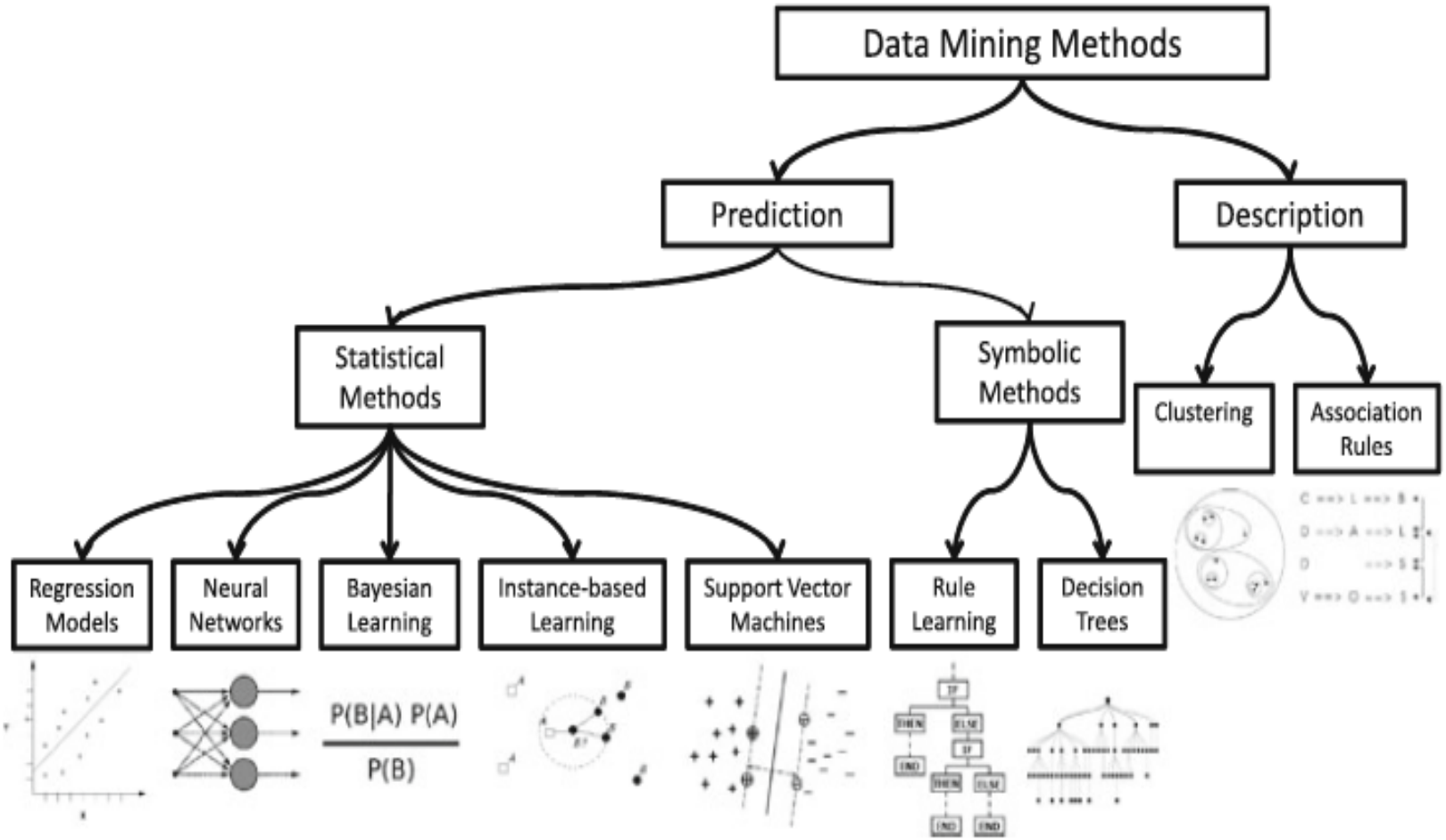
Fundamentals of ML

Fundamentals of ML

1. Data Mining Methods
2. Supervised Learning
3. Unsupervised Learning
4. Data Preprocessing
5. Types of ML Algorithms

Data Mining Methods

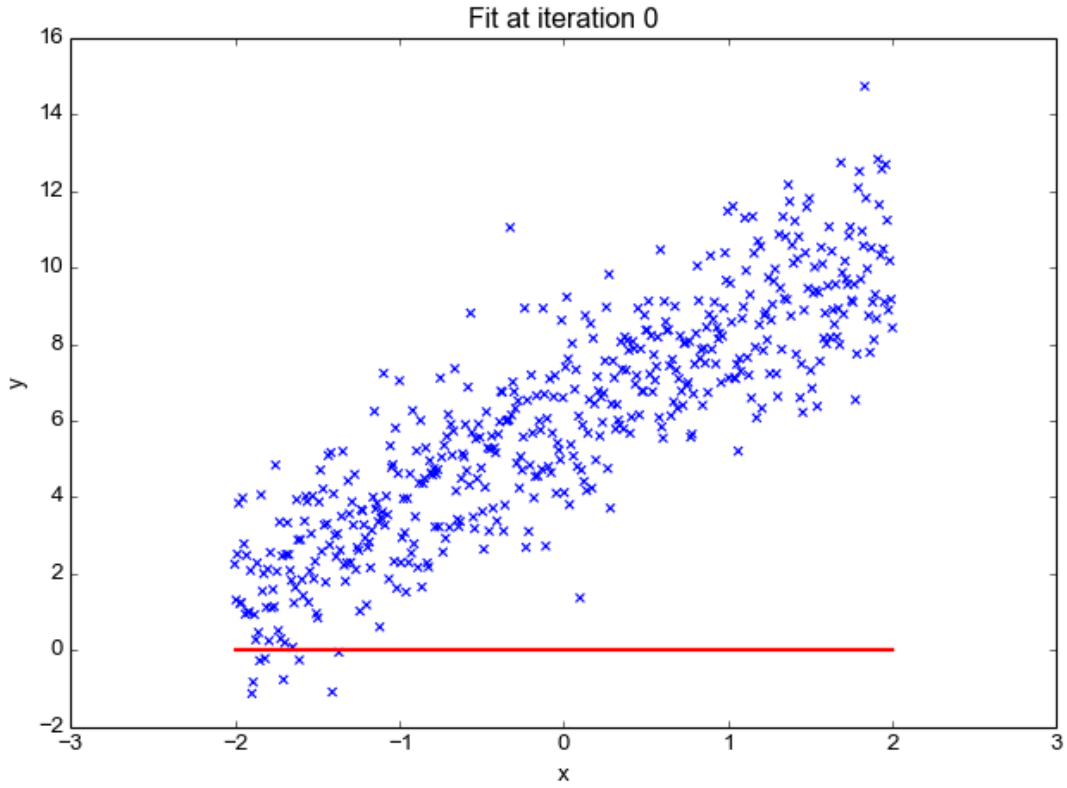


Data Mining Methods

- Statistical Methods:
  - Regression Models:
    - They are used in estimation tasks
    - Linear, quadratic and logistic regression are the most well known.
    - They may have problems with missing values, outliers and redundant/harmful features.

Data Mining Methods

- Statistical Methods:
  - Regression Models:

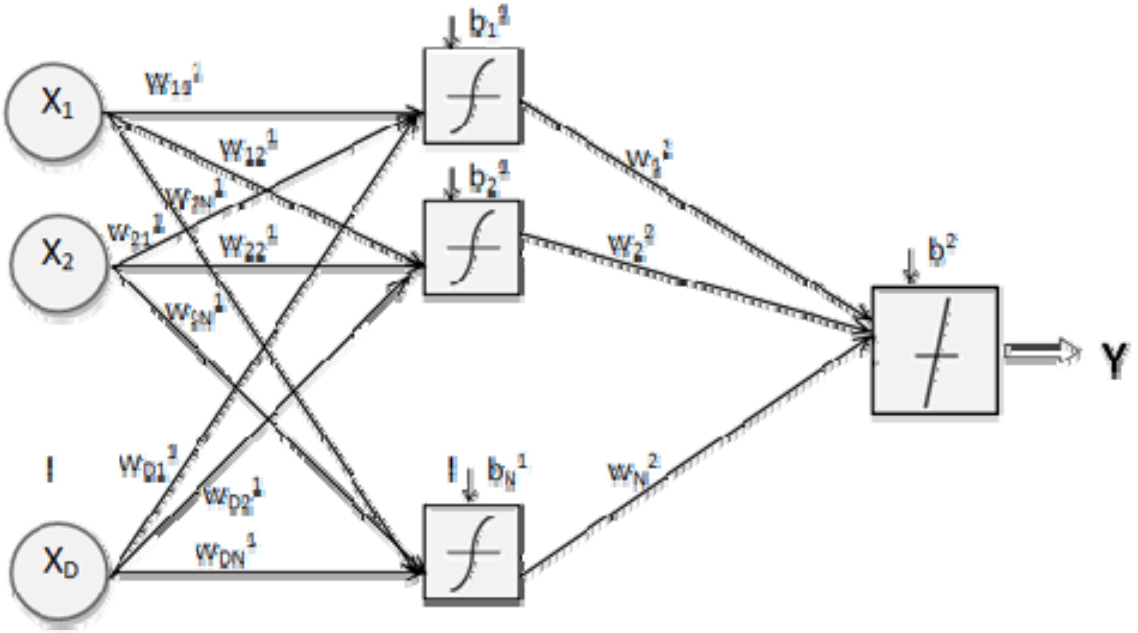


Data Mining Methods

- Statistical Methods:
  - Artificial Neural Networks (ANNs):
    - They are powerful mathematical models suitable for almost all ML tasks, especially predictive one.
    - Multi-layer perceptron (MLP), Radial Basis Function Networks (RBFNs) and Learning Vector Quantization (LVQ) are the most well known.
    - They require numeric attributes and may have problems with missing values.
    - They are robust to outliers and noise.

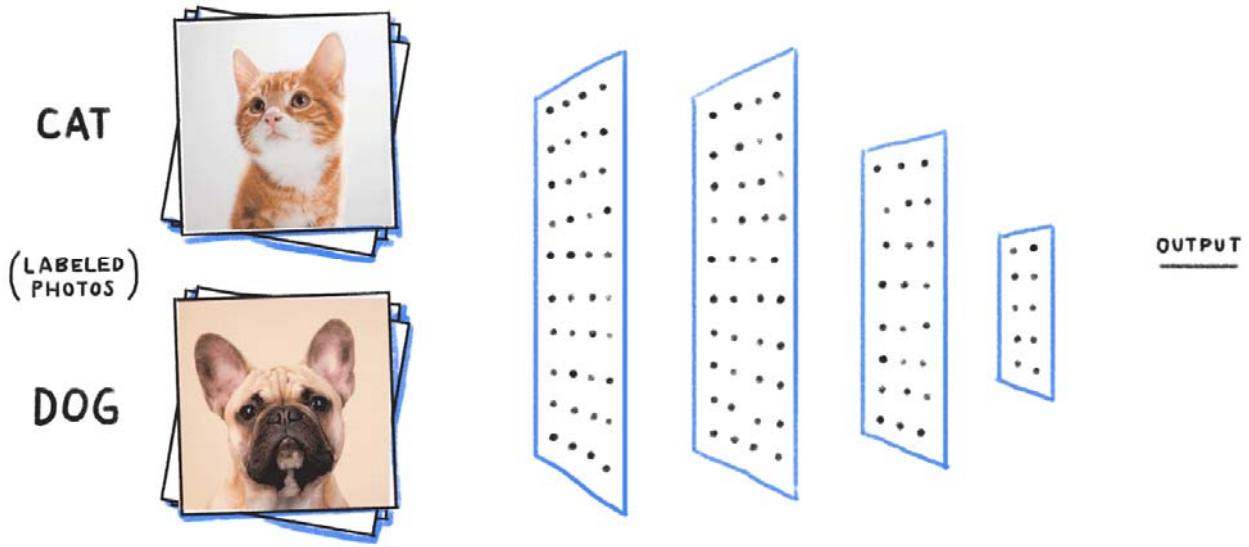
Data Mining Methods

- Statistical Methods:
  - Artificial Neural Networks (ANNs):



Data Mining Methods

- Statistical Methods:
  - Artificial Neural Networks (ANNs):



Data Mining Methods

- Statistical Methods:
  - Bayesian Learning:
    - It uses the probability theory as a framework for making rational decisions under uncertainty.
    - Naïve Bayes is the most well known technique.
    - They are very sensitive to the redundancy and usefulness of some of the attributes and examples from the data, together with noisy and outliers examples.
    - Example: P(X="cat"|A1,A2,…) = ?

Data Mining Methods

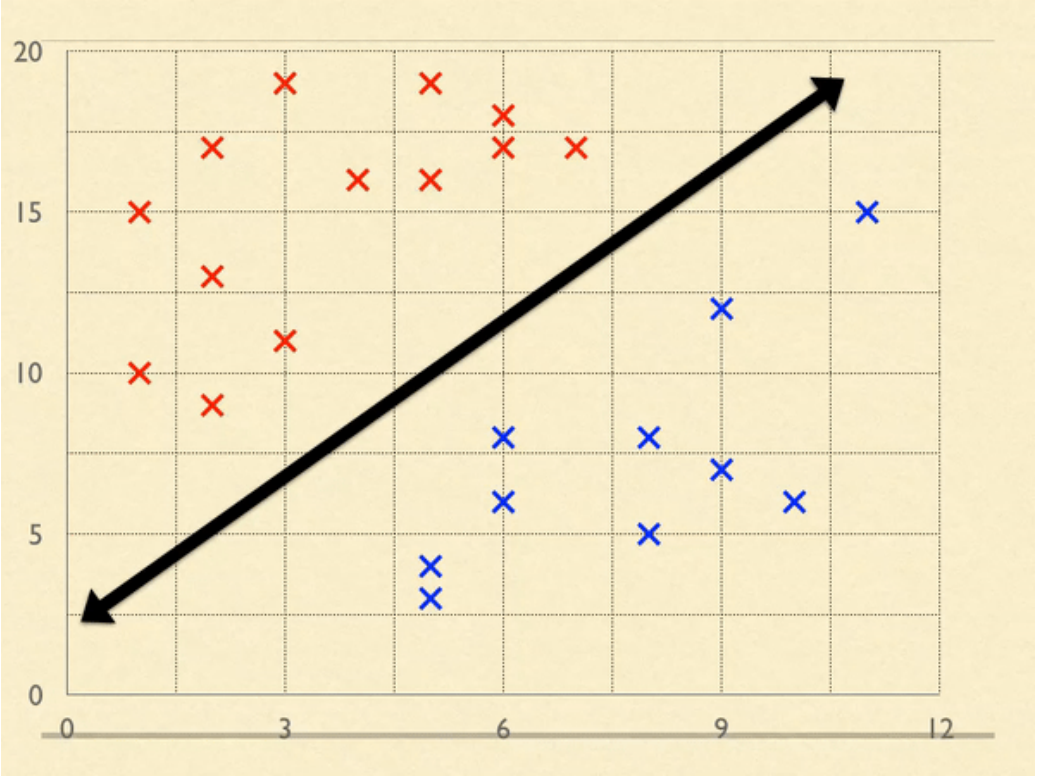
- Statistical Methods:
  - Instance-based Learning:
    - distance function is used to determine which members of the dataset are closest to a new example with a desirable prediction.
    - The K-Nearest Neighbor (KNN) is the most representative method.

Data Mining Methods

- Statistical Methods:
  - Support Vector Machines (SVMs):
    - They are machine learning algorithms based on learning theory and similar to ANNs in the sense that they are used for estimation and perform very well when data is linearly separable.
    - They require numeric, non-missing data and are commonly robust against noise and outliers.

Data Mining Methods

- Statistical Methods:
  - Support Vector Machines (SVMs):



Data Mining Methods

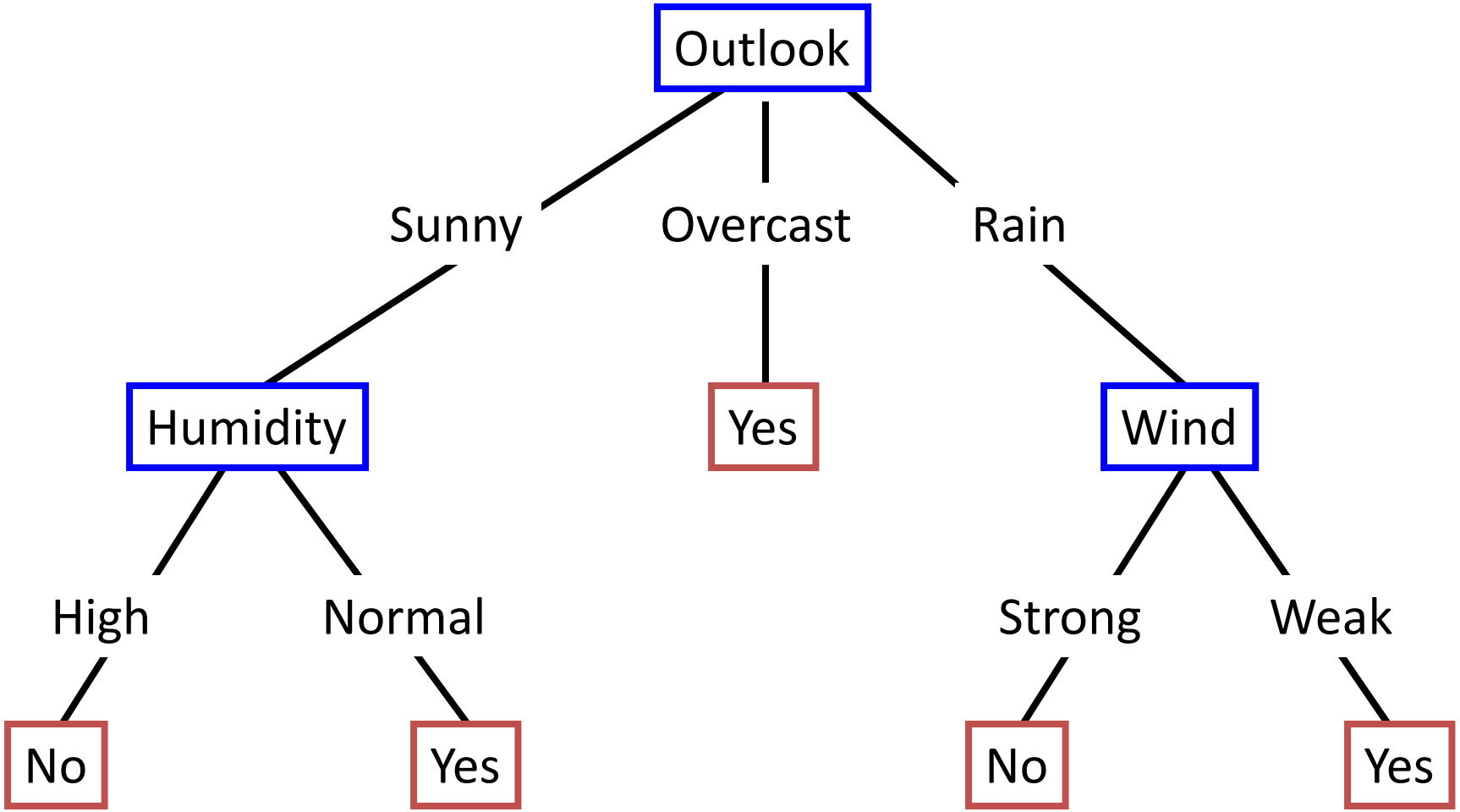
- Symbolic Methods:
  - Decision Trees:
    - They construct predictive models formed by iterations of a divide and conquer scheme of hierarchical decisions.
    - They are closely related to rule learning methods and suffer from the same disadvantage as them.

Example

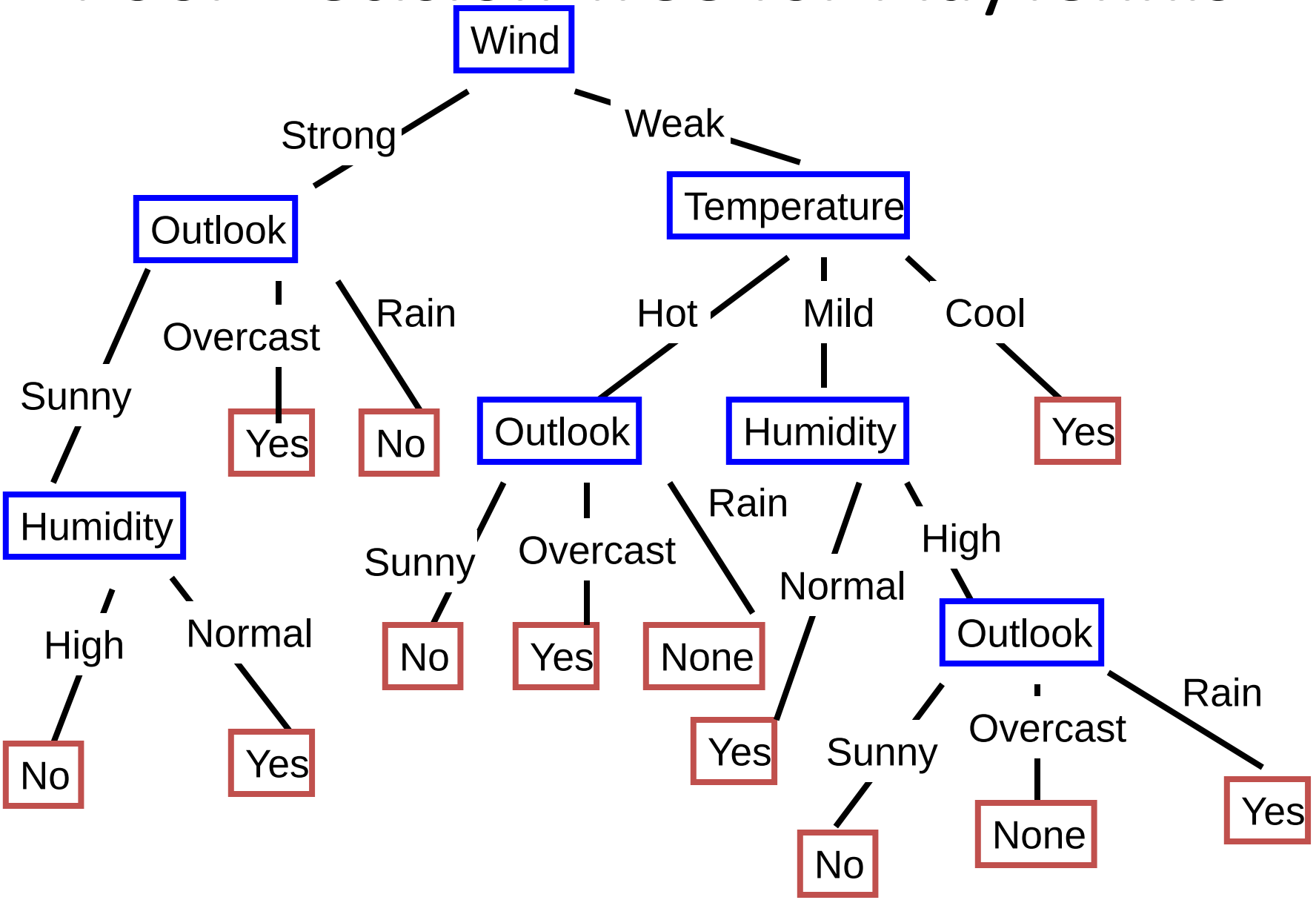
Day	Outlook	Temp.	Humidity	Wind	Play (Tennis)
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

A New Day: Outlook= sunny, Temperature=cool, humidity=high and windy= strong, play tennis?

Decision Tree for PlayTennis



Poor Decision Tree for PlayTennis

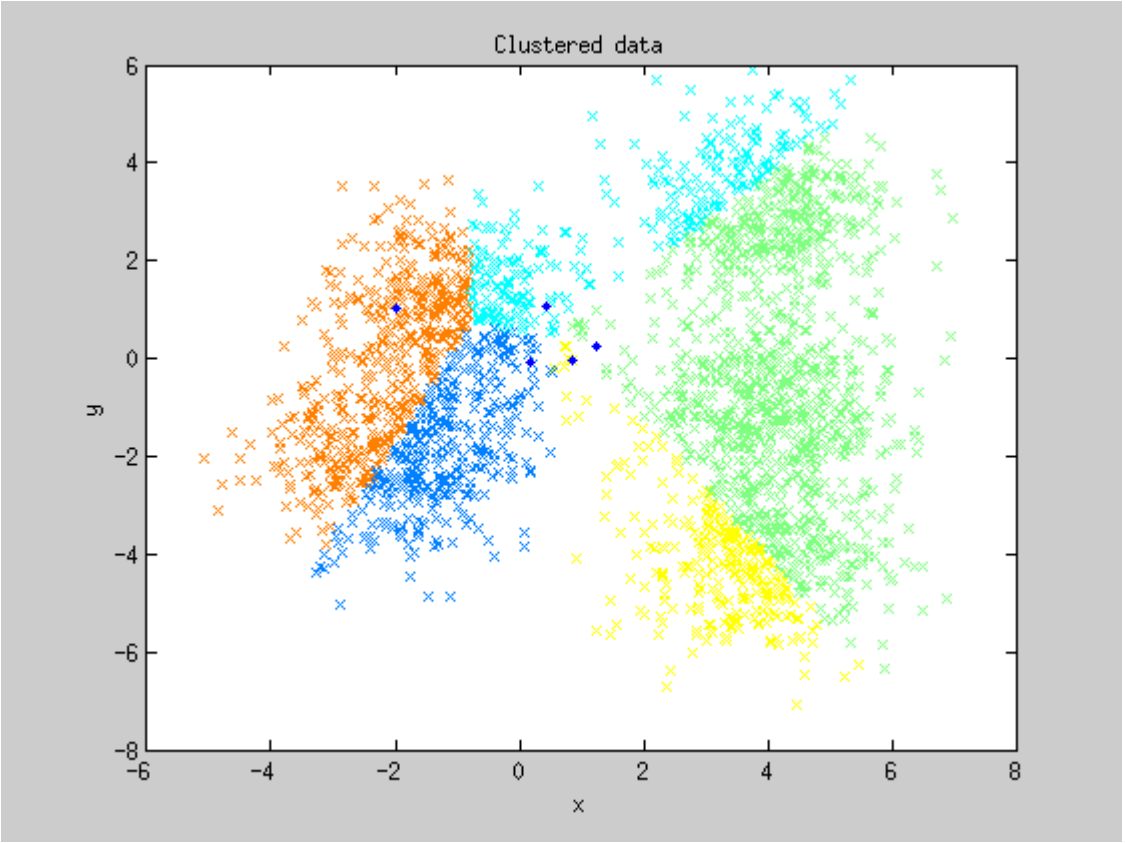


Data Mining Methods

- Data descriptive tasks:
  - Clustering:
    - It appears when there is no class information to be predicted but the examples must be divided into natural groups or clusters.
    - k-Means : Well known clustering algorithm
    - They prefer numeric data together with no-missing data and the absence of noise and outliers.

Data Mining Methods

- Data descriptive tasks:
  - Clustering:



Supervised Learning

- Prediction methods are commonly referred to as supervised learning. Supervised methods are thought to attempt the discovery of the relationships between input attributes and a target attribute.
- A training set is given and the objective is to form a description that can be used to predict unseen examples.

Supervised Learning

- Problems:
  - Classification
    - The domain of the target attribute is finite and categorical.
    - A classifier must assign a class to a unseen example.
  - Regression
    - The target attribute is formed by infinite values.
    - To fit a model to learn the output target attribute as a function of input attributes.
  - Time Series Analysis
    - Making predictions in time.

Unsupervised Learning

- There is no supervisory data and only input data is available.

- The aim is now to find regularities, irregularities, relationships, similarities and associations in the input.

Unsupervised Learning

- Problems:
  - Clustering
  - Pattern Mining
  - Outlier Detection

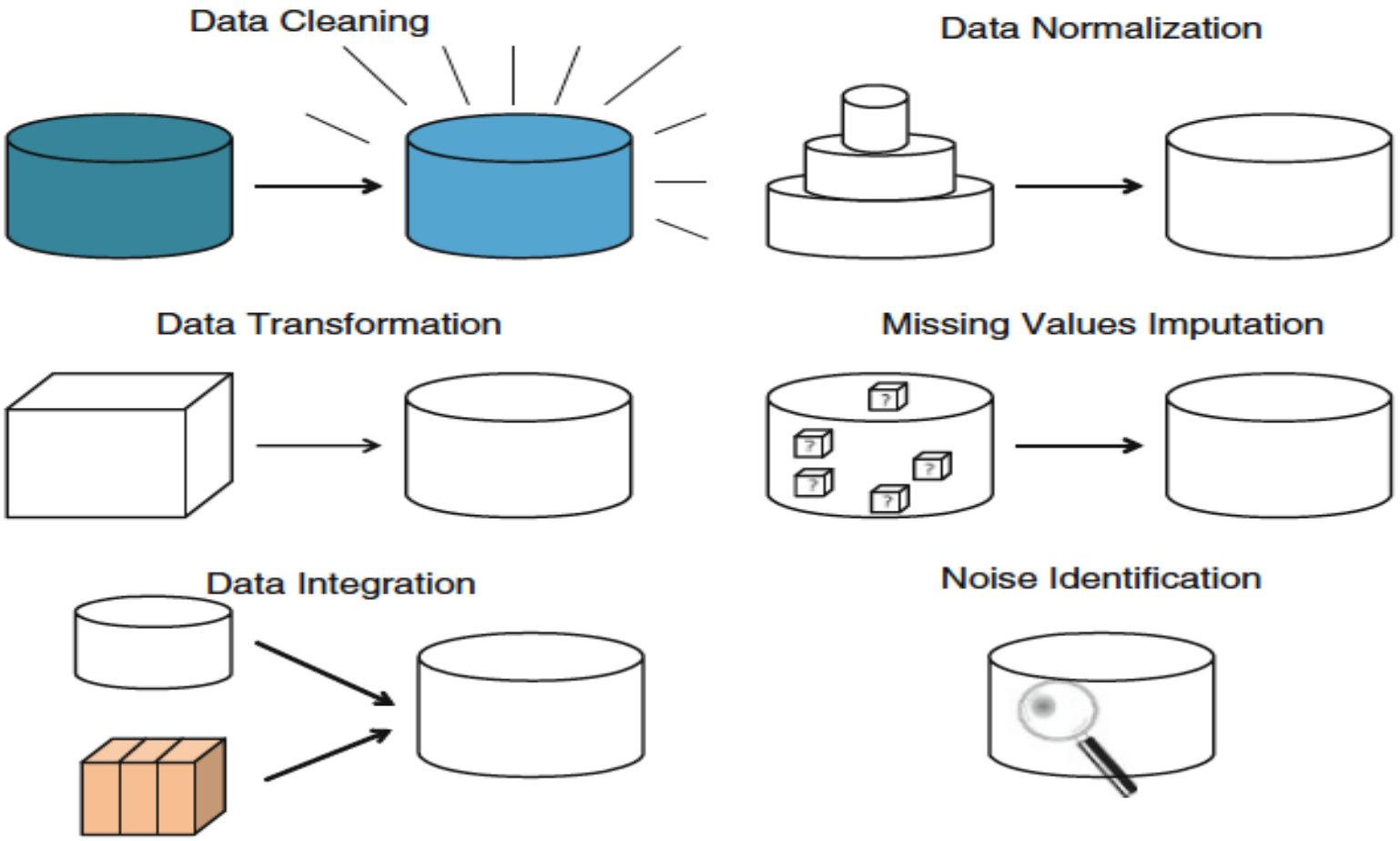
Data Preprocessing

Introduction to Data Preprocessing

- Unfortunately, real-world databases are highly influenced by negative factors:
  - noises
  - missing values
  - inconsistent and redundant data
  - huge sizes: both dimensions and features
- Low-quality data will lead to low-quality ML performance.

Introduction to Data Preprocessing

- Forms of Data Preparation



Introduction to Data Preprocessing

- Data Cleaning
  - Correct bad data
  - filter some incorrect data out of the data set
  - reduce the unnecessary detail of data.
- Data Transformation
  - The data is consolidated so that the mining process result could be applied or may be more efficient.
- Data Integration
  - Merging of data from multiple data stores.

Introduction to Data Preprocessing

- Data Normalization and Standardization
  - **Normalization:** To express data in the same measurements units, scale or range.
  - **Standardization:** To transforms data to have a mean of zero and a standard deviation of 1
- Missing Data Imputation
  - To fill the variables that contain missing values with some intuitive data: mean values, median, mode, default values, etc.
  - Detect and handle missing values in a column

Introduction to Data Preprocessing

- Normalization
  - To be done if the given data does not conform to any distribution

$$\text{scaled value} = \frac{value - min}{max - min}$$

- Standardization or z-score
  - To be done if the given data conforms or close to Normal distribution

$$\text{standardized\_value}_i = \frac{\sum_{i=1}^{n}(value_i - mean)}{stdev}$$

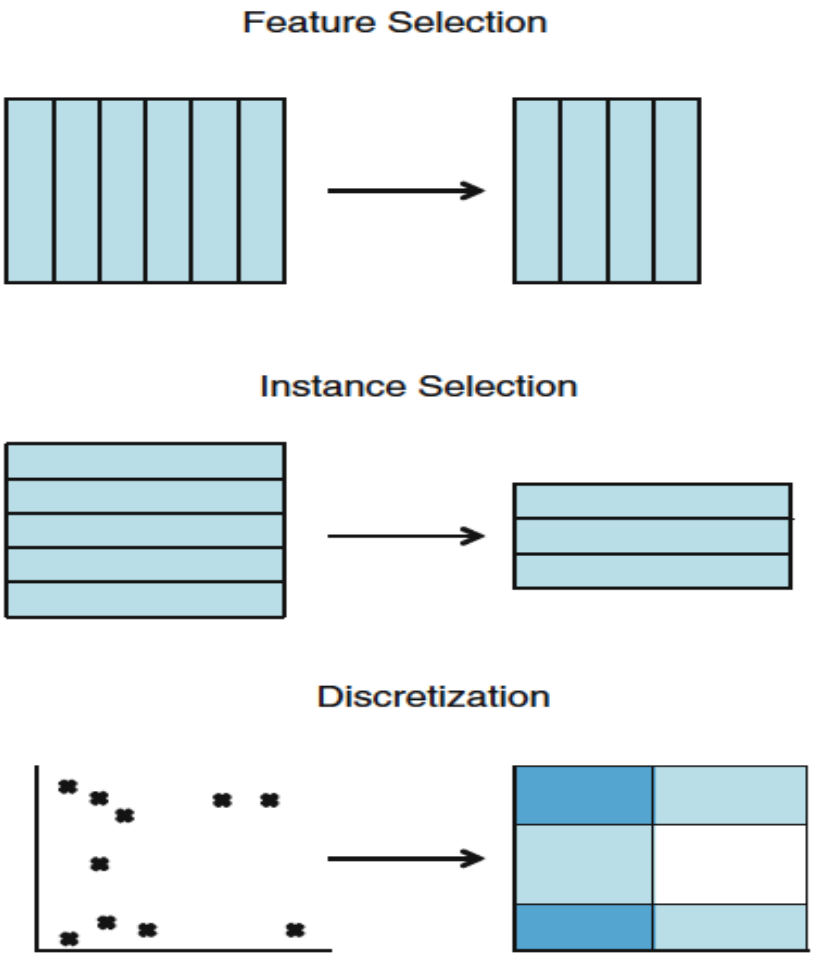
- Plot the columns of your data to understand whether you need to do normalization or standardization

Introduction to Data Preprocessing

- Noise Identification
  - To detect random errors or variances in a measured variable
  - Due to sensor noise

Introduction to Data Preprocessing

Forms of Data Reduction:



Introduction to Data Preprocessing

- Feature Selection
  - Achieves the reduction of the data set by removing irrelevant or redundant features (or dimensions).
- Instance Selection
  - Consists of choosing a subset of the total available data to achieve the original purpose of the ML application as if the whole data had been used.

Introduction to Data Preprocessing

- Discretization
  - Transforms quantitative data into qualitative data
  - numerical attributes into categorical attributes with a finite number of intervals.
- Feature Extraction/Instance Generation
  - Extends both the feature and instance selection by allowing the modification of the internal values that represent each example or attribute.

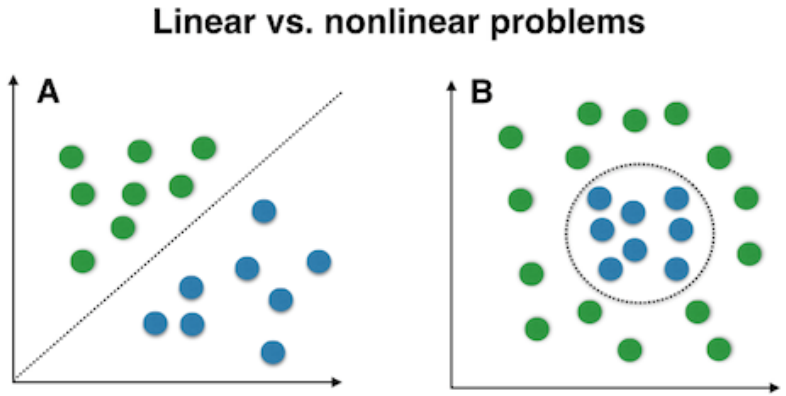
Introduction to Data Preprocessing

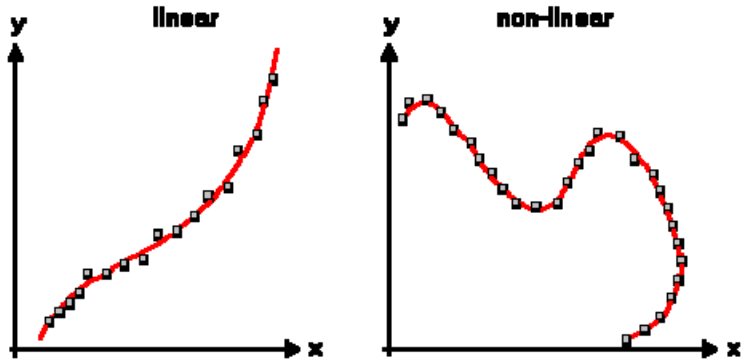
- Other preprocessing steps:
  - Detect and remove empty lines at the top or bottom of the file.
  - Detect and handle rows that do not match expectations for the rest of the file.
- **Hint:** use loadtxt() function from your NumPy to load data files into NumPy arrays.

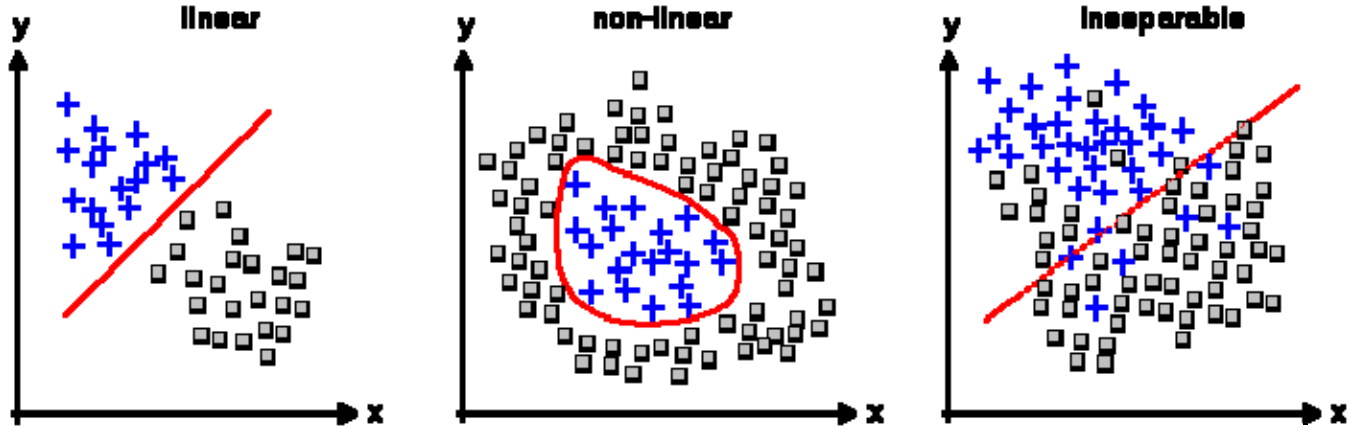
ML Algorithms

- Linear Algorithms
  - separate the output classes as linear combination of the input variables
  - Gradient descent optimization; Linear regression; Logistic regression
- Non-linear Algorithms
  - Output classes are not linear combination of input variables
  - Naive Bayes; K-NN Algorithm; Support vector machines
- Ensemble Algorithms
  - Combine predictions/classifications from multiple models
  - Random Forests; Boosting ensemble; AdaBoost algorithm

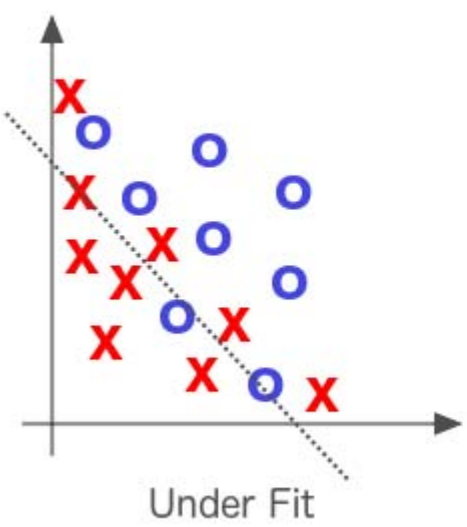
Linearly Separable data

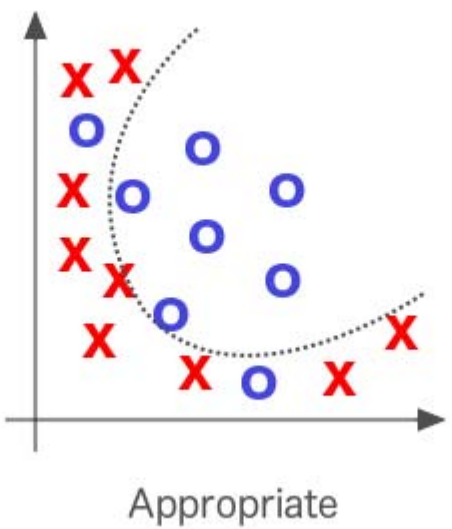


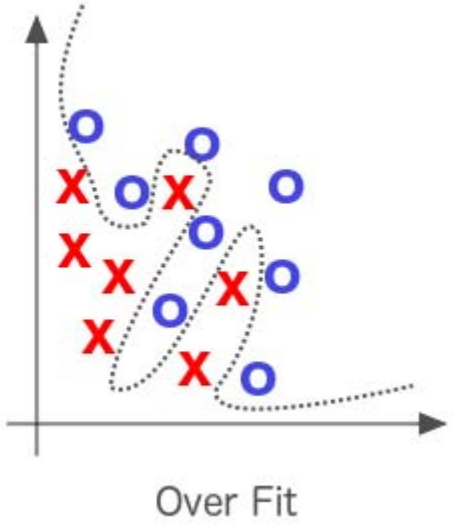




Overfitting and Under-fitting







ML Algorithms

- Parametric ML Algorithms

  1. Select a form for the function (probability distributions)

  2. Learn the coefficients for the function from the training data

  Adv: Simpler, fast and less data

  Disadv: Constrained by function and poor fit.

- Non-parametric ML Algorithms
  - No assumption about the form of data
  - Is Good when you have a lot of data and no prior knowledge
  - Nonparametric methods seek to best-fit training data in constructing the mapping

  Adv: Flexible, No assumptions on data, Performance

  Disadv: More data, slower, overfitting

Evaluation Methods for ML Algorithms

- We need to create models that does GOOD classification/predictions
- GOOD – choosing good parameters for your model.
- During training phase, we don’t have access to the new data to our algorithm.
- How can we make sure that our model does a GOOD job for the unknown data?
  - Use statistical methods to estimate the performance of our models

Evaluation Methods for ML Algorithms

- Train and Test Split
- k-fold Cross-validation
- Leave-one-out cross validation

Evaluation Methods for ML Algorithms

- Train and Test Split
  - Spit the dataset into training set and test set
  - Choose the data randomly during the split
  - How much to % split?
    - Training: 60%
    - Test: 40%
  - Noisy estimate of the algorithm's performance

Evaluation Methods for ML Algorithms

- k-fold Cross-Validation Split
  - Gold standard for estimating ML algorithms
  - Spit the dataset into k-groups
  - Train and test the algorithm for k-times
    - Train the algorithm on (k-1) group and evaluate on the k^{th} group
    - Each k-group is called as folds, and should have equal number of data samples
  - The overall performance is the mean of performance in k-folds
  - Choose value of k?
    - k=3 for small and k=10 for large datasets
    - Depends on the dataset size and computation time

Evaluation Methods for ML Algorithms

- Leave-one-out cross validation
  - Very powerful and expensive evaluation method
  - Set k=N in k-fold cross-validation method
  - N is the total data samples in the dataset
  - Every data sample will be test data once, and training data for k-1 times
  - Mean of all N performance
  - Computationally expensive method