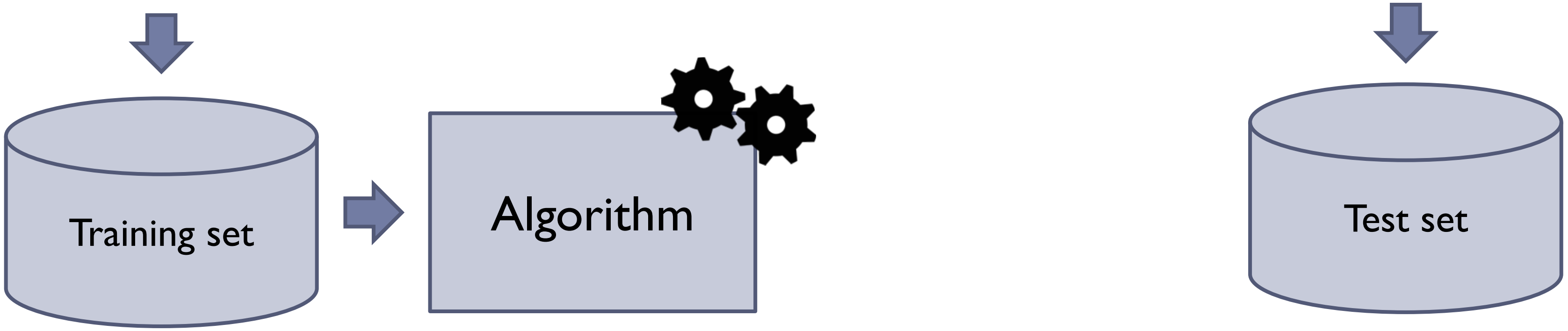
Re.	Status	Salary	Fraud
No	Single	75,000€	?
Yes	Married	50,000€	?
No	Married	150,000€	?
Yes	Divorced	90,000€	?
…	…	…	…

Test set

Classification

Re.	Status	Salary	Fraud
Yes	Single	125,000€	No
No	Married	100,000€	No
No	Single	70,000€	No
Yes	Married	120,000€	No
No	Divorced	95,000€	Yes
No	Married	60,000€	No
…	…	…	…

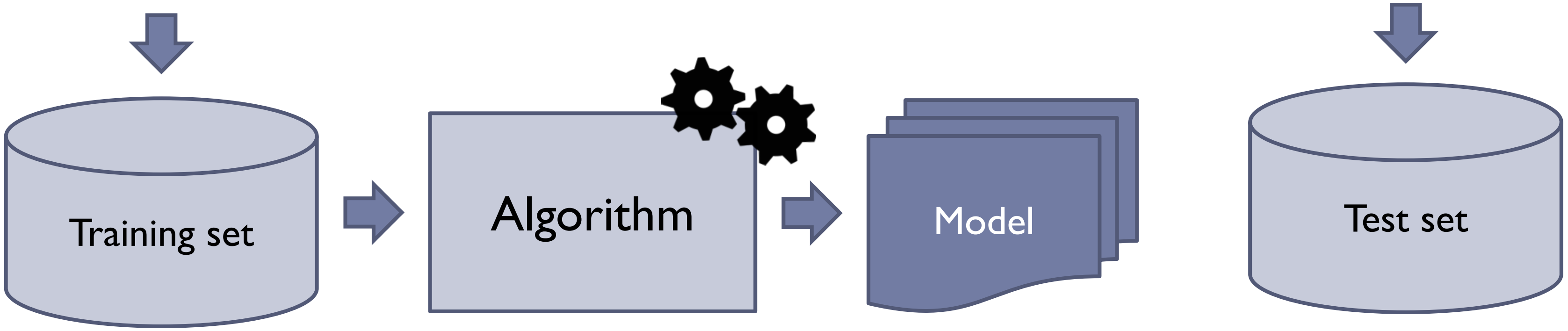
Re.	Status	Salary	Fraud
No	Single	75,000€	?
Yes	Married	50,000€	?
No	Married	150,000€	?
Yes	Divorced	90,000€	?
…	…	…	…



Classification

Re.	Status	Salary	Fraud
Yes	Single	125,000€	No
No	Married	100,000€	No
No	Single	70,000€	No
Yes	Married	120,000€	No
No	Divorced	95,000€	Yes
No	Married	60,000€	No
…	…	…	…

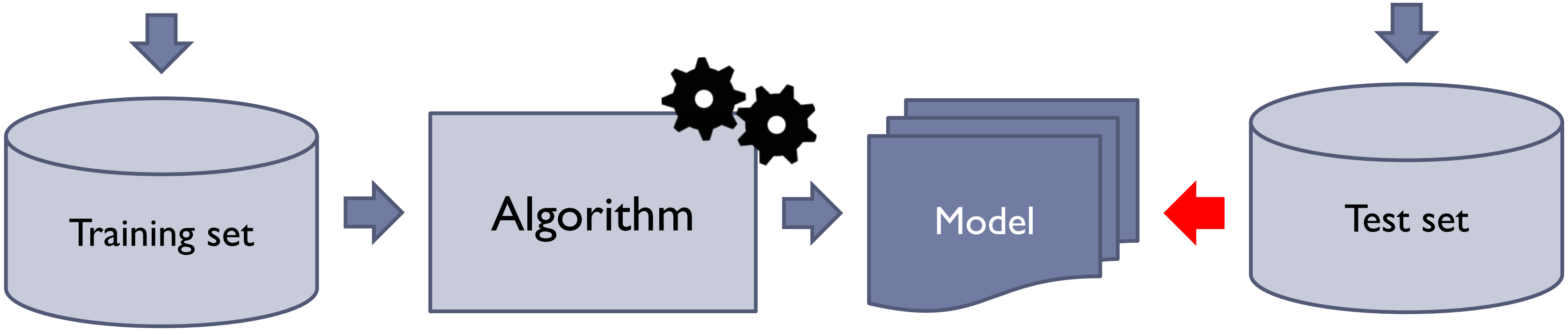
Re.	Status	Salary	Fraud
No	Single	75,000€	?
Yes	Married	50,000€	?
No	Married	150,000€	?
Yes	Divorced	90,000€	?
…	…	…	…



Classification

Re.	Status	Salary	Fraud
Yes	Single	125,000€	No
No	Married	100,000€	No
No	Single	70,000€	No
Yes	Married	120,000€	No
No	Divorced	95,000€	Yes
No	Married	60,000€	No
…	…	…	…

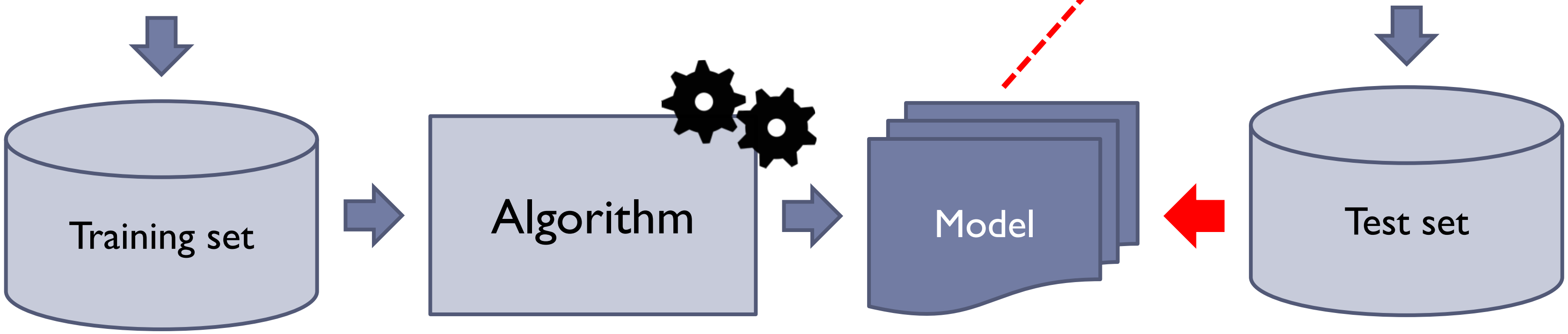
Re.	Status	Salary	Fraud
No	Single	75,000€	?
Yes	Married	50,000€	?
No	Married	150,000€	?
Yes	Divorced	90,000€	?
…	…	…	…



Classification

Re.	Status	Salary	Fraud
Yes	Single	125,000€	No
No	Married	100,000€	No
No	Single	70,000€	No
Yes	Married	120,000€	No
No	Divorced	95,000€	Yes
No	Married	60,000€	No
…	…	…	…

Re.	Status	Salary	Fraud
No	Single	75,000€	No
Yes	Married	50,000€	Yes
No	Married	150,000€	No
Yes	Divorced	90,000€	No
…	…	…	…



Classification

- Example applications
  - Direct marketing
    - Objective
      - □ Reduce the cost of sending mail by selecting a set of customers who are candidates to buy a new model of mobile phone
    - Approximation
      - □ Use data from a previously existing similar product
      - □ We know which customers bought it and who didn't
      - □ The decision `{buy, not buy}` constitutes the *class* attribute we want to predict
      - □ Collect demographic, lifestyle, business type, salary, etc. information for each potential customer
      - □ Use that information as input features to train the classifier

Classification

- Example applications
  - Customer loyalty
    - Objective
      - □ Predict when a company may lose a customer
    - Approximation
      - □ Use instances of past and present customer transactions
      - □ How often the client calls, where they calls, at what time of day, economic situation, marital status, etc.
      - □ Label customers as `{loyal, disloyal}` (this will be the *class*)
      - □ Find a model for predicting customer loyalty

Classification

Let’s practice!

https://teachablemachine.withgoogle.com/

Classification

- Algorithms
  - There are numerous classification algorithms
    - Some work better for certain tasks
    - Depending on the number of instances
    - Depending on the number of features
  - Types
    - Naïve Bayes
    - Decision trees
    - Neural networks
    - Example-based
    - Linear separators
    - …

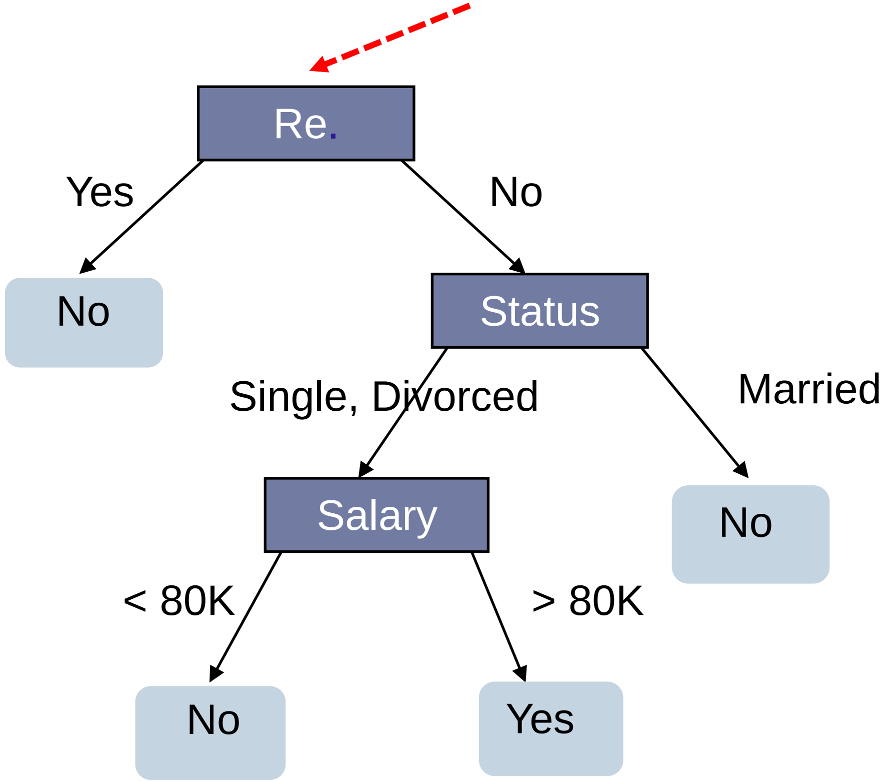
Classification

- Decision trees
  - Training phase
    - There are different algorithms to build these trees (models) from the training set
      - Hunt
      - CHAID
      - CART
      - ID3
      - C4.5
      - …

Classification

- Decision trees
  - Prediction phase

Start in the root

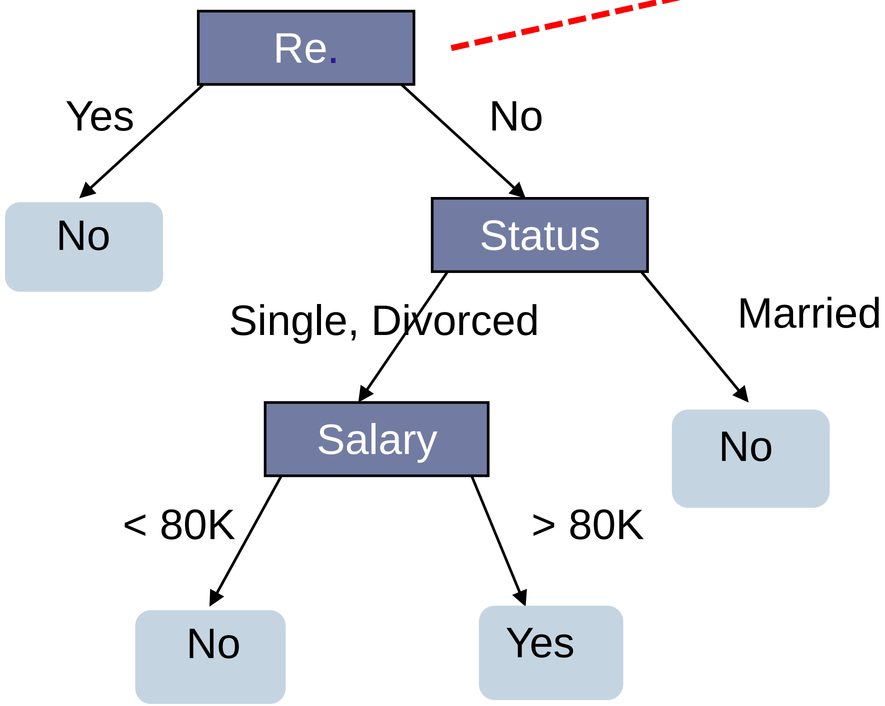


New instance

Re.	Status	Salary	Fraud
No	Married	80.000€	?

Classification

- Decision trees
  - Prediction phase



New instance

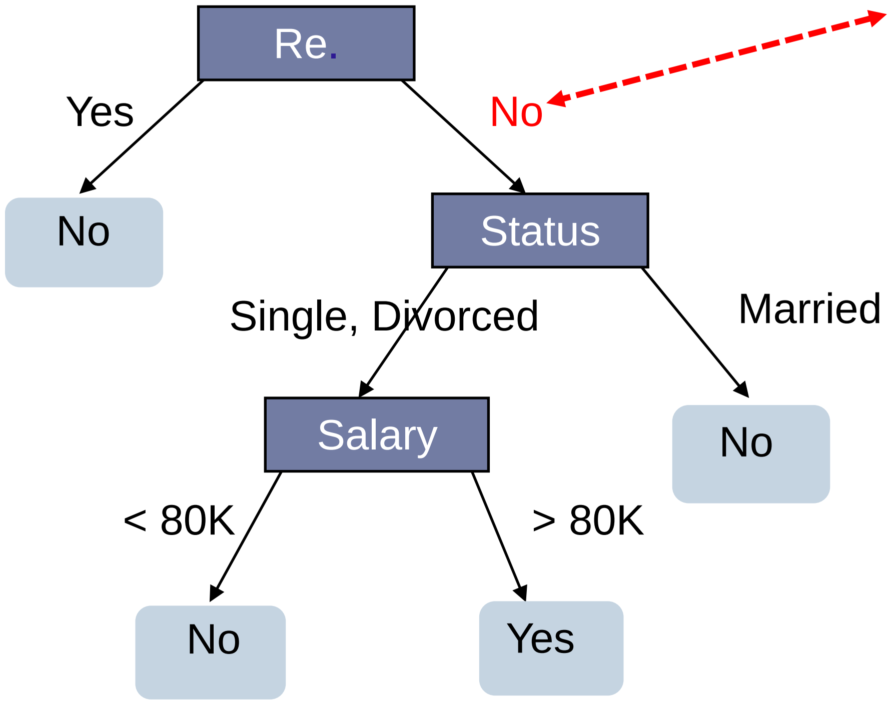
Re.	Status	Salary	Fraud
No	Married	80.000€	?

Classification

- Decision trees
  - Prediction phase

New instance

Re.	Status	Salary	Fraud
No	Married	80.000€	?

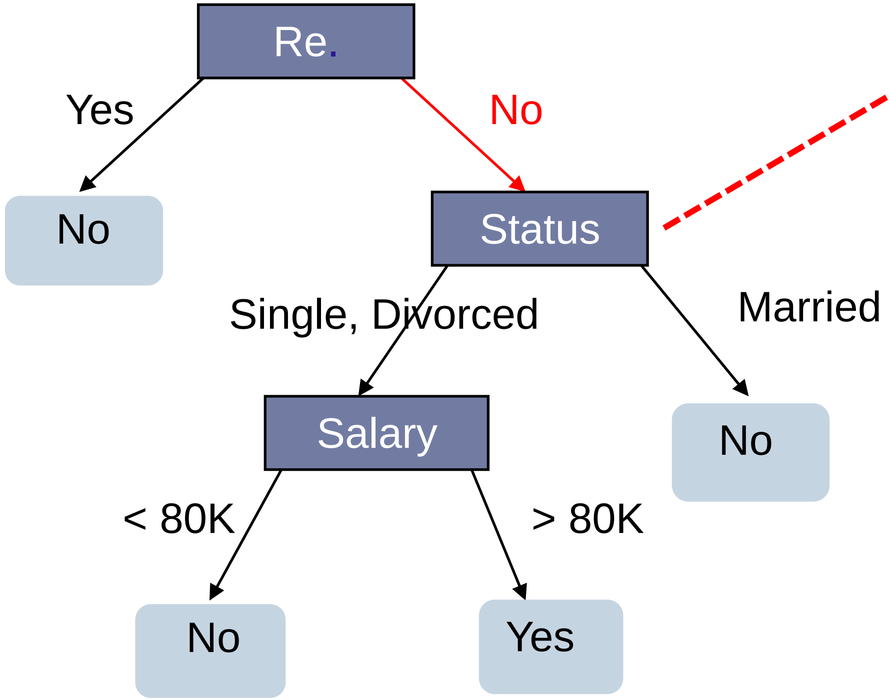


Classification

- Decision trees
  - Prediction phase

New instance

Re.	Status	Salary	Fraud
No	Married	80.000€	?

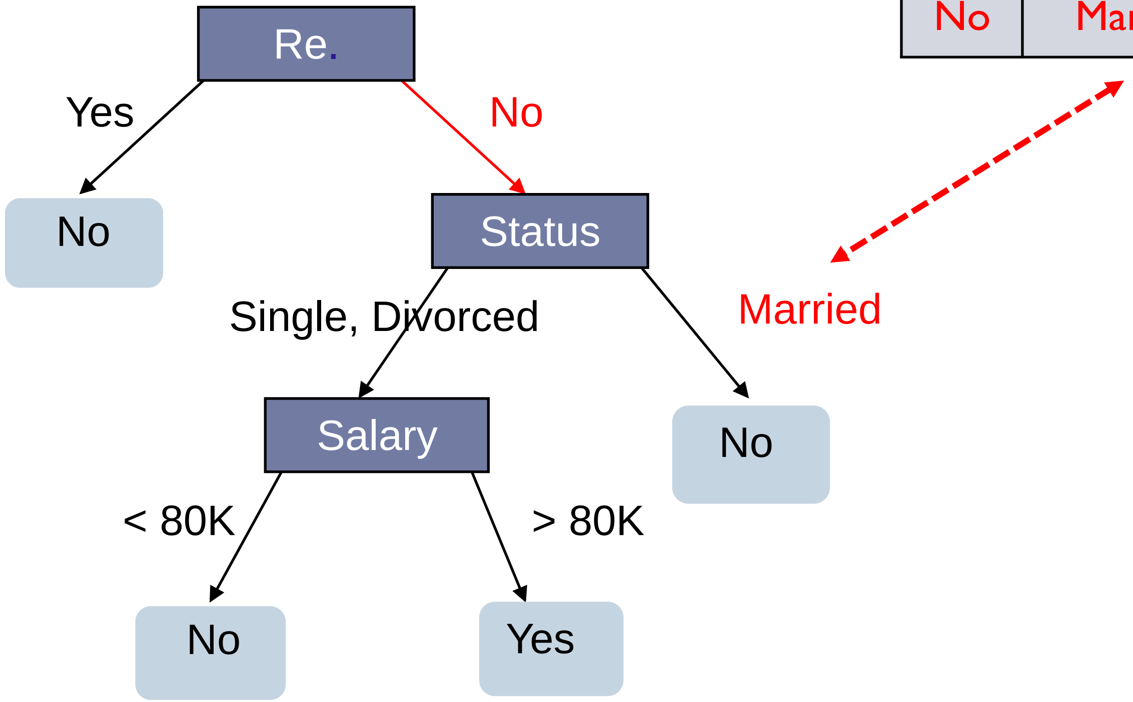


Classification

- Decision trees
  - Prediction phase

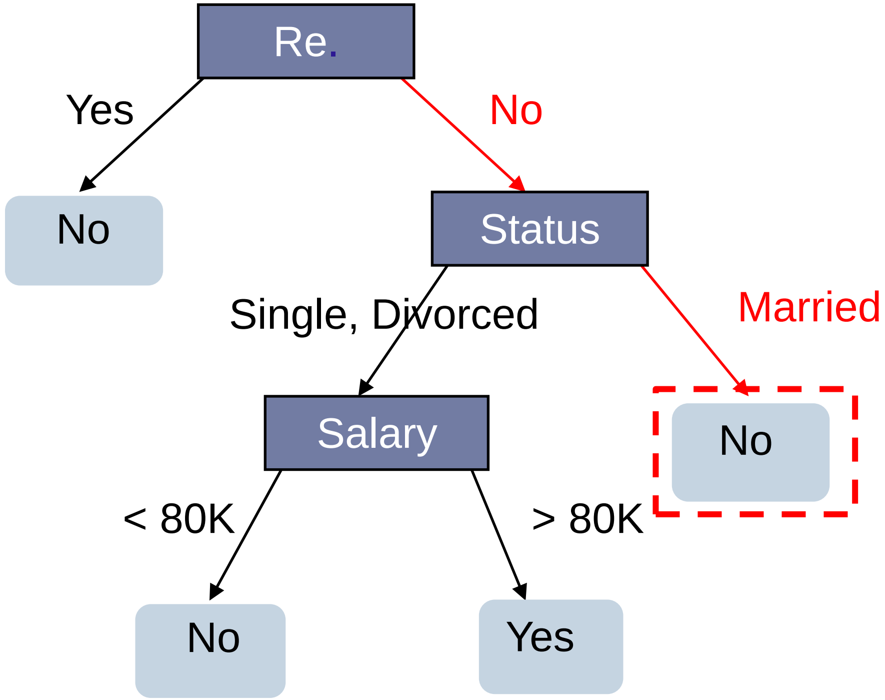
New instance

Re.	Status	Salary	Fraud
No	Married	80.000€	?



Classification

- Decision trees
  - Prediction phase



New instance

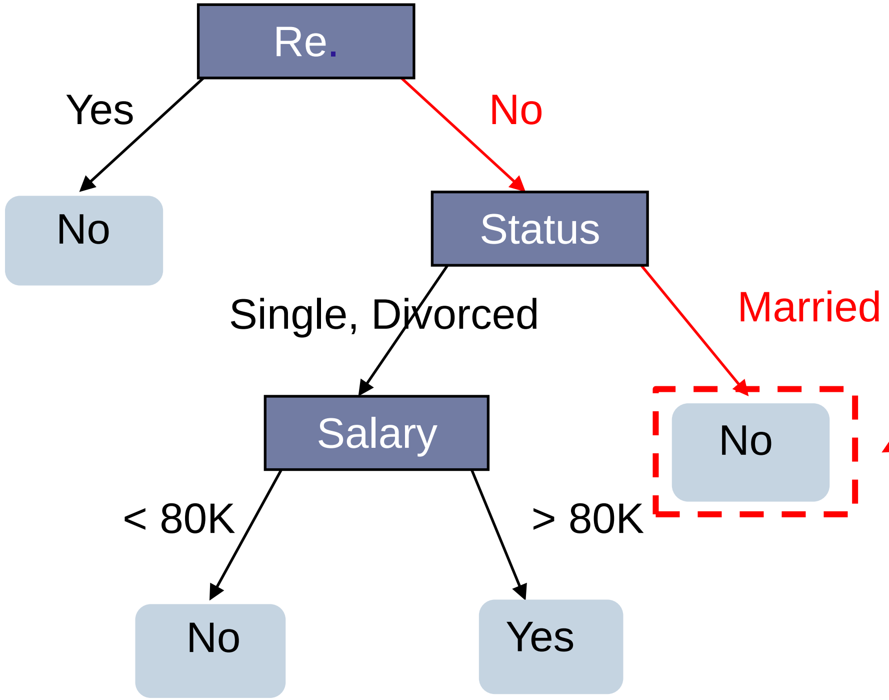
Re.	Status	Salary	Fraud
No	Married	80.000€	?

Classification

- Decision trees
  - Prediction phase

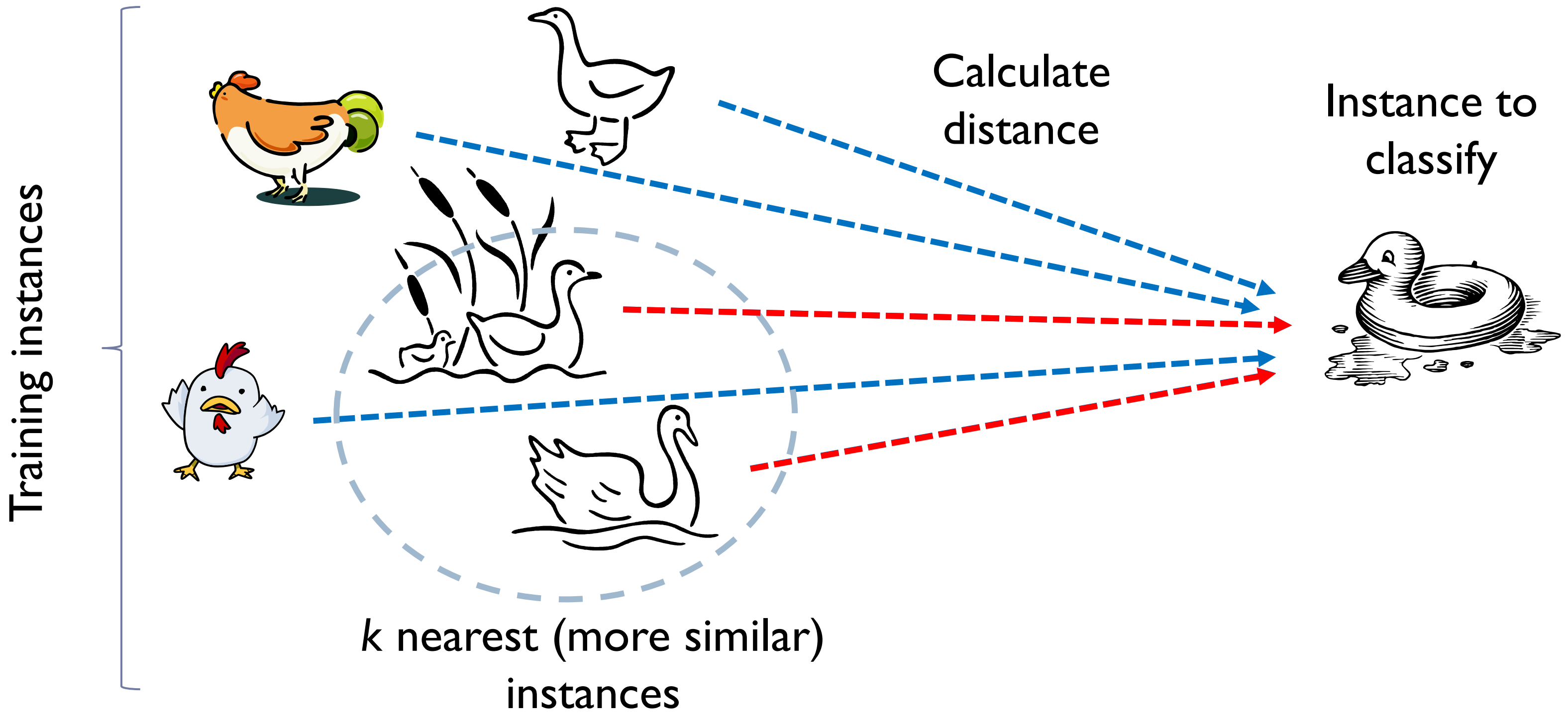
New instance

Re.	Status	Salary	Fraud
No	Married	80.000€	No



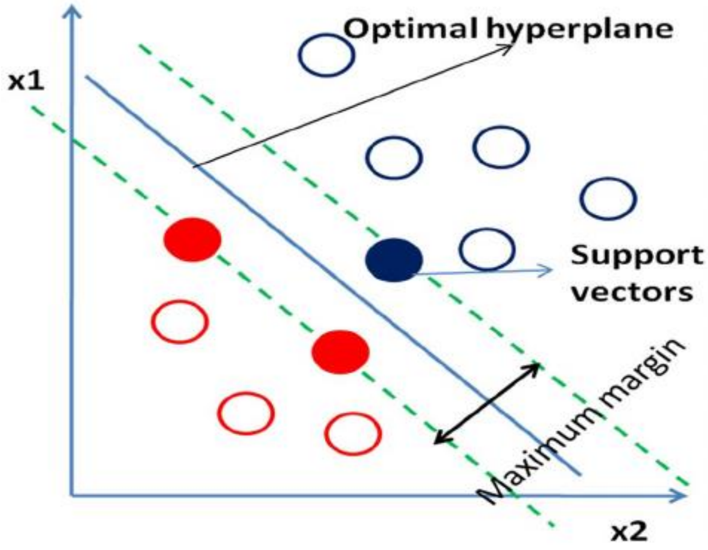
Classification

- *k-Nearest Neighbors* (k-NN)
  - Intuitive idea: "*If it walks like a duck and quacks like a duck, it's probably a duck*"



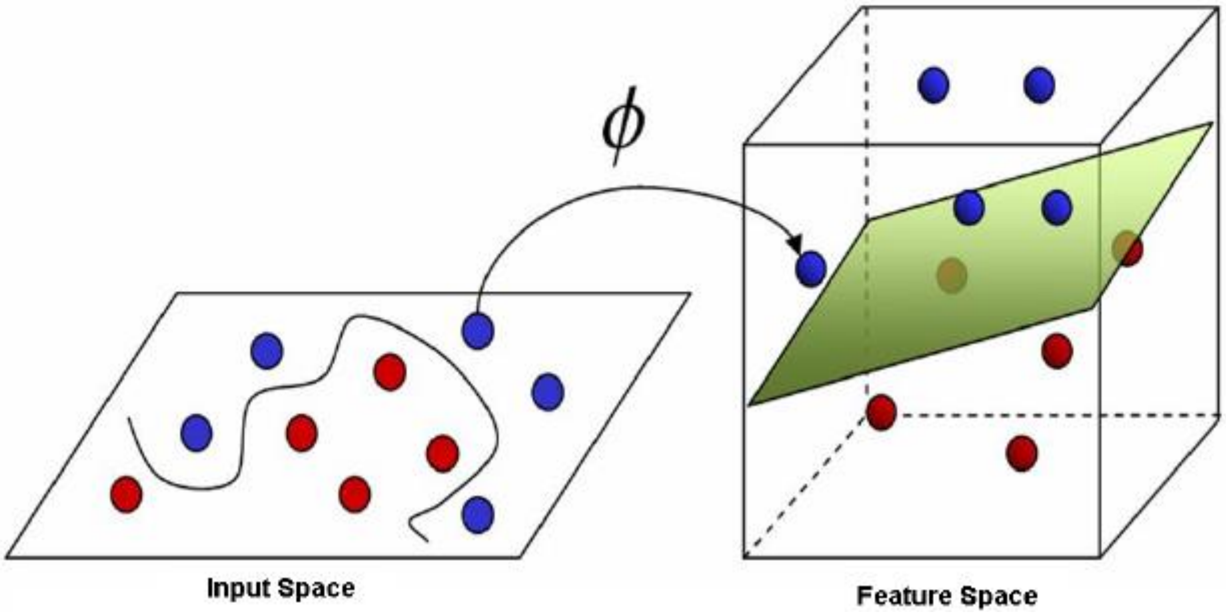
Classification

- Support Vector Machines (SVM)
  - Linear classifiers provide high performance
  - Work well in high dimensional spaces
  - Slow building the model but fast classifying
  - Find the optimal hyperplane (boundary) that maximizes the margin between two classes



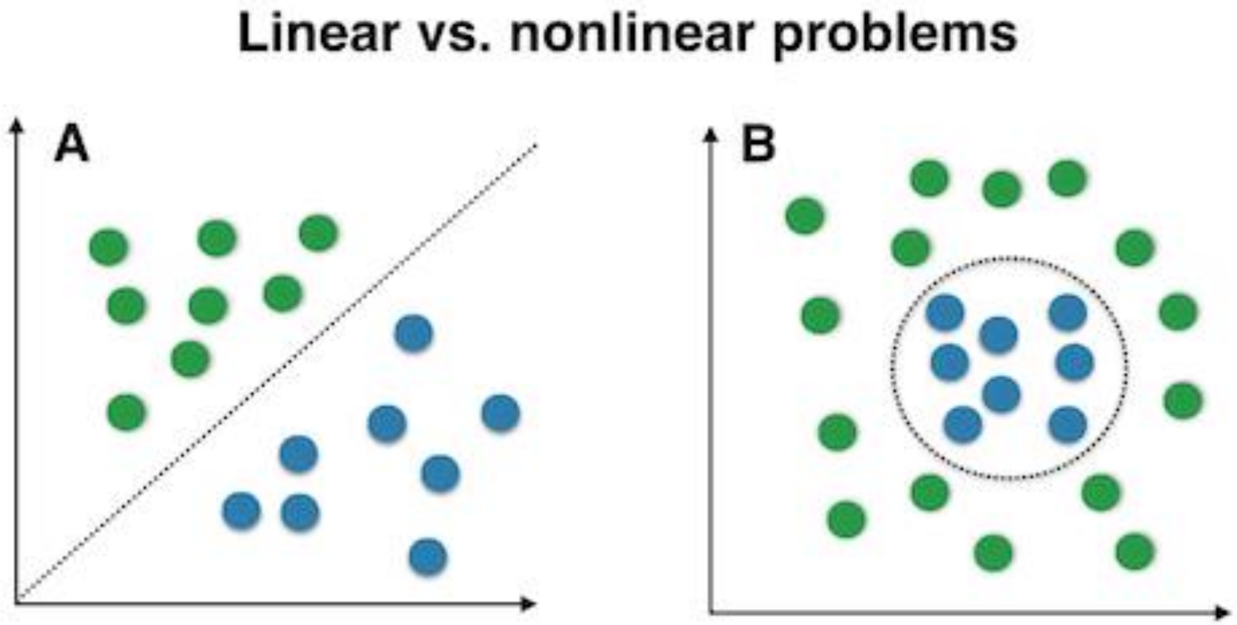
Classification

- Support Vector Machines (SVM)
  - If the dataset is not linearly separable, the algorithm can be extended by means of non linear transformations ∅(x) to a new feature space



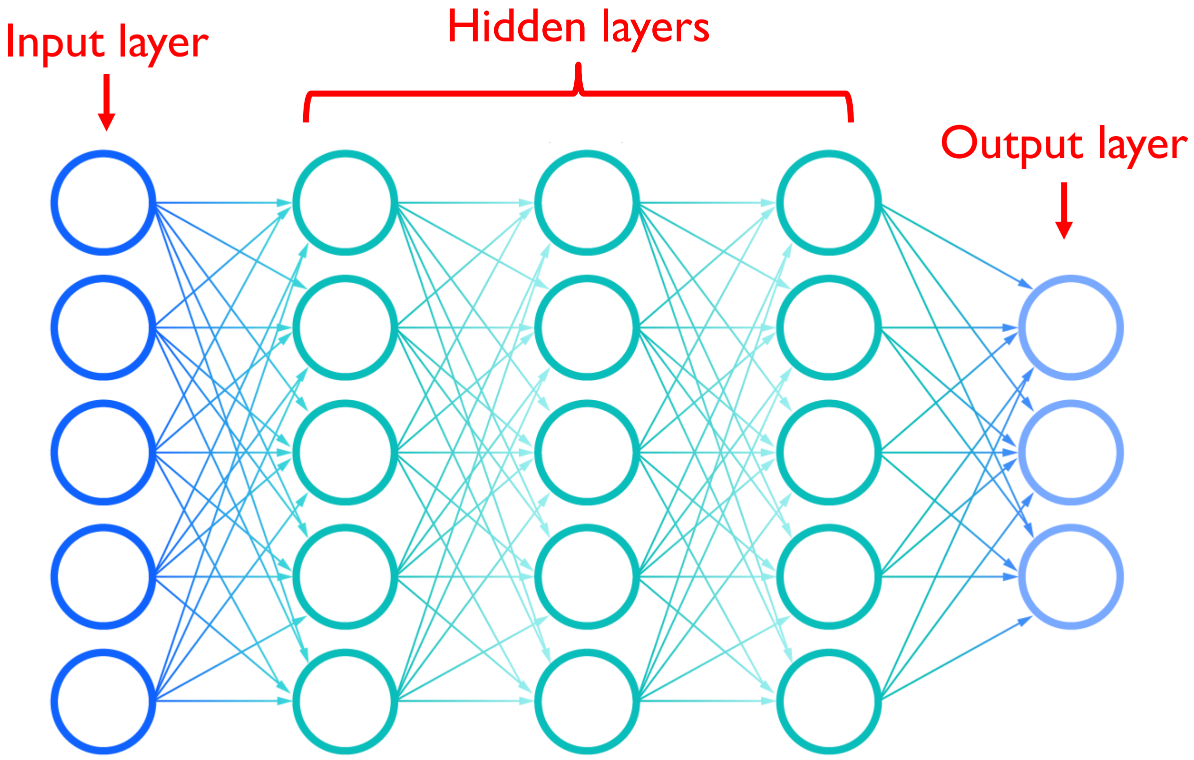
Classification

- Neural Networks
  - Separate samples in a multidimensional space



Classification

- Neural Networks
  - When there are multiple hidden layers, we talk about Deep Neural Networks (a.k.a. *Deep Learning*)



Classification

Let’s practice!

https://playground.tensorflow.org/

Contents

- What is machine learning?
- Components
  - Data
  - Features
  - Algorithms
- Classification
- **Regression**
- Evaluation

Regression

- Works with labelled data (such as *classification*)
- The class attribute is of continuous type
  - Contains a set of numeric values
  - E.g., estimated price of a house, of a share, ...
- We must find a model for the *class* attribute based on the values of the other attributes
- The goal is to predict the value of the class continuous attribute for previously unseen instances

Regression

- Uses the same datasets as for classification
  - Training set
  - Validation set
  - Test set
- Algorithms
  - Multilayer perceptron
  - k-NN
  - Support Vector Machines (SVM)
  - Decission trees (M5P)
  - …

Regression

- Application examples
  - Predict the number of sales of a new product based on advertising expenses
  - Predict wind speed as a function of temperature, humidity, air pressure, etc.
  - Time series prediction in stock market indices

Contents

- What is machine learning?
- Components
  - Data
  - Features
  - Algorithms
- Classification
- Regression
- **Evaluation**

Evaluation

- The main reason to build a classifier is to learn to classify previously unseen (*unlabelled*) instances
- The most obvious criteria to estimate the performance of the classifier is accuracy
  - Proportion of new instances correctly classified

$$c = r/n$$

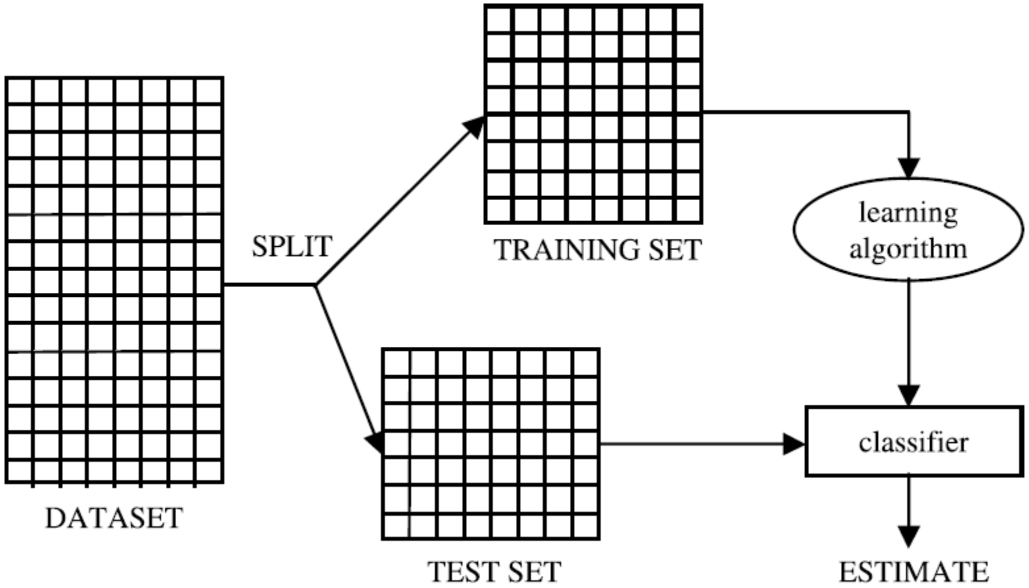
c: classification accuracy
r: number of test documents correctly classified
n: total number of test documents

Evaluation

- In most domains the number of new samples is potentially huge
  - E.g. weather forecast for every possible future day
- Estimate the predicting capability of the classifier by measuring the precision for a set of samples not used during the training process
- Three strategies
  - Train and test set
  - K-fold cross validation
  - Leave-one-out

Evaluation

- Train and test set
  - Data is split in two sets: training and test
    - E.g., 80% training and 20% test
  - The training set is used to build the model (classifier)
  - The test set is used to predict the performance of the model

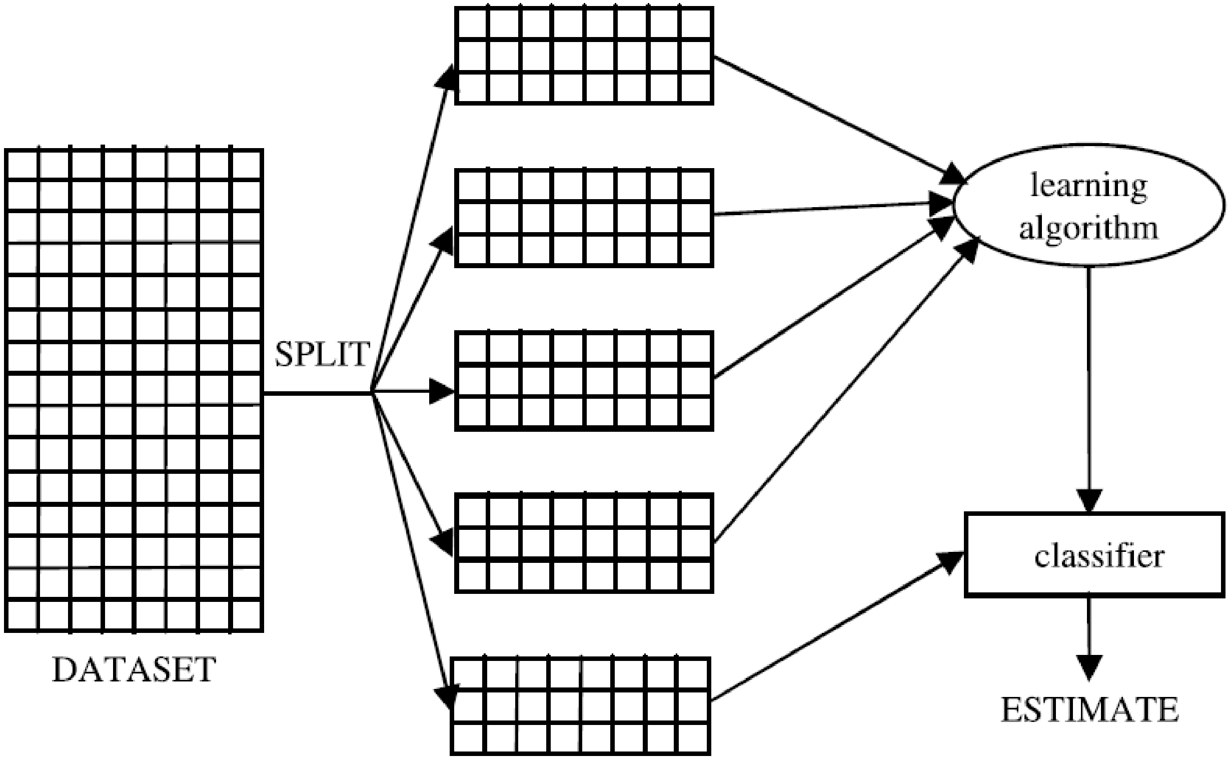


Evaluation

- K-fold cross validation
  - Used when the number of instances is small and do not want to split into training and test sets
  - *N* instances are divided in *k* equal groups
  - Typically *k=5* or *k=10*
  - Generates *k* different classifiers
  - Each one uses *1* fold as test and *k-1* as training
  - Performance is given by the total number of correct answers in the *k* iterations divided by the total number of instances

Evaluation

- K-fold cross validation

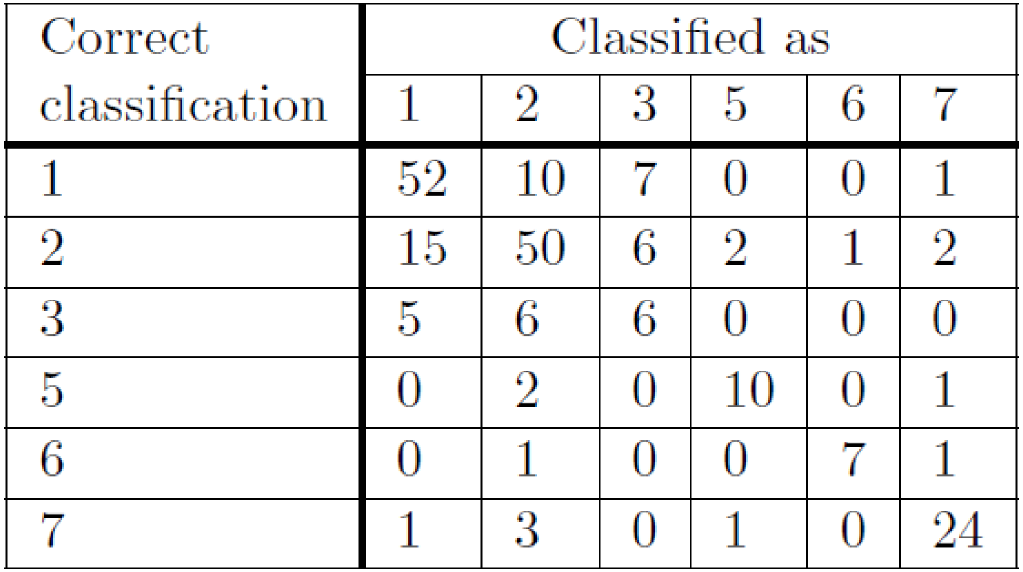


Evaluation

- Leave-one-out
  - A.k.a. n-fold cross validation
  - “Extreme” case of cross validation
  - The dataset is divided in as many sets as instances (*N*)
  - *N* classifiers are generated, each one trained on *N-1* samples and evaluated in the remaining one
  - The computational cost is huge for large amounts of data
  - The classifier performance is given by the total number of correct answers in the *N* iterations divided by the total number of instances

Evaluation

- Confusion matrix
  - Shows the performance of the classifier per class
  - Shows how frequent *class X* is correctly labelled or was confused with *class Y*



Correct classification	Classified as					
	1	2	3	5	6	7
1	52	10	7	0	0	1
2	15	50	6	2	1	2
3	5	6	6	0	0	0
5	0	2	0	10	0	1
6	0	1	0	0	7	1
7	1	3	0	1	0	24

Evaluation

Let's practice!

https://bit.ly/3Ndw4fR

@d_tomas

David Tomás Díaz