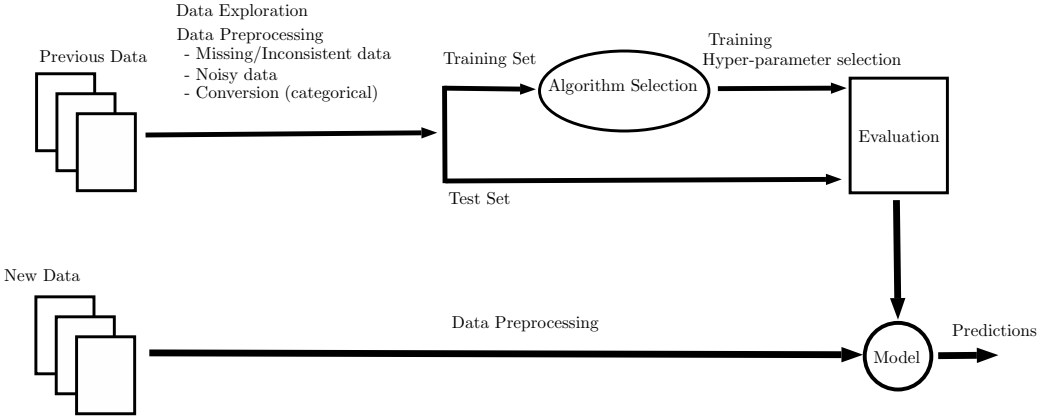
BABD

Masters in Business Analytics and Big Data

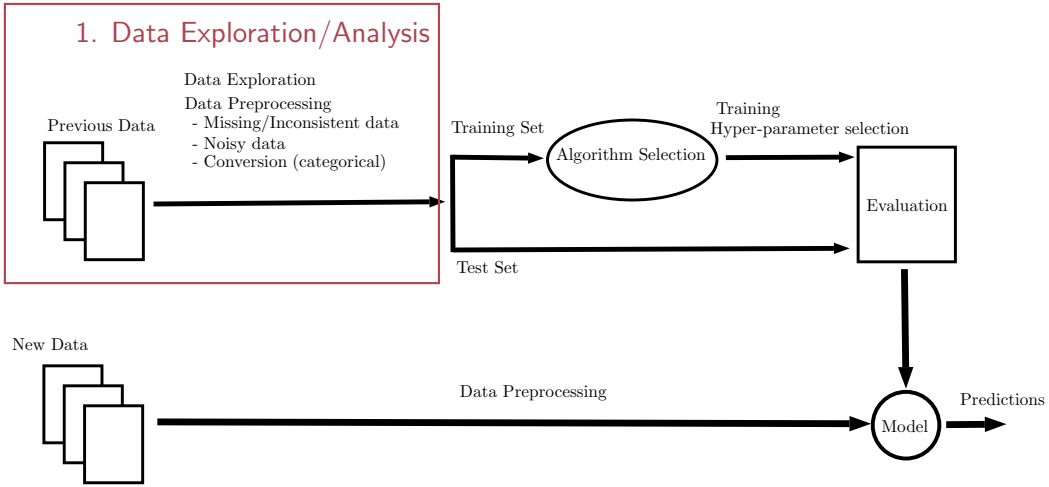
Data Preparation

Mauricio Soto - mauricioabel.soto@polimi.it

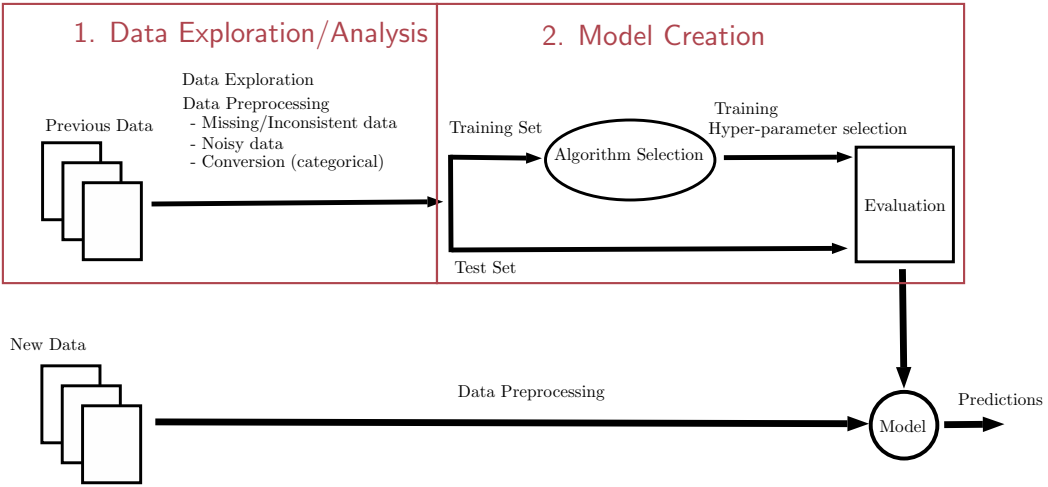
Workflow


Data Exploration
Data Preprocessing
- Missing/Inconsistent data
- Noisy data
- Conversion (categorical)
Previous Data
Training Set
Algorithm Selection
Training
Hyper-parameter selection
Evaluation
Test Set
New Data
Data Preprocessing
Model
Predictions

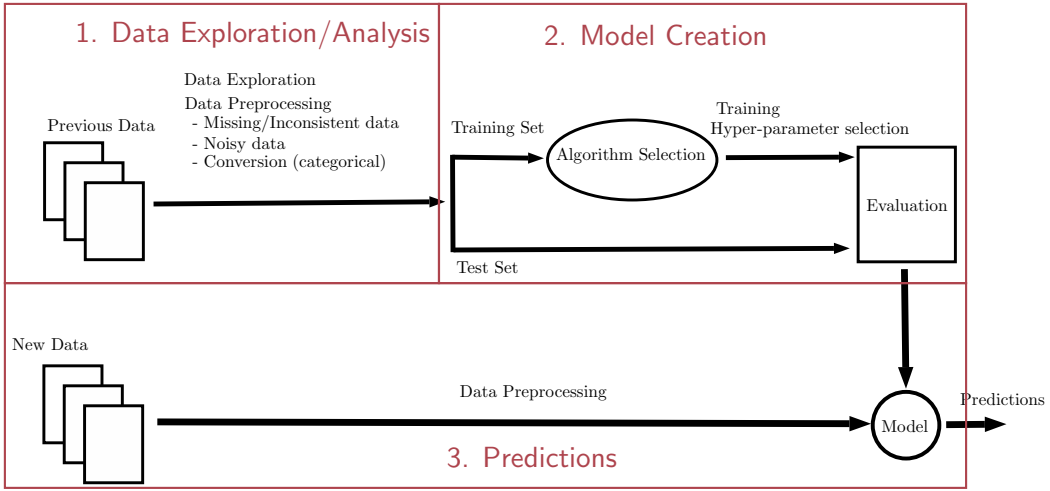
Workflow


1. Data Exploration/Analysis
Data Exploration
Data Preprocessing
- Missing/Inconsistent data
- Noisy data
- Conversion (categorical)
Previous Data
Training Set
Algorithm Selection
Training
Hyper-parameter selection
Evaluation
Test Set
New Data
Data Preprocessing
Model
Predictions

Workflow


1. Data Exploration/Analysis
2. Model Creation
Data Exploration
Data Preprocessing
- Missing/Inconsistent data
- Noisy data
- Conversion (categorical)
Previous Data
Training Set
Algorithm Selection
Training
Hyper-parameter selection
Evaluation
Test Set
New Data
Data Preprocessing
Model
Predictions

Workflow


1. Data Exploration/Analysis
Data Exploration
Data Preprocessing
- Missing/Inconsistent data
- Noisy data
- Conversion (categorical)
Previous Data
2. Model Creation
Training Set
Algorithm Selection
Training
Hyper-parameter selection
Evaluation
Test Set
New Data
Data Preprocessing
Model
Predictions
3. Predictions

Incomplete Data

- Inspection
- Elimination
- Identification
- Replacement
  - mean value of numerical attributes
  - mean value of the target class
  - value estimated sing statistical inference

Noisy Data

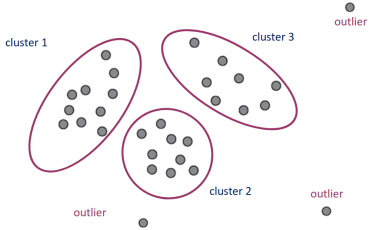
- Univariate
  - Normal-like distribution

    $$[\bar{\mu} - 2\bar{\sigma}, \bar{\mu} + 2\bar{\sigma}]$$

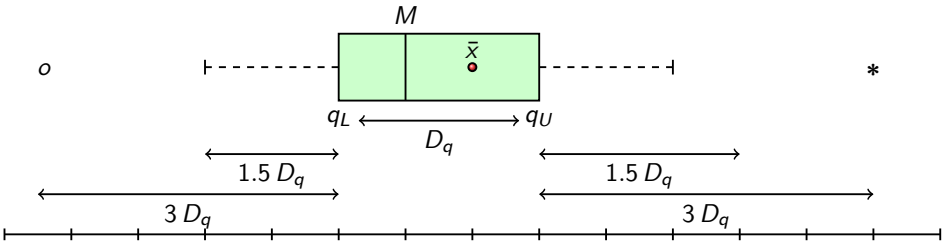
    contains about 96% of the data
  - In the general case, Tchebysheff theorem states taht for $\gamma > 1$

    $$[\bar{\mu} - \gamma\bar{\sigma}, \bar{\mu} + \gamma\bar{\sigma}]$$

    contains $1 - 1/\gamma^2$ proportion of the observations
- Multi variate
  - Clustering techniques



Box-plot



- $D_q = q_U - q_L = q_{0.75} - q_{0.25}$
- internal lower edge$= q_L - 1.5\, D_q$
- external lower edge$= q_L - 3\, D_q$

Data transformation

- **Decimal Scaling**

$$x'_{ij} = \frac{x_{ij}}{10^k}$$

- **Min-Max** in the interval $[x'_{\text{min},j}, x'_{\text{max},j}]$

$$x'_{ij} = \frac{x_{ij} - x_{\text{min},j}}{x_{\text{max},j} - x_{\text{min},j}}(x'_{\text{max},j} - x'_{\text{min},j}) + x'_{\text{min},j}$$

- z-**index**

$$x'_{ij} = \frac{x_{ij} - \bar{\mu}_j}{\bar{\sigma}_j}$$

Data reduction

- **Sampling**
  - Simple sampling
  - Stratified sampling
- **Selection**
  - Filter methods
  - Wrapper methods
  - Embedded methods
- **Discretization, Aggregation**
- **Projection** (ex. PCA)

PCA: Principal Component Analysis

- Covariance data matrix $V = X'X$
- $\bar{x}_{ij} = x_{ij} - \bar{\mu}_j$
- New components p_j obtained as a linear transformation of original data $p_j = X\, w_j$
- Variance of $p_j = w_j'\, X'X\, w_j = w_j' V w_j$
- Maximizing the variance:

$$\max_{w_1} w_1' V w_1 \text{ s.t. } w_1' w_1 = 1$$

- w_j is the j-th eigenvector of V, which explains a variance λ_j which is the j-th eigenvector

Nonlinear reduction