Politecnico di Bari

Politecnico di Bari

Dipartimento di Ingegneria Elettrica e dell'Informazione

Corso di Laurea Triennale in Ingegneria dei Sistemi Medicali

DIPARTIMENTO DI INGEGNERIA ELETTRICA E DELL'INFORMAZIONE

Bioinformatics and Big Data Analytics

Machine Learning Basics

Eng. Nicola **Altini**, *Ph.D. Student*
Eng. Giacomo Donato **Cascarano**, *Ph.D. Student*
Prof. *Eng*. Vitoantonio **Bevilacqua**, *Ph.D.*

INDUSTRIAL INFORMATICS LAB
LABORATORIO DI INFORMATICA INDUSTRIALE

Anno Accademico 2019/2020

apulian bioengineering company

Outline

- Machine Learning
- Supervised Learning
  - Linear Regression
  - Polynomial Regression
  - Neural Networks
  - Classification Trees
- Unsupervised Learning
  - Cluster Analysis
  - Principal Component Analysis
- Data Normalization
  - Min-max normalization
  - Z-score normalization

Machine Learning

- Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead.
- Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop a conventional algorithm for effectively performing the task.
- *"A computer program is said to learn from experience (E) with some class of tasks (T) and a performance measure (P) if its performance at tasks in T as measured by P improves with E"*. [Tom Mitchell]

Supervised Learning

- Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples.
- In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way.

Supervised Learning

Given a training set of m example input-output pairs:

$$(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})$$

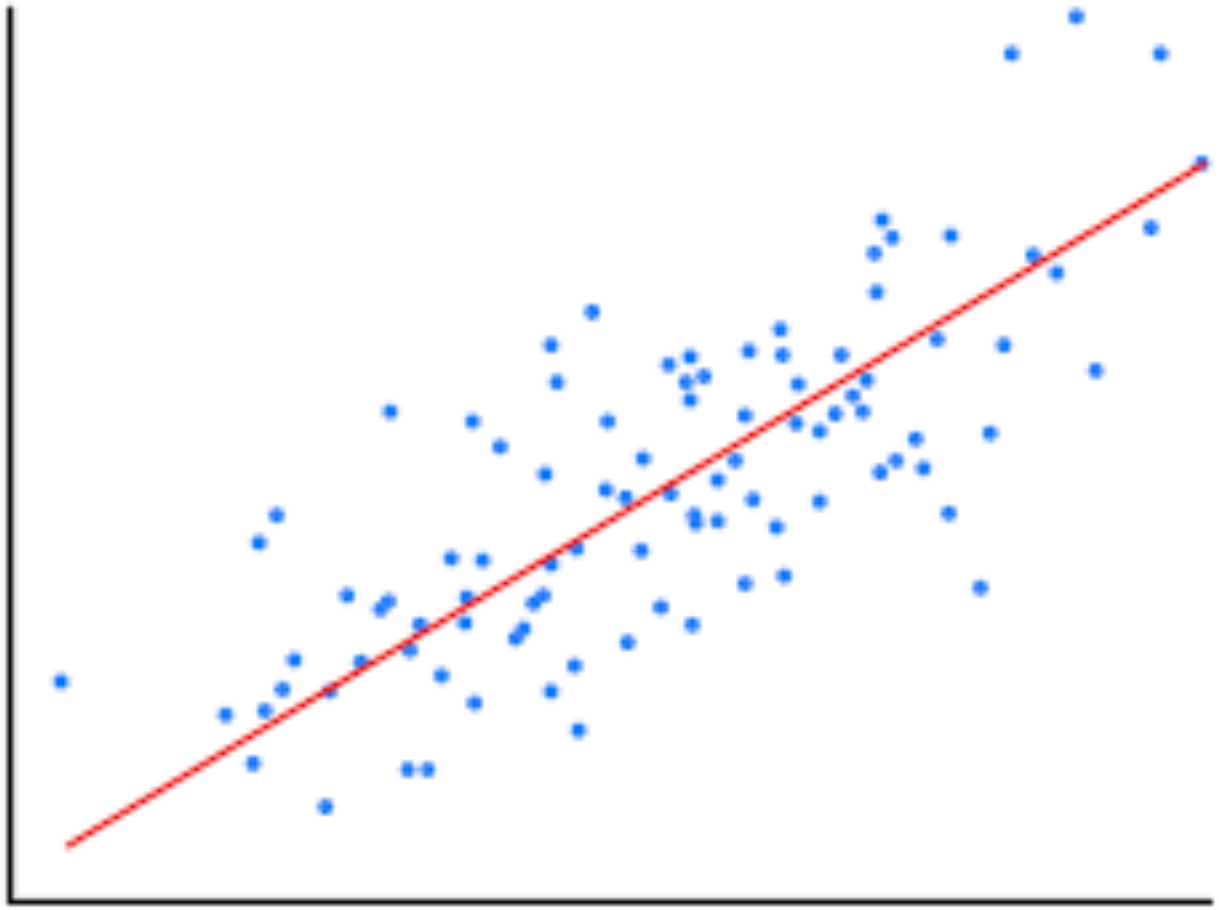
Where each y_i was generated by an unknown function $y = f(x)$, discover a function h that approximates the true function f.

- To measure the accuracy of a hypothesis we use a test set of examples different from the training set. We say that a hypothesis generalizes well if it correctly predicts the value of y for unseen samples.
- In real applications of ML the f is usually stochastic, i.e. it is not strictly a function of x, and we have to learn a conditional probability distribution $P(y \mid x)$.

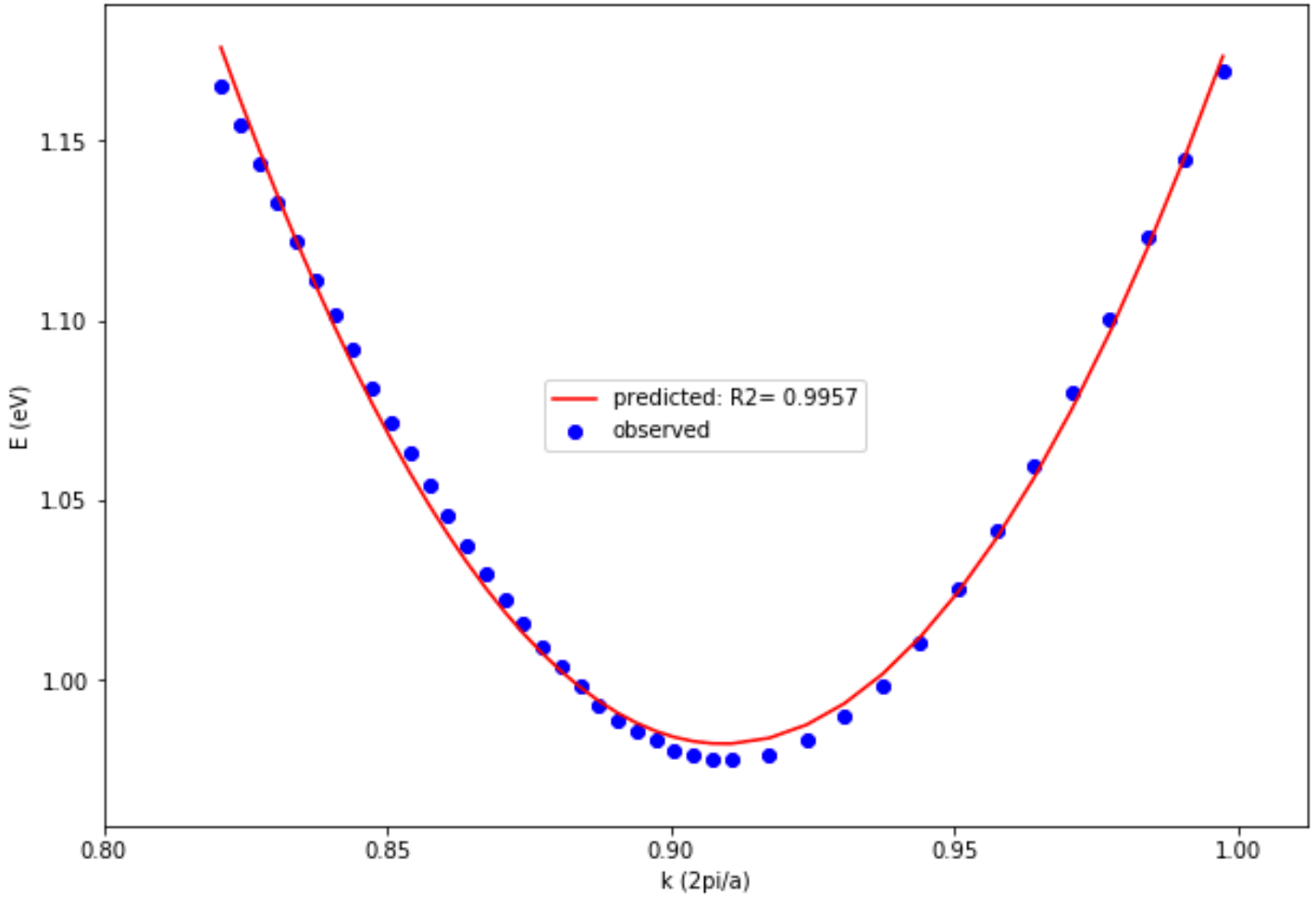
Supervised Learning

- **Regression**. When the output y is a number from a continuous range of vales (e.g., finding the bounding box coordinates of an object of interest), the learning problem is called regression.
- Examples:
  - Linear Regression
  - Polynomial Regression
- **Classification**. When the output y is one of a finite set of values (e.g. healthy glomerulus vs sclerotic glomerulus) the learning problem is called classification (binary classification if there are two classes, or multi-class classification if there are more).
- Examples:
  - Neural networks
  - Classification trees

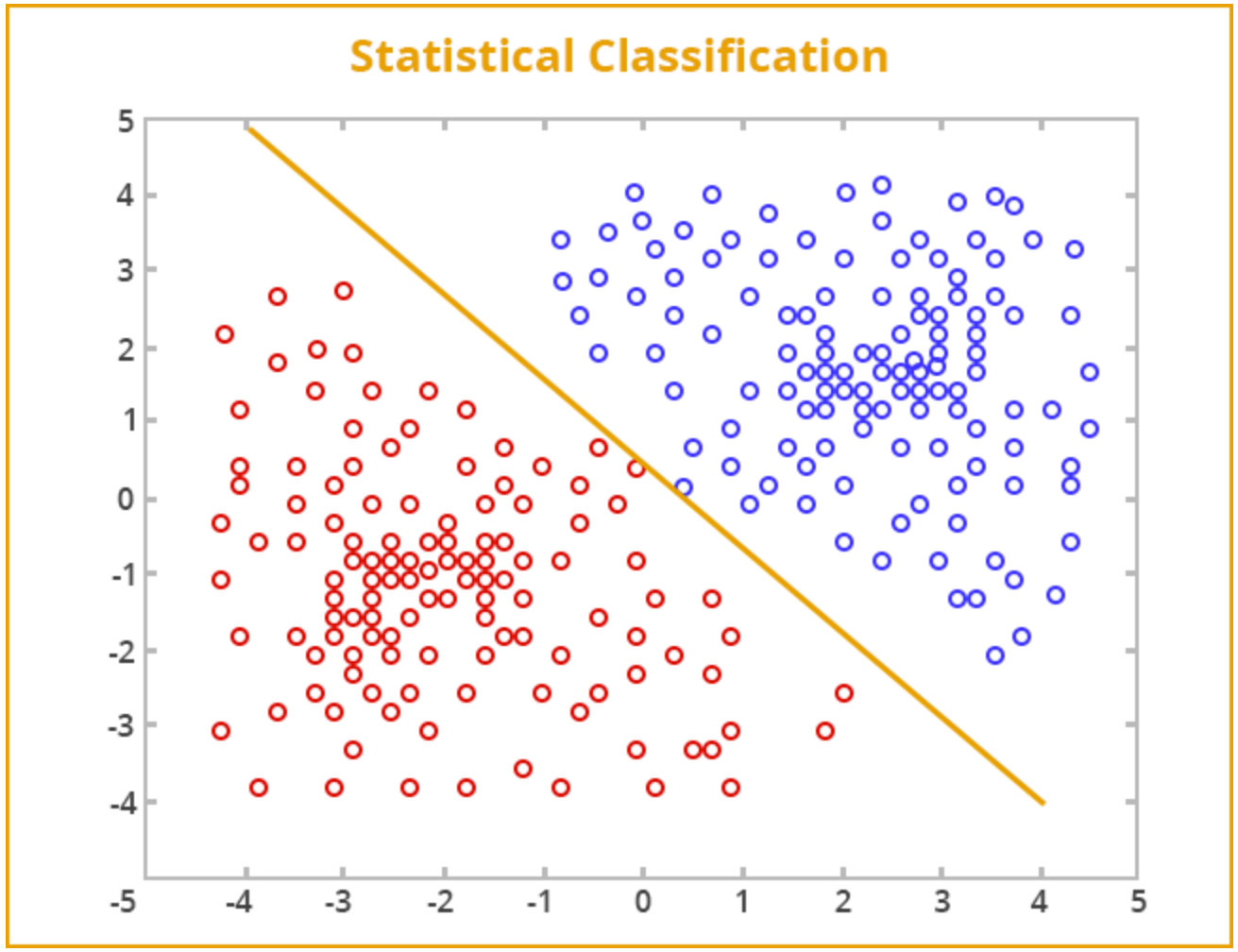
Linear Regression



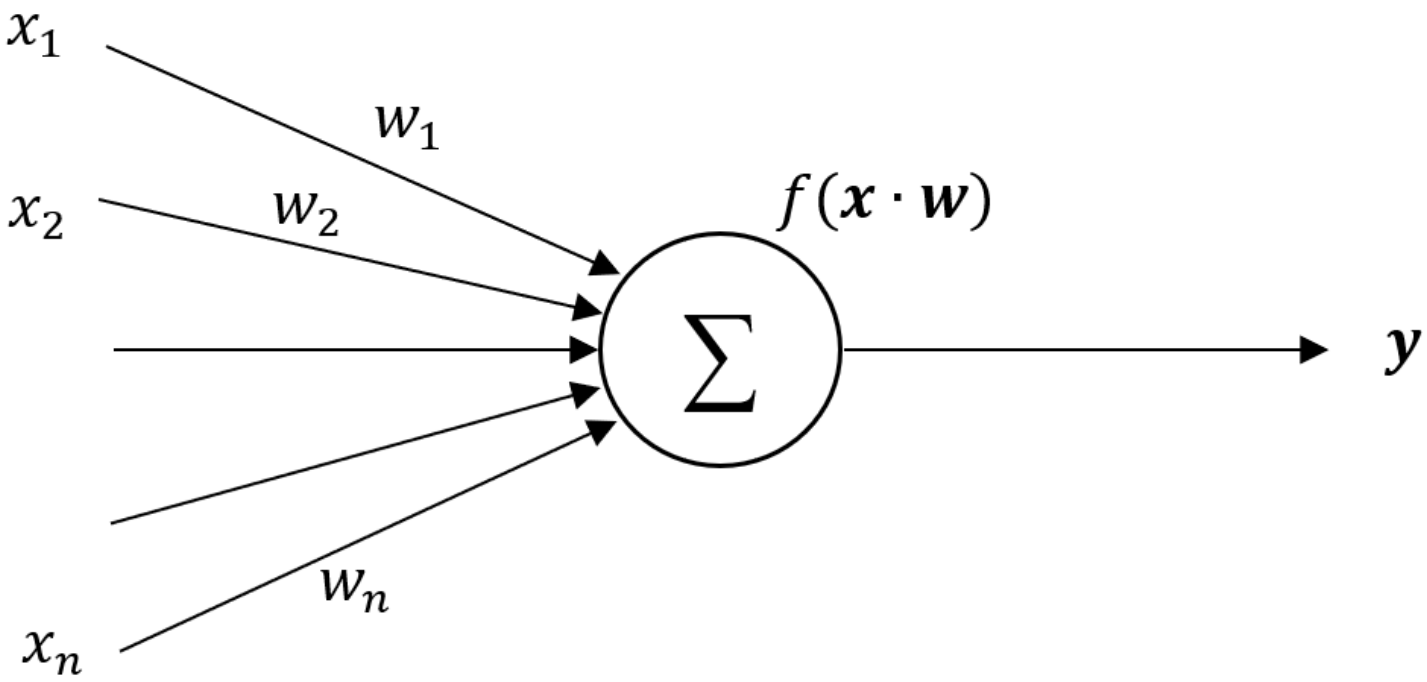
Polynomial Regression



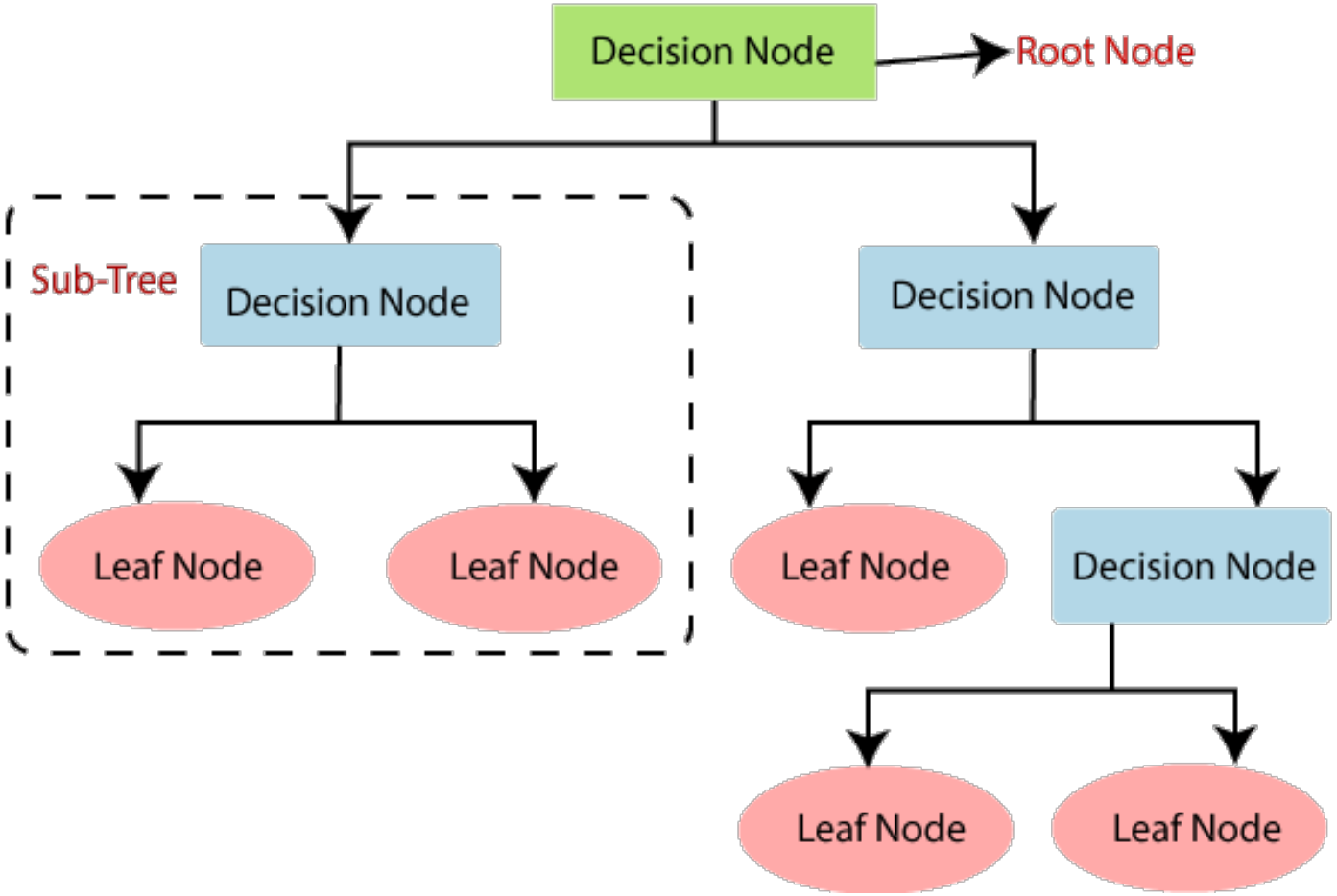
Classification



Neural Network



Decision Tree



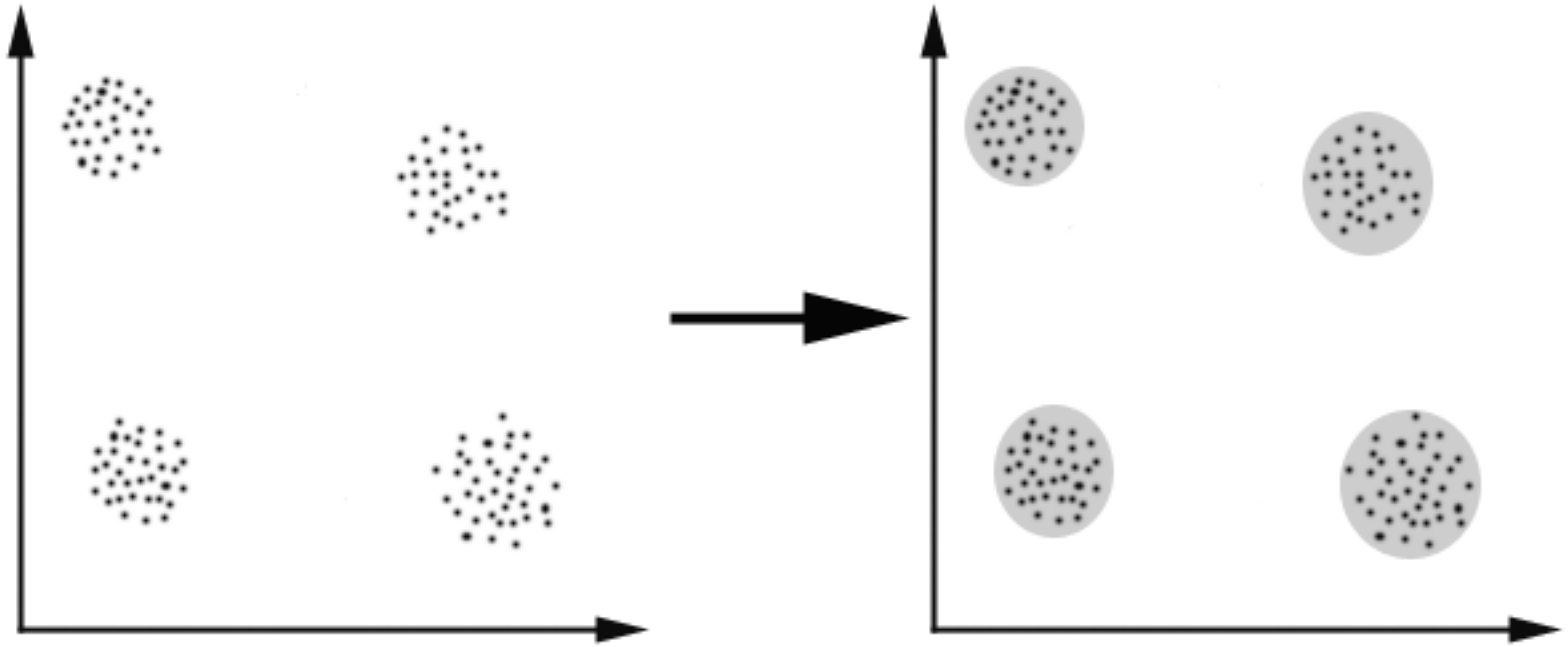
Unsupervised Learning

- A central application of unsupervised learning is in the field of density estimation in statistics, though unsupervised learning encompasses many other domains involving summarizing and explaining data features.
- It could be contrasted with supervised learning by saying that whereas supervised learning intends to infer a conditional probability distribution $p_X(x \mid y)$ conditioned on the label y of input data; unsupervised learning intends to infer an a priori probability distribution $p_X(x)$.
- Examples:
  - **Cluster Analysis**
  - **Principal Component Analysis**

Unsupervised Learning

- Two of the main methods used in unsupervised learning are principal component and cluster analysis. Cluster analysis is used in unsupervised learning to group, or segment, datasets with shared attributes in order to extrapolate algorithmic relationships.
- Cluster analysis is a branch of machine learning that groups the data that has not been labelled, classified or categorized. Instead of responding to feedback, cluster analysis identifies commonalities in the data and reacts based on the presence or absence of such commonalities in each new piece of data. This approach helps detect anomalous data points that do not fit into either group.

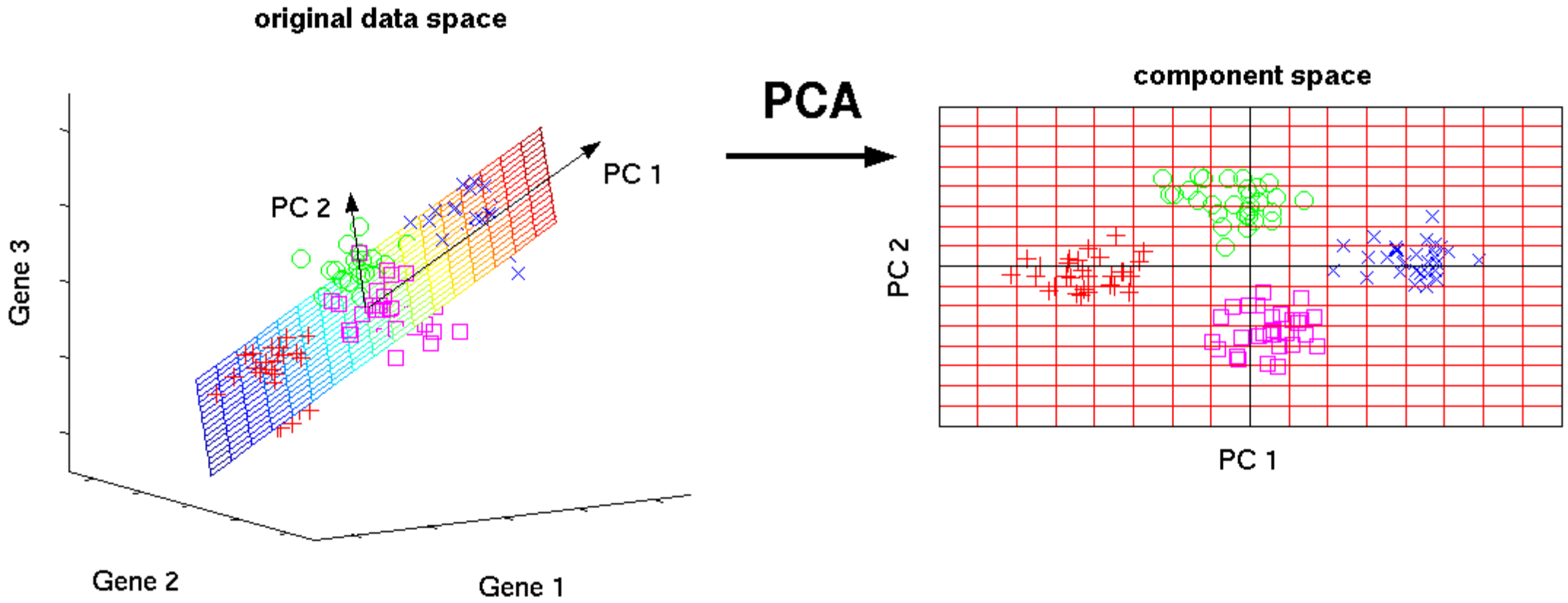
Cluster Analysis



Principal Component Analysis

- Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called principal components.
- This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components.
- The resulting vectors (each being a linear combination of the variables and containing n observations) are an uncorrelated orthogonal basis set. PCA is sensitive to the relative scaling of the original variables.

Principal Component Analysis



Data Normalization

- Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.
- Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, many classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.
- Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it.

Data Normalization

- Rescaling or min-max normalization
  - Rescaling in range [0,1]:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

  - Rescaling in range [a,b]:

$$x' = a + \frac{(x - \min(x))(b - a)}{\max(x) - \min(x)}$$

- Standardization or z-score normalization
  - We get a zero mean distribution through subtraction of the mean and dividing by the standard deviation:

$$x' = \frac{x - \mu(x)}{\sigma(x)}$$