Brief review of ML basics

Jorge Acosta Hernández
jorge.acosta@upm.es

With some slides from Jesús Montes

Dec. 2024

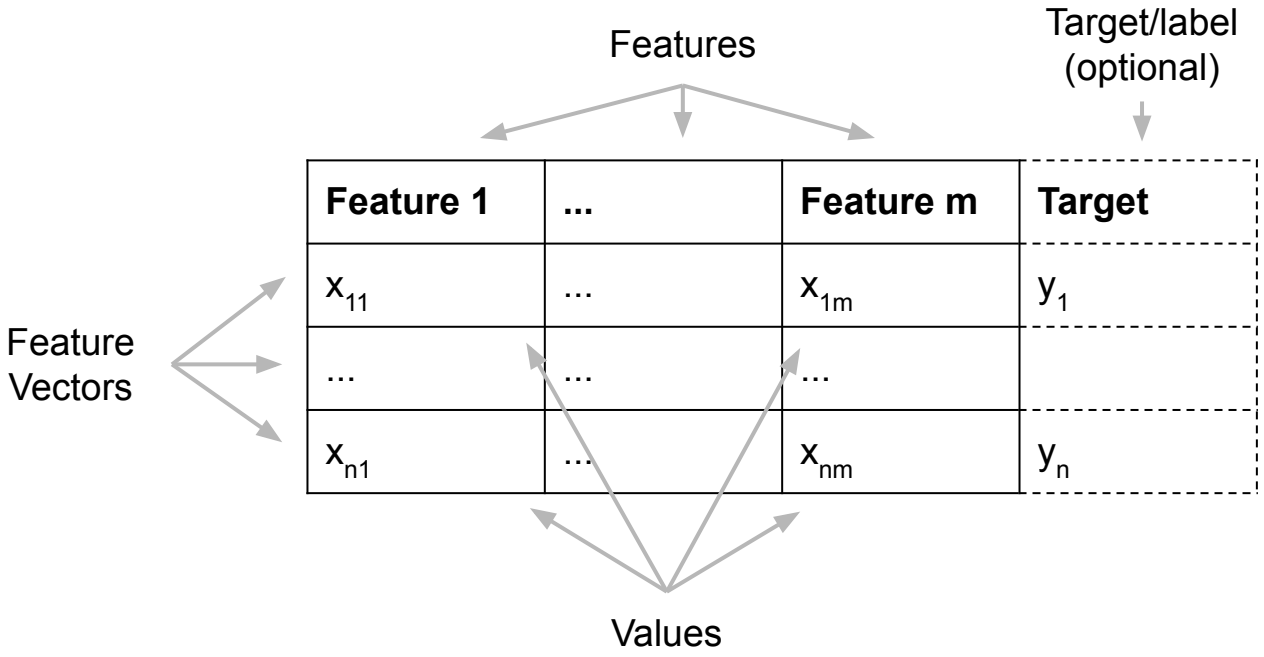
What is Machine Learning? (just in case...)

- Related to/part of:
  - Mathematics (Linear Algebra, Differential Calculus, Statistics, Optimization...)
  - Data mining
  - Artificial Intelligence (AI)
- Providing machines with new, advanced capabilities.
- Some capabilities:
  - Database analysis
  - Autonomous driving
  - Natural language processing
  - Recommendation systems
  - Medical Diagnosis
  - Predictive Analysis

What is Machine Learning? (just in case...)

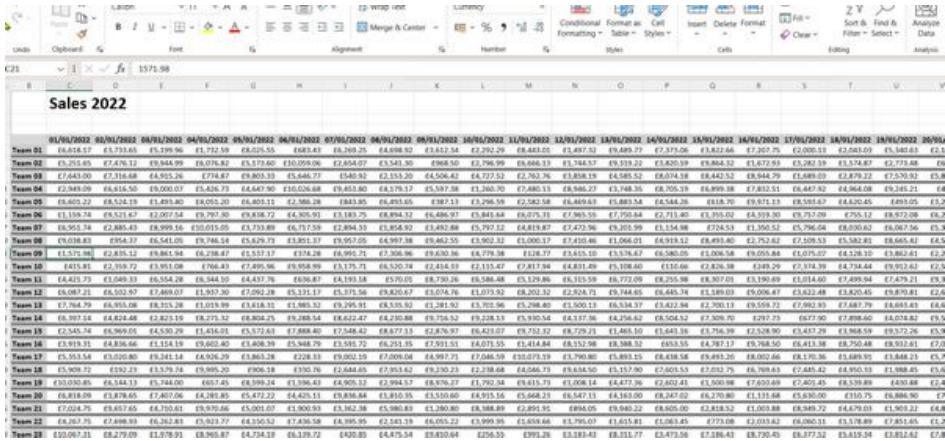
- The subfield of computer science that "gives computers the ability to learn without being explicitly programmed" (Arthur Samuel, 1959).
- A more formal definition: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." (Tom M. Mitchell, 1998).
  - Experience (E)
  - Task (T)
  - Performance (P)
- Can you identify E, T and P in a typical automated SPAM filter problem?

Typical input data in Machine Learning



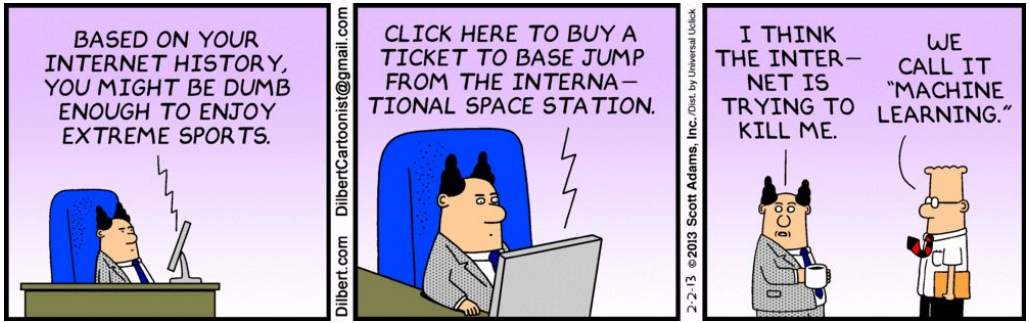
Typical input data in Machine Learning

Year	City	Amount
1990	New York City	$1,123,456.00
1995-96		2.2 mil
2000s	NYC	No data
2020	New_York	50000000+



Machine Learning techniques

- Supervised learning
- Unsupervised learning
- Other:
  - Semi-supervised learning
  - Reinforcement learning
  - Self-Supervised learning
  - ...



Supervised Learning

Extract knowledge from data that contains a target variable.

Once the learning process takes place the system will be able to predict the expected value of the target variable (more or less accurately).

Types of supervised learning:

- Regression
  - Continuous target variable
  - Examples: stock market values, temperature prediction...
  - Techniques: Generalized Linear Models, Artificial Neural Networks...
- Classification
  - Discrete/categorical target variable
  - Examples: Disease diagnosis, error detection in production systems...
  - Techniques: Logistic Regression, Decision Trees, Naïve-Bayes...

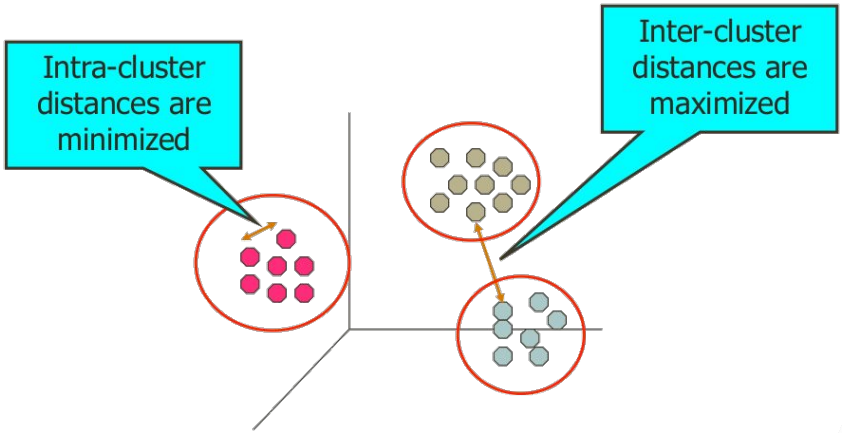
Unsupervised Learning

Extract knowledge from data without a target variable available.

The learning result is new knowledge about the organization of the information contained in the initial data.

Most typical case is **clustering** (a.k.a. unsupervised classification).

Typical techniques: K-Means, Expectation Maximization, Hierarchical clustering, Density-based clustering...



Evaluating the learning process

"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, 1998).

To evaluate T we need to define the performance measure P.

Ideally, P should be a numerical measurement that allows objective evaluation of the learning process.

Regression:

- Mean squared error
- Coefficient of determination (R^2)

Classification:

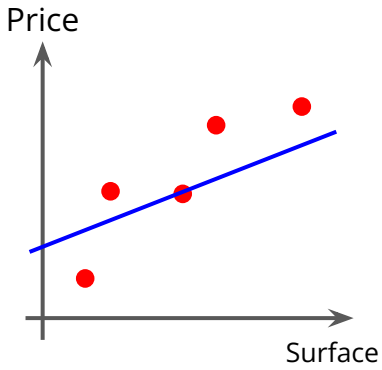
- Correctly classified percentage
- True pos/neg + False pos/neg
- Precision + Recall

Clustering:

- Level of agreement (Rand index...)

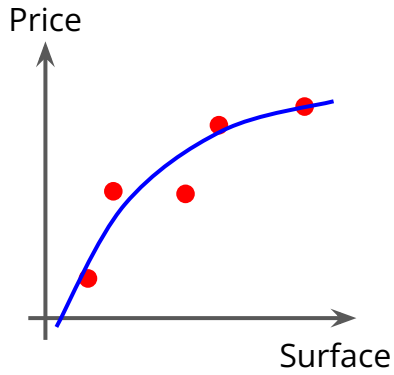
Learning and evaluation

High bias



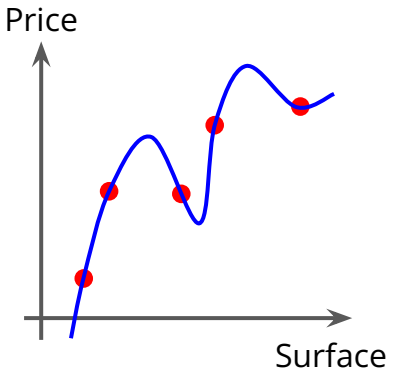
- The model is too simple
- High prediction error
- **Underfitting**

OK



- Prediction within acceptable margins

High variance



- The model is too complex
- High prediction error
- **Overfitting**

Learning and evaluation

Typically, the performance measure is also used during the learning process, to guide it.

- At the end of the learning process, we obtain a performance measurement for the data used during the training (training dataset).

WARNING: **overfitting**

- The model created could only be able to correctly predict the values it has already “seen” (the training dataset).

We need (at least) two datasets with the same structure:

- Training set: Used to train the model.
- Test set: Used to evaluate the model.

The test set is only used at the end, and the resulting value of the performance measure is the final performance of our model.

The test set must **NEVER** be used for training.

Machine Learning with Spark

Data Science and Big Data are strongly related fields.
So far, we have learned how Spark can be used to load and process data:

- Data load (Spark Core and/or Spark SQL)
- Processing/cleaning (Spark Core and/or Spark SQL)

but Spark can be used for many steps of a Data Science project:

- Model training (**MLlib)**
- Model evaluation (**MLlib)**

Machine Learning in the Spark Stack