Machine Learning in the context of Big Data

Luis Fernando Lago Fernández

EPS - UAM

CÁTEDRA UAM/IBM

GRANDES SISTEMAS, SOLUCIONES COGNITIVAS Y ANALÍTICA BIG DATA

www.catedrauamibm.com

Outline

CÁTEDRA UAM/IBM
GRANDES SISTEMAS, SOLUCIONES COGNITIVAS Y ANALÍTICA BIG DATA

- Introduction to ML in the context of Big Data
- The ML design cycle

PART 1: MACHINE LEARNING IN THE CONTEXT OF BIG DATA

Overview

CÁTEDRA UAM/IBM
GRANDES SISTEMAS, SOLUCIONES COGNITIVAS Y ANALÍTICA BIG DATA

- **Big Data** = **Data** + **Analytics**
- What do we mean by **"Big"**?
  - Volume
  - Velocity
  - Variety
  - But also, veracity and value
- **Predictive Analytics** uses data to make future predictions with the help of **Machine Learning** algorithms

Is Big Data just data?

CÁTEDRA UAM/IBM
GRANDES SISTEMAS, SOLUCIONES COGNITIVAS Y ANALÍTICA BIG DATA

Without analytics Big Data is just data

Data Analysis and Data Analytics

- **Data analysis** is the process of examining data to find facts, relationships, patterns, insights and/or trends. The overall goal is to suport better decision making.
- **Data analytics** is a broader term than encompasses data analysis. It includes the management of the complete data lifecycle: collecting, cleansing, organizing, storing, analyzing and governing

(From *Big Data Fundamentals: Concepts, Drivers and Techniques*, Prentice Hall, 2016)

Data Analytics

CÁTEDRA UAM/IBM
GRANDES SISTEMAS, SOLUCIONES COGNITIVAS Y ANALÍTICA BIG DATA

- Descriptive analytics
- Diagnostic analytics
- Predictive analytics
- Prescriptive analytics

Descriptive Analytics

CÁTEDRA UAM/IBM
GRANDES SISTEMAS, SOLUCIONES COGNITIVAS Y ANALÍTICA BIG DATA

Answer questions about events that have already ocurred

- What was the sales volume over the past 6 months?
- How many new customers in the last year?

Diagnostic Analytics

CÁTEDRA UAM/IBM
GRANDES SISTEMAS, SOLUCIONES COGNITIVAS Y ANALÍTICA BIG DATA

Determine the cause of something that happened in the past

- Why where sales in June less than sales in May?
- Why the demand for product A has decreased in the last 2 months?

Predictive Analytics

CÁTEDRA UAM/IBM
GRANDES SISTEMAS, SOLUCIONES COGNITIVAS Y ANALÍTICA BIG DATA

Determine the outcome of an event that might occur in the future.
Models generate future predictions based on past events

- If a customer has purchased products A and B, what are the chances that he will also purchase product C?
- What is the probability of my customers going to another company for the same service?

Prescriptive Analytics

CÁTEDRA UAM/IBM
GRANDES SISTEMAS, SOLUCIONES COGNITIVAS Y ANALÍTICA BIG DATA

Built upon the results of predictive analytics by prescribing actions that should be taken.

- Which products should I offer to customers that have purchased products A and B?
- Which actions should I take to keep my customers?

Predictive Analytics and Machine Learning

CÁTEDRA UAM/IBM
GRANDES SISTEMAS, SOLUCIONES COGNITIVAS Y ANALÍTICA BIG DATA

In this course we will focus on Predictive Analytics and Machine Learning

- How to build models to extract useful information from data
- How to learn from data in order to make predictions

Machine Learning

CÁTEDRA UAM/IBM
GRANDES SISTEMAS, SOLUCIONES COGNITIVAS Y ANALÍTICA BIG DATA

"Field of study that gives computers the ability to learn without being explicitly programmed."

(Attributed to A. Samuel, 1959)

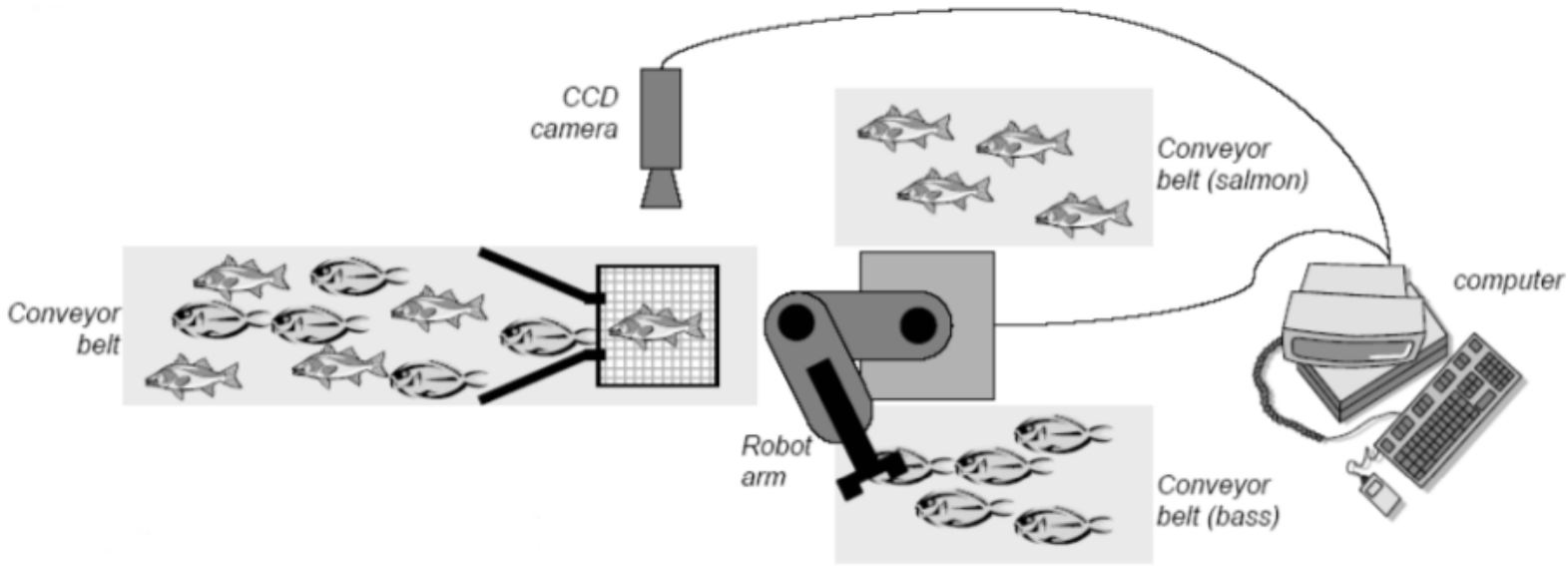
"Machine learning explores the study and construction of algorithms that can learn from and make predictions on data."

(From Wikipedia: https://en.wikipedia.org/wiki/Machine_learning)

Machine Learning - An example

CÁTEDRA UAM/IBM
GRANDES SISTEMAS, SOLUCIONES COGNITIVAS Y ANALÍTICA BIG DATA

- A fishing company wants to automate the process of separation of fish (salmon vs sea bass), using images recorded by a CCD camera

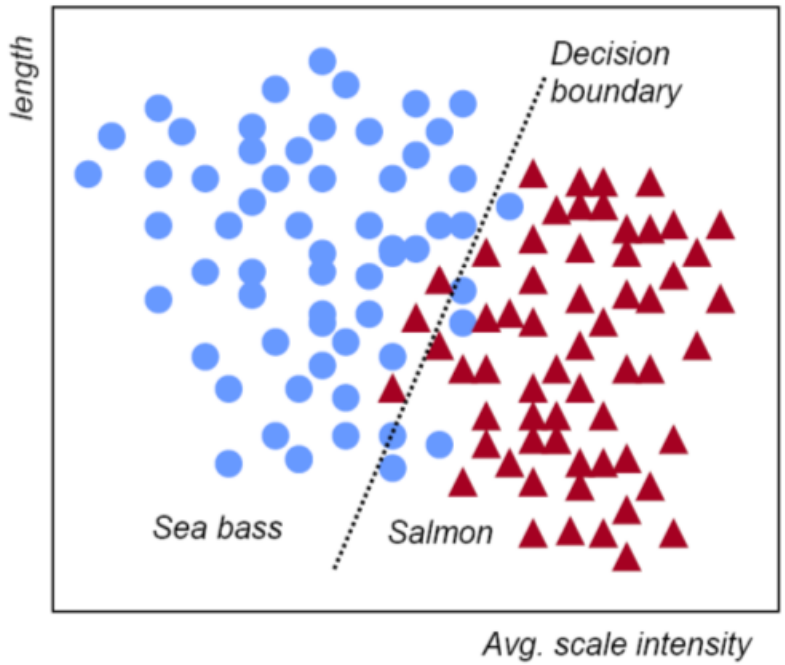


(From Duda, Hart and Stork, *Pattern Classification*, 2001)

Machine Learning - An example

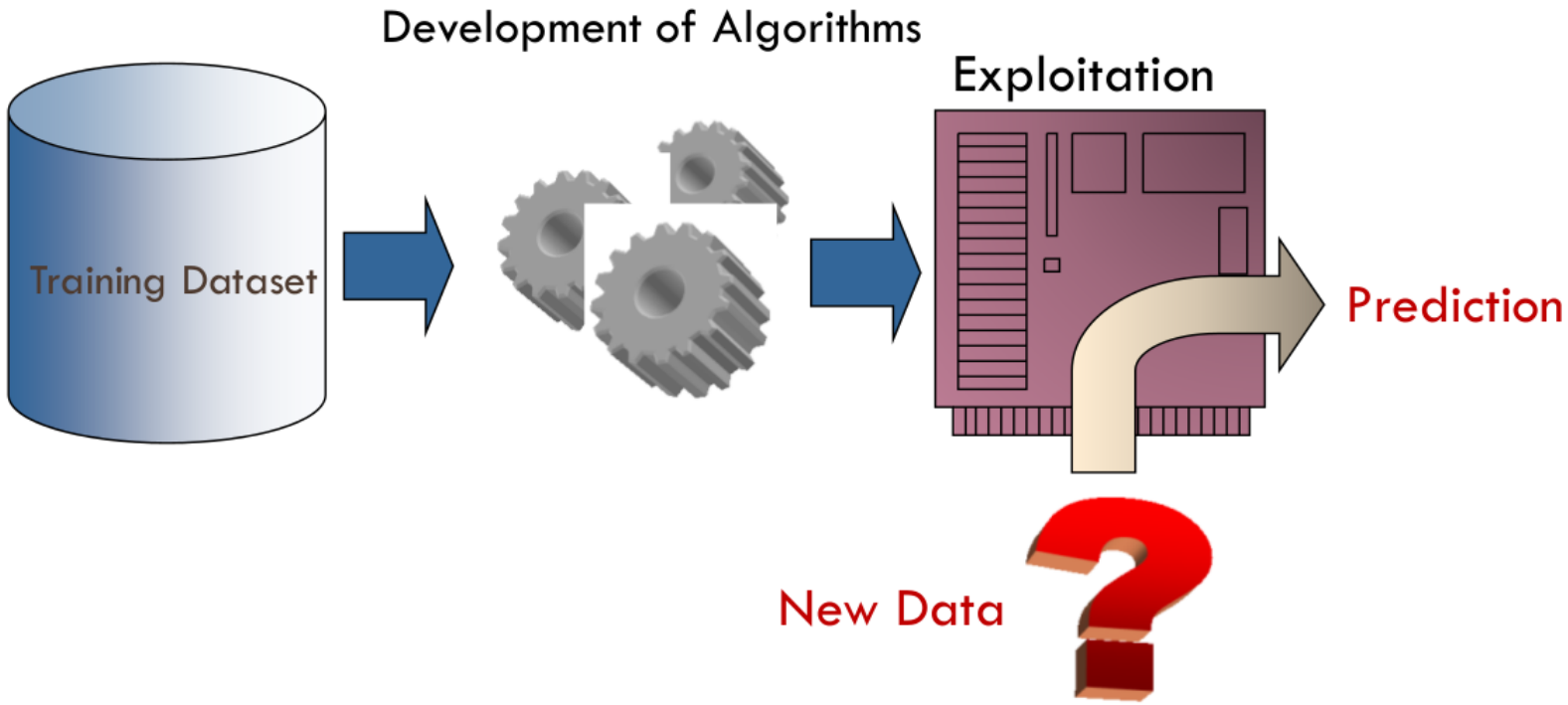
CÁTEDRA UAM/IBM
GRANDES SISTEMAS, SOLUCIONES COGNITIVAS Y ANALÍTICA BIG DATA

- Combining fish length and scale intensity they build the following linear classifier, which achieves a $95{,}7\,\%$ classification accuracy on the *training* data
- What do you expect to happen on new, unlabeled data?
- Will this model make good predictions?



(From Duda, Hart and Stork, *Pattern Classification*, 2001)

The Machine Learning design cycle



Machine Learning types

CÁTEDRA UAM/IBM
GRANDES SISTEMAS, SOLUCIONES COGNITIVAS Y ANALÍTICA BIG DATA

- Supervised learning
- Unsupervised learning
- (Semi-supervised learning)
- Reinforcement learning

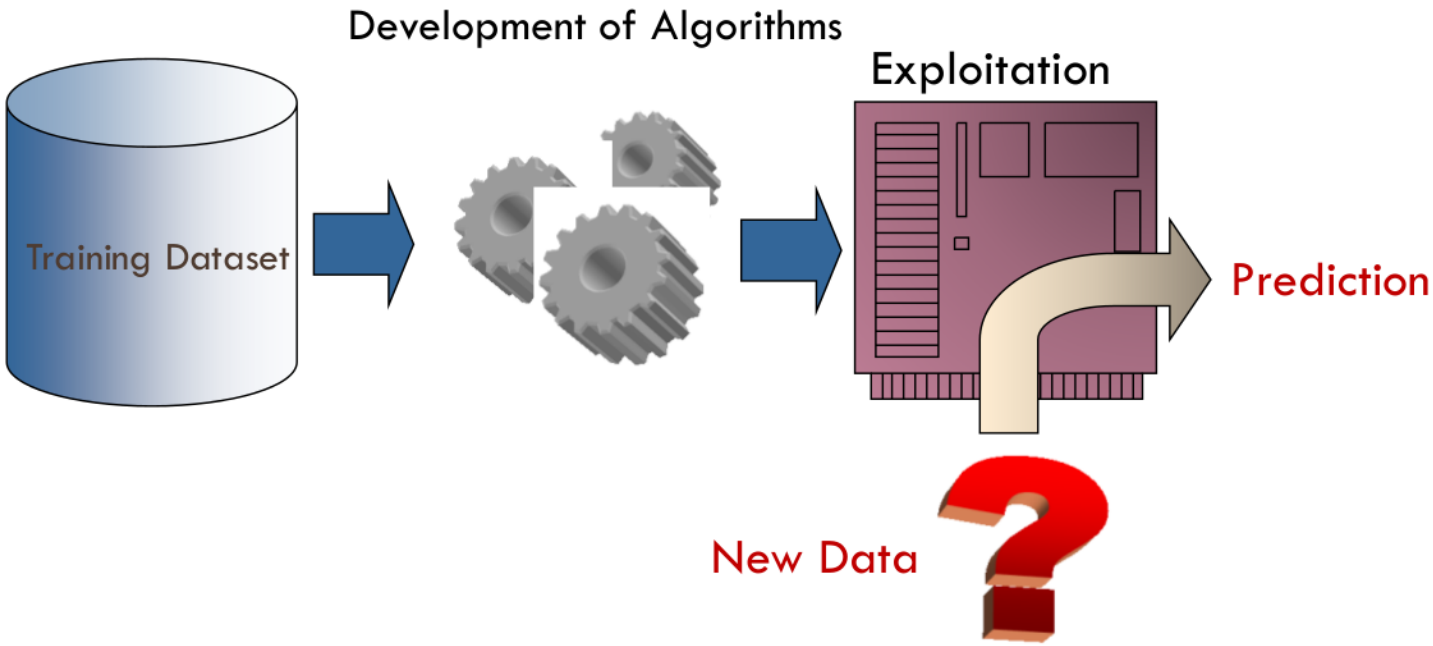
PART 2: THE ML DESIGN CYCLE

Outline

CÁTEDRA UAM/IBM
GRANDES SISTEMAS, SOLUCIONES COGNITIVAS Y ANALÍTICA BIG DATA

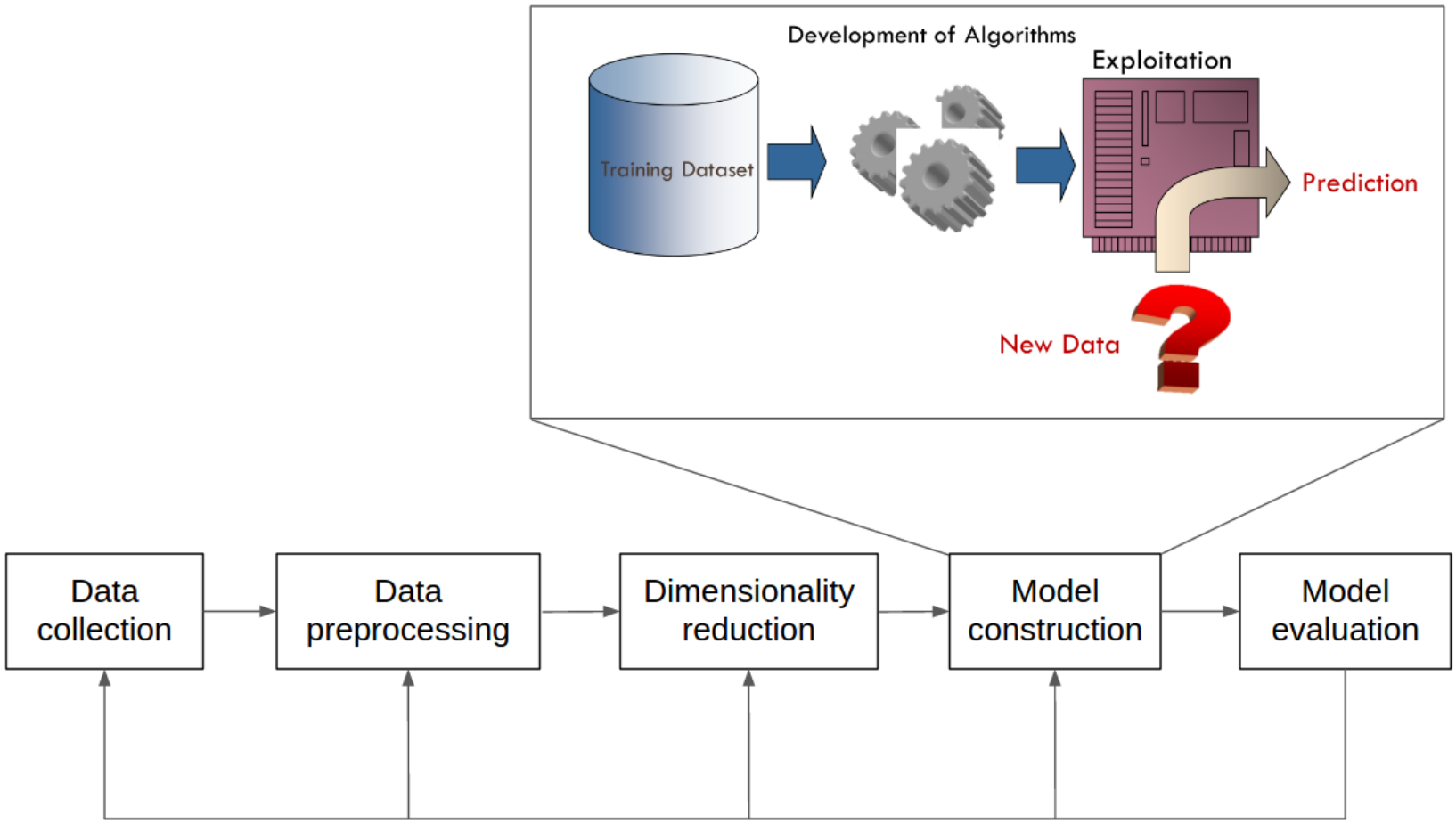
- Overview of the ML design cycle
- Pattern classification
- Model evaluation and selection
- Data preparation and audit
- Attribute selection and dimensionality reduction

The Machine Learning design cycle



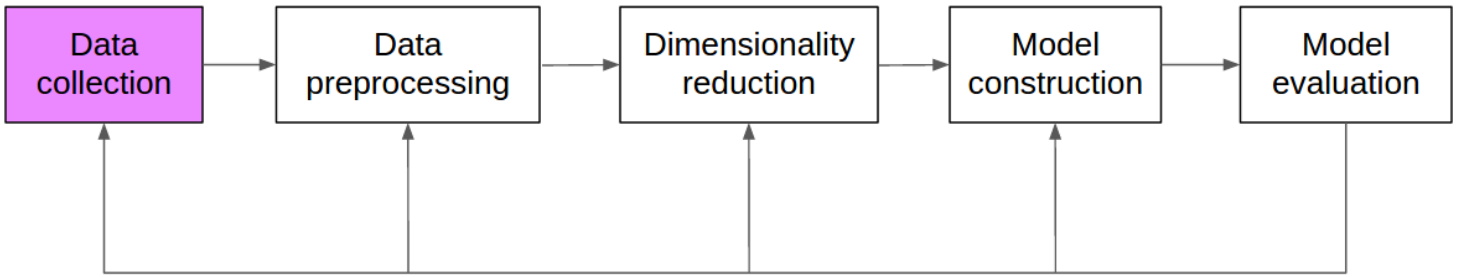
- Building the model is just one of the steps

The Machine Learning design cycle



Data collection

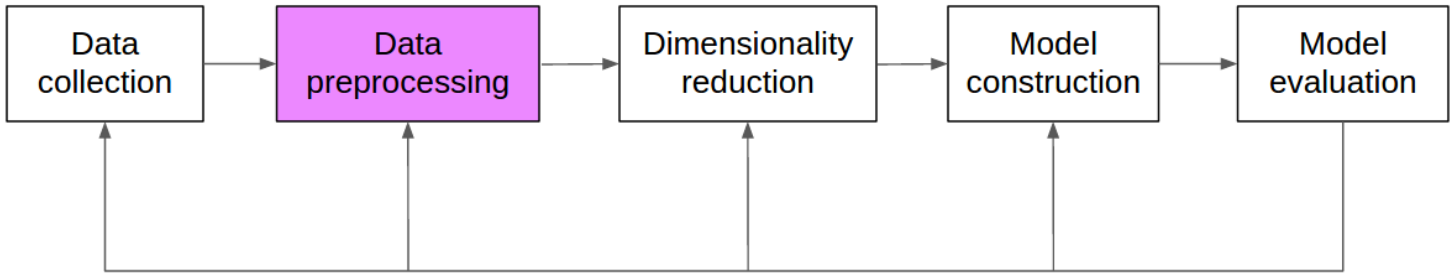
CÁTEDRA UAM/IBM
GRANDES SISTEMAS, SOLUCIONES COGNITIVAS Y ANALÍTICA BIG DATA



Data Collection

- Probably the most time-consuming part
- How much data?
  - Sufficiently large number of instances
  - Representative

Data preprocessing

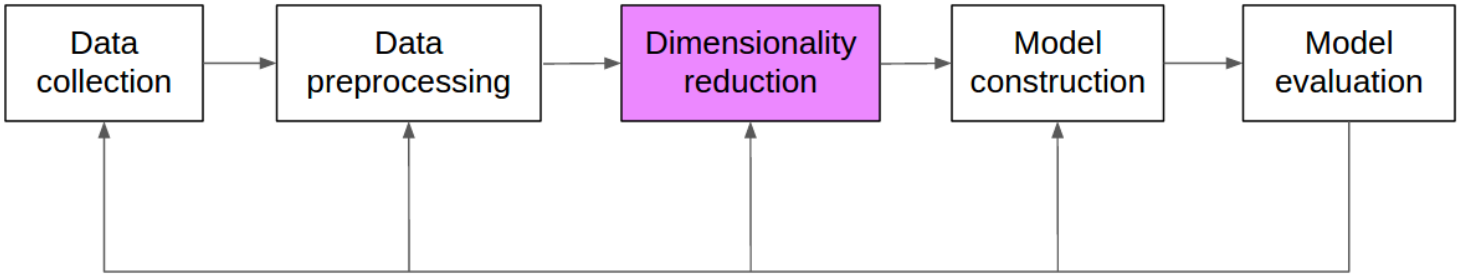


Data Preprocessing

- Correct inconsistencies in data
- Data Cleansing: missing values, outliers, noise, etc.
- Data Transformation: normalization, smoothing, segmentation, etc.

Feature selection and dimensionality reduction

CÁTEDRA UAM/IBM
GRANDES SISTEMAS, SOLUCIONES COGNITIVAS Y ANALÍTICA BIG DATA

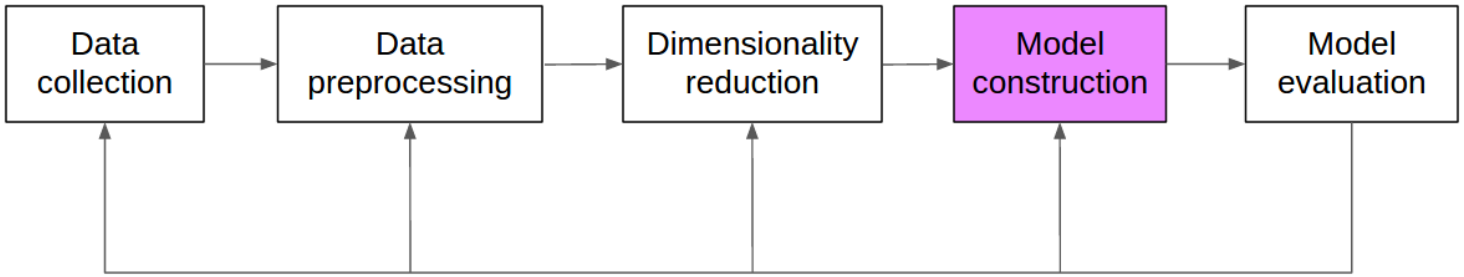


Feature selection

- Critical in any Pattern Recognition problem
- Requires a basic understanding of the problem
- Ideal features:
  - Simple to extract
  - Invariant to irrelevant transformation
  - Insensitive to noise

Model construction (pattern classification)

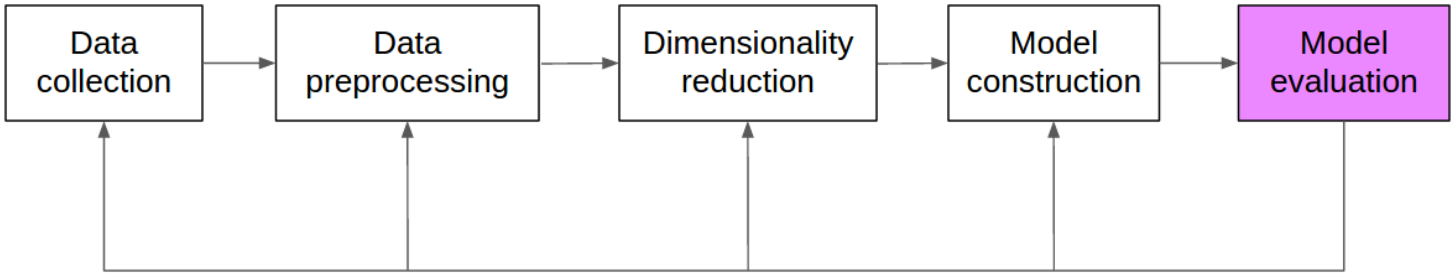
CÁTEDRA UAM/IBM
GRANDES SISTEMAS, SOLUCIONES COGNITIVAS Y ANALÍTICA BIG DATA



Model construction

- Select the model: linear discriminant, neural net, decision tree, etc.
- Use the data to train the classifier

Model evaluation

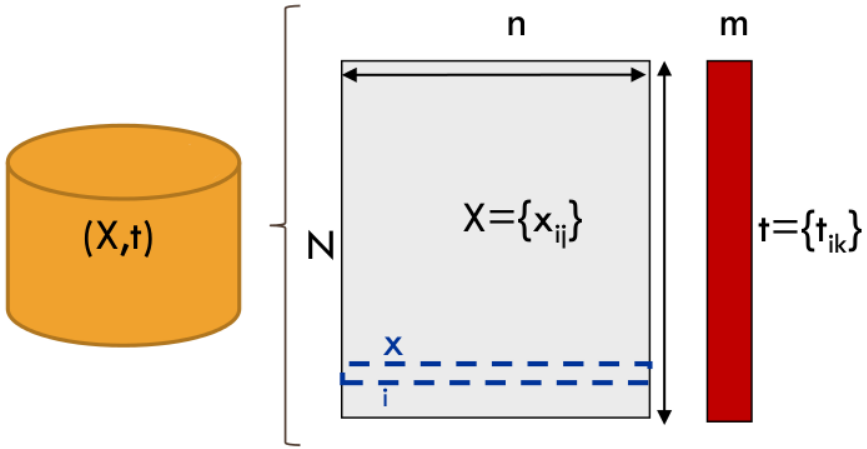


Evaluation

- How good is the trained model?
- Measure error rate to obtain model performance
- Compare the performances of different models
- Overfitting versus generalization

Pattern classification

- The problem data is the set of patterns:
  $\{(\mathbf{x_1}, t_1), (\mathbf{x_2}, t_2), ..., (\mathbf{x_n}, t_n)\}$
  - n is the number of training patterns
  - $\mathbf{x_i}$ is the attribute vector for pattern i
  - t_i is the class label for pattern i



- The goal is to predict the class of each pattern

Pattern classification

CÁTEDRA UAM/IBM
GRANDES SISTEMAS, SOLUCIONES COGNITIVAS Y ANALÍTICA BIG DATA

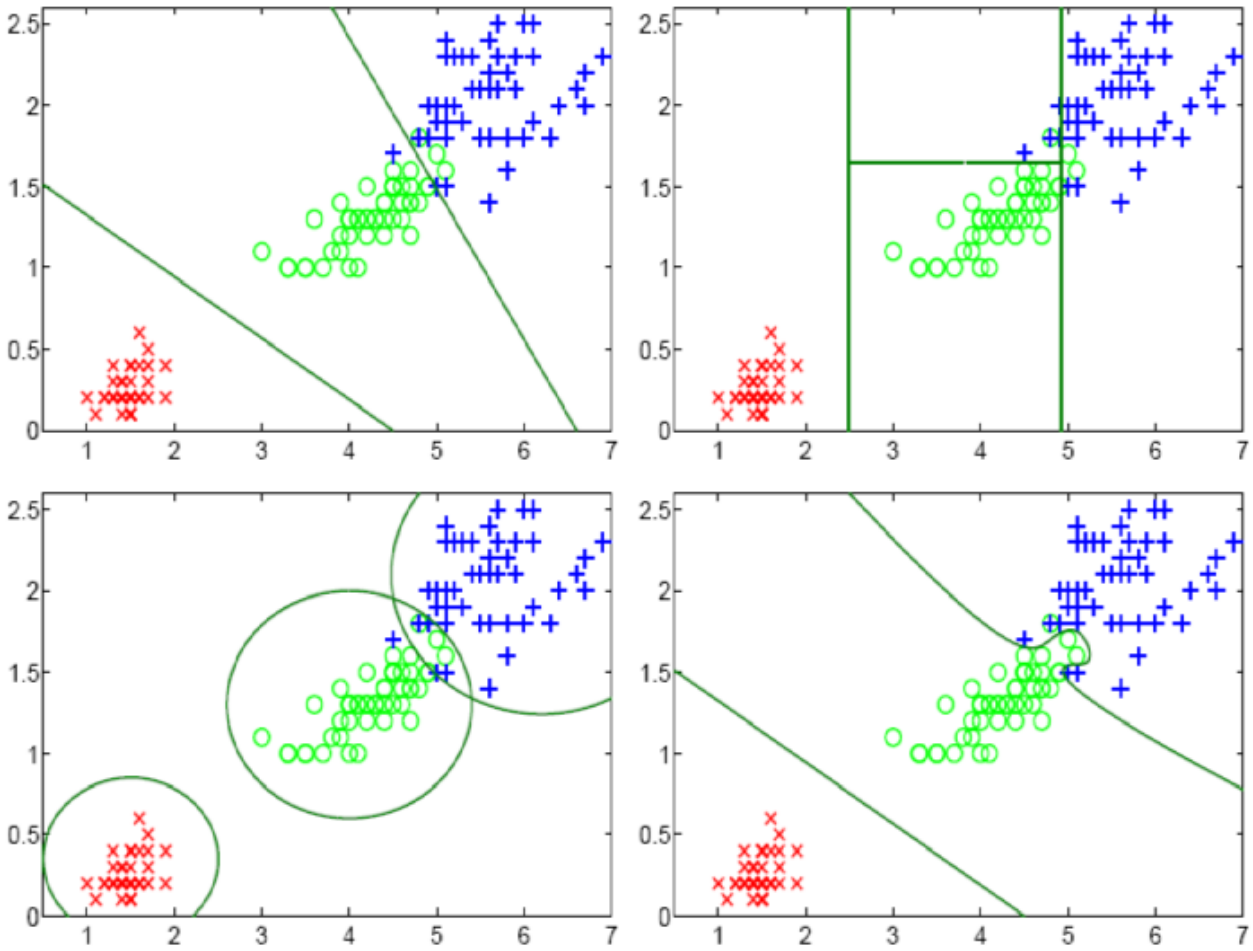
- A classifier is a function $f(\mathbf{x}, \Theta)$ that assigns each pattern $\mathbf{x_i}$ an estimation of its class $y_i = f(\mathbf{x_i}, \Theta)$
- Training the classifier means tuning the parameters Θ in order to minimize an error function that measures the discrepancy between the real classes t_i and the predictions y_i.
- Different function families define different types of classifiers: linear discriminant analysis, neural networks, decision trees, Bayesian methods, support vector machines, etc.

Pattern classification

CÁTEDRA UAM/IBM
GRANDES SISTEMAS, SOLUCIONES COGNITIVAS Y ANALÍTICA BIG DATA

Linear discriminant

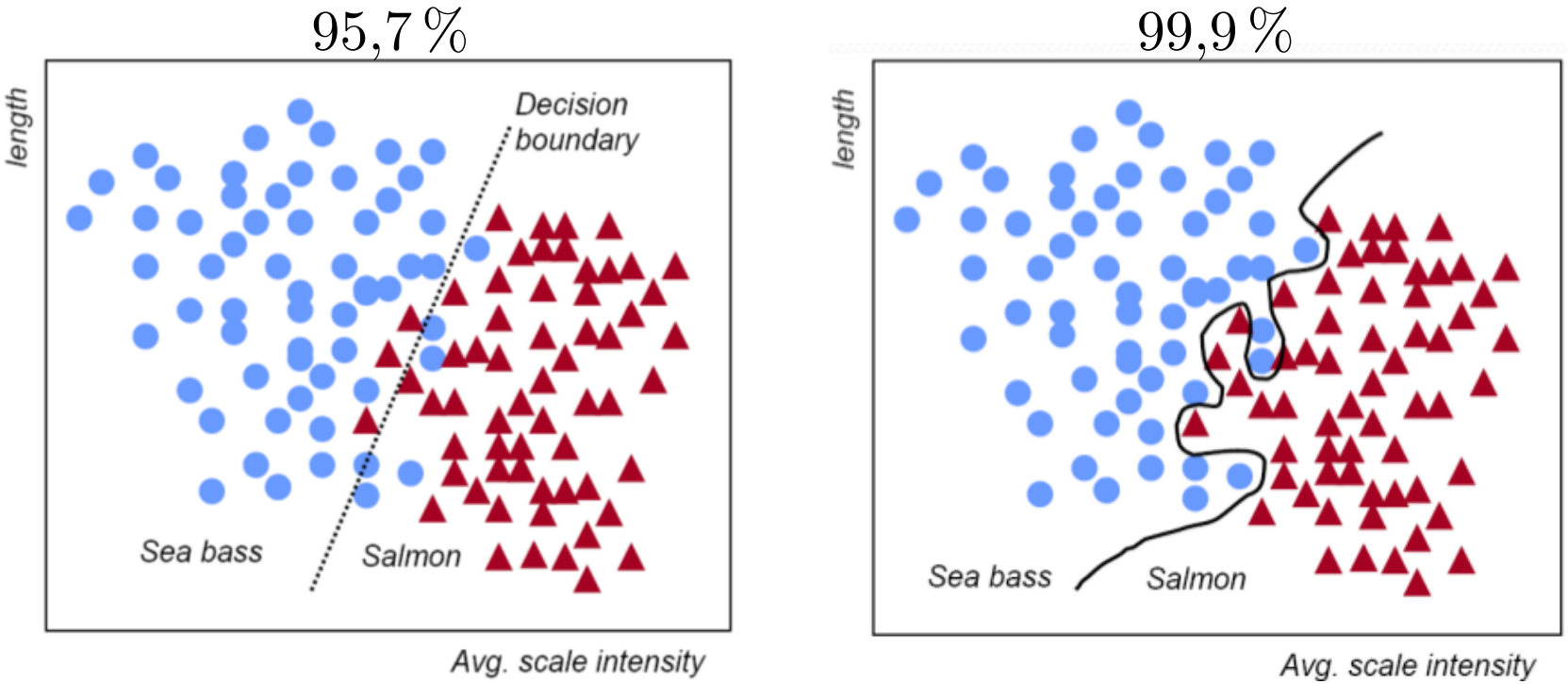
Decision tree



Gaussian mixture

Support vector machine

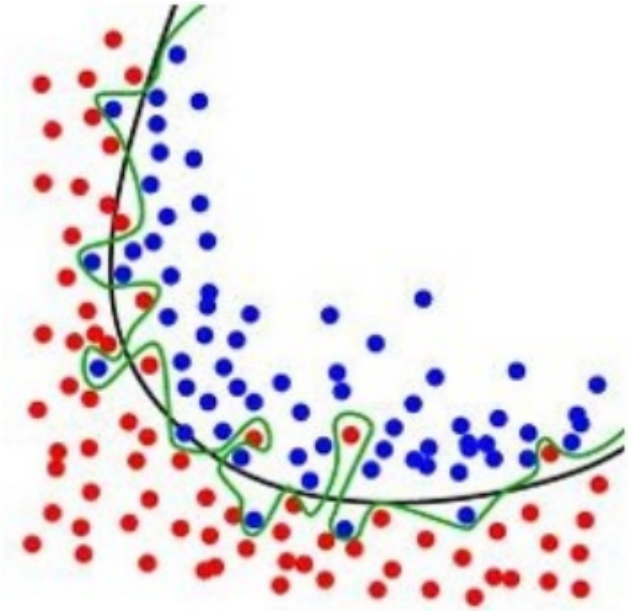
Model evaluation and selection



(From Duda, Hart and Stork, *Pattern Classification*, 2001)

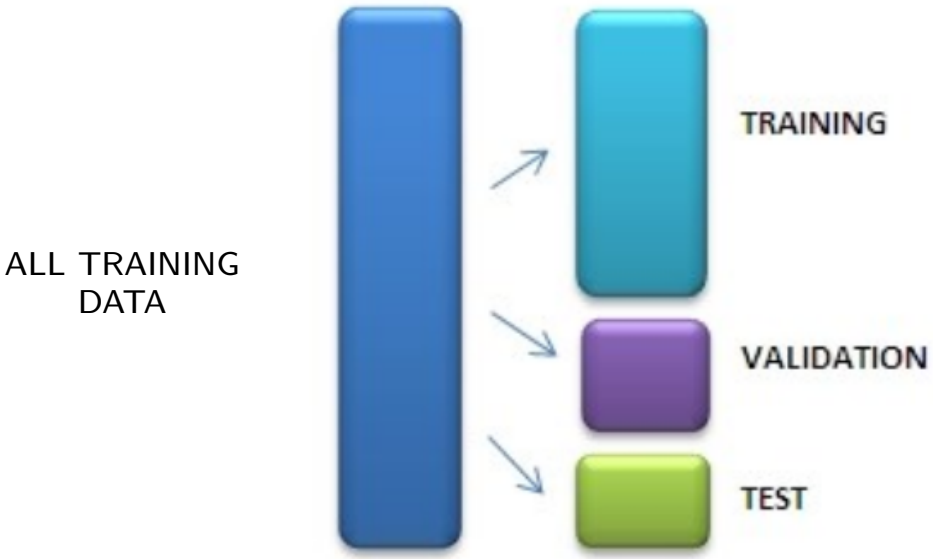
- Which model is better?

Overfitting vs generalization



- Measuring the classification accuracy on the training data is not the best way to evaluate the model
- It is better to use a different data set not used during the training phase

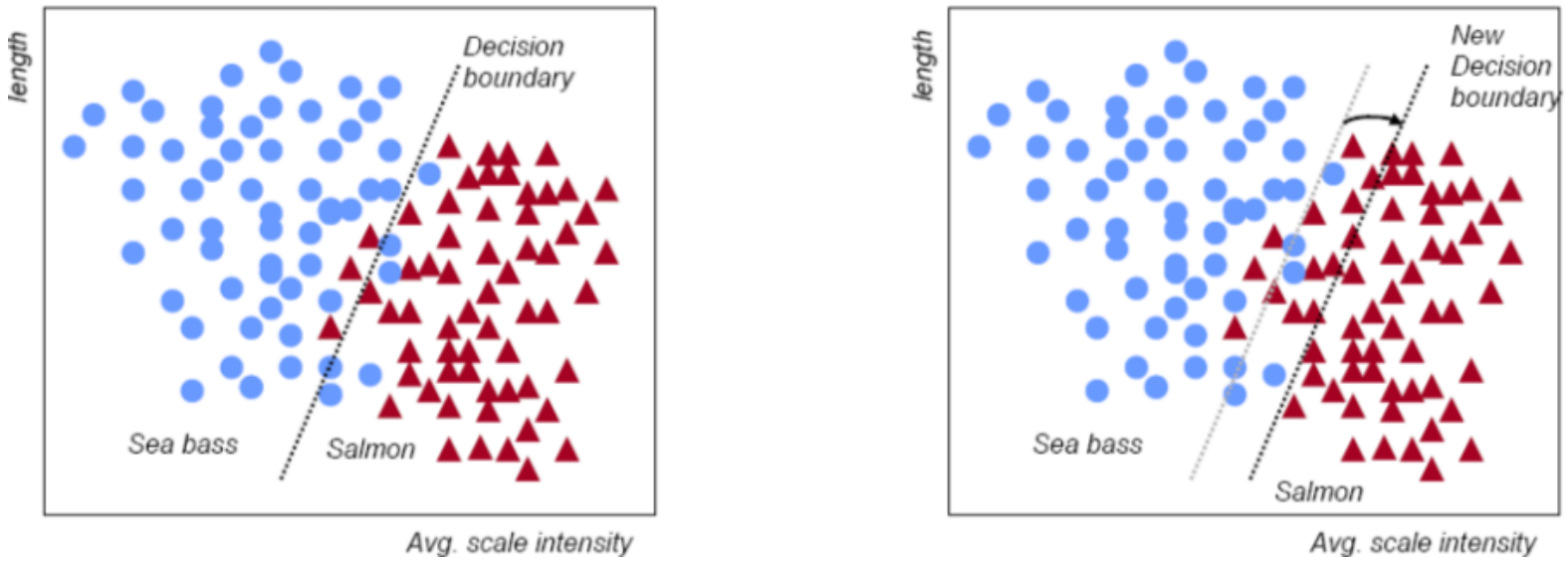
Training, validation, test



- Training set: used to train the models
- Validation set: used to validate the models and to select the final classifier
- Test set: used to test the model performance

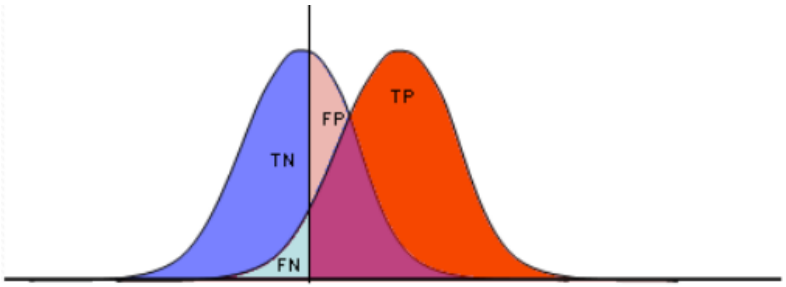
Cost versus classification rate

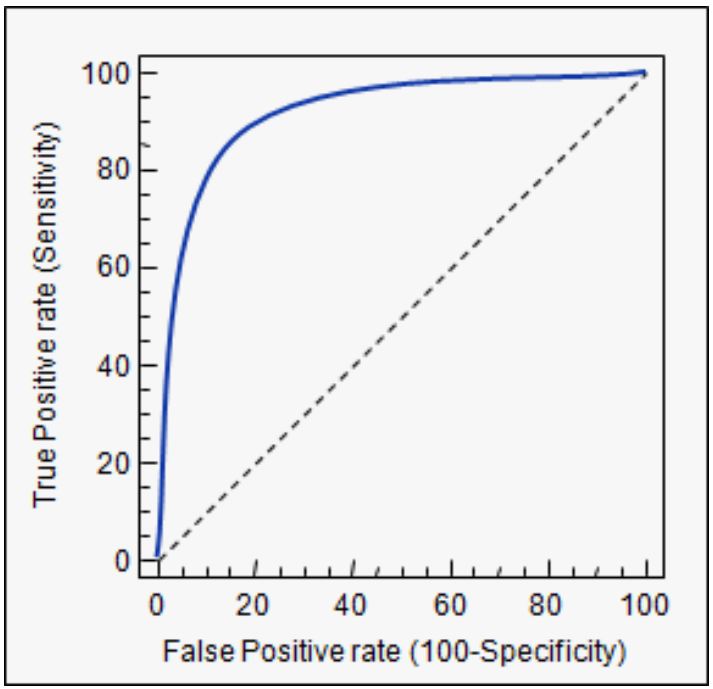
- Usually misclassified patterns from different classes imply different costs
- In those cases we might want to adjust the decision boundary in order to minimize the cost associated to misclassifications



(From Duda, Hart and Stork, *Pattern Classification*, 2001)

ROC analysis





- True positive rate:

$$TPR = \frac{TP}{TP + FN}$$

- False positive rate:

$$FPR = \frac{FP}{FP + TN}$$

- The area under the curve (AUC) is a good measure of model performance

Data preparation and audit

CÁTEDRA UAM/IBM
GRANDES SISTEMAS, SOLUCIONES COGNITIVAS Y ANALÍTICA BIG DATA

- Process that transforms the raw input data into the set of patterns used to train the models
- Initial data may have any form
- Final data used to train the models:
  - Attribute matrix (one row per pattern, one column per attribute), usually numeric data
  - Vector of targets, usually symbolic data
- Note that
  - Training data must be measured in the same conditions as evaluation data
  - Any bias in the training data should be avoided

Data preprocessing

CÁTEDRA UAM/IBM
GRANDES SISTEMAS, SOLUCIONES COGNITIVAS Y ANALÍTICA BIG DATA

- Correct inconsitencies in data that may negatively affect the training phase
- Data preprocessing has a strong impact in the model performance
- Data preprocessing includes
  - Data cleansing: filter data that is not useful, correct errors in the database
  - Data transformation: change coding into a more appropriate format that facilitates learning

Data cleansing and transformation

CÁTEDRA UAM/IBM
GRANDES SISTEMAS, SOLUCIONES COGNITIVAS Y ANALÍTICA BIG DATA

- Cleansing:
  - Fill in missing values
  - Filter noise
  - Remove inconsistencies
  - Remove correlations
  - Remove outliers
- Transformation:
  - Normalization and/or standardization
  - Smoothing
  - Data aggregation
  - Segmentation
  - Compression of temporal series

Attribute selection and dimensionality reduction

CÁTEDRA UAM/IBM
GRANDES SISTEMAS, SOLUCIONES COGNITIVAS Y ANALÍTICA BIG DATA

- Avoid the *curse of dimensionality*
- Select the most informative attributes, those which convey the most information about the target
- Avoid false predictors: attributes that are strongly correlated with the target, but that are not available in a realistic prediction scenario
- Reduce dimensionality by making projections in the attribute space
  - Principal Component Analysis
  - Linear Discriminant Analysis

Summary

CÁTEDRA UAM/IBM
GRANDES SISTEMAS, SOLUCIONES COGNITIVAS Y ANALÍTICA BIG DATA

- Machine Learning is in the core of Big Data
- Models/classifiers are trained using labeled data, and used to make predictions on unlabeled data
- Prevent overfitting: models should be evaluated using data that was not used for training
- Building the model is just a single step in the whole process (data collection, preprocessing, dimensionality reduction, etc.)

Bibliography

CÁTEDRA UAM/IBM
GRANDES SISTEMAS, SOLUCIONES COGNITIVAS Y ANALÍTICA BIG DATA

- *Pattern Classification.* R.O. Duda, P.E. Hart, D.G. Stork. Wiley, 2001.
- *Pattern Recognition and Machine Learning.* C. Bishop. Springer, 2006.
- *Machine Learning in Python: Essential Techniques for Predictive Analysis.* M. Bowles. Wiley, 2015.
- *Introduction to Machine Learning with Python. A Guide for Data Scientists.* A.C. Mueller, S. Guido. O'Reilly, 2016.