MANAGEMENT ACADEMY

MACHINE LEARNING WITH PHYTON

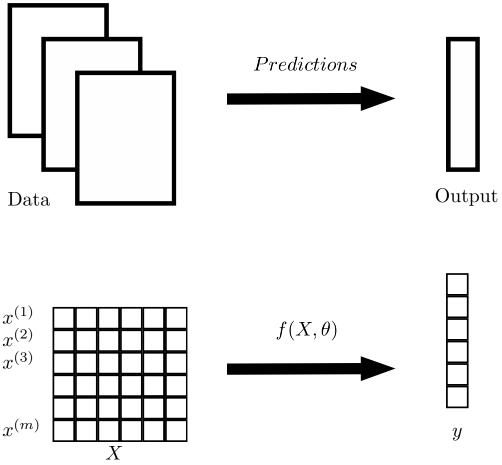
Mauricio Soto

POLITECNICO DI MILANO
GRADUATE SCHOOL
OF BUSINESS

The program

- Day 1
  - Pre-processing
  - Classification
- Day 2
  - Regression
  - Clustering

The objective

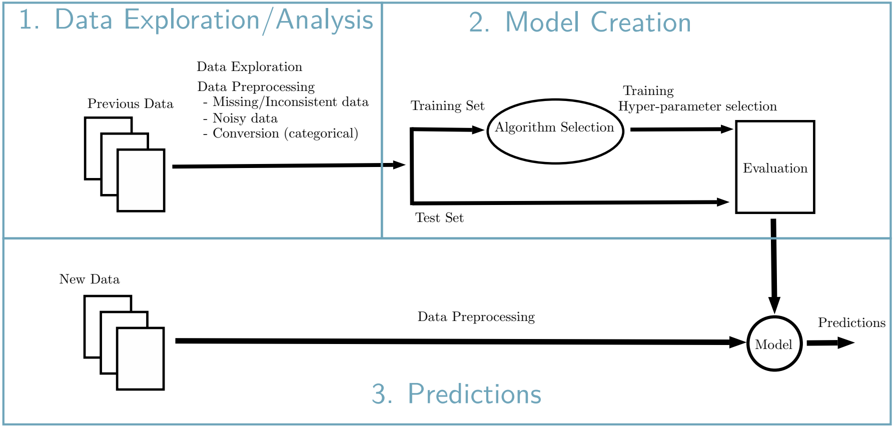


The problem: Bank telemarketing[1]

Attribute		Type	Description/Values
Personal	age	num	Age of the potential client
	job	cat	admin., blue- collar, entrepreneur, housemaid,... ,unknown
	marital_status	cat	divorced, married, single, unknown
	education	cat	basic.4y, basic.6y, basic.9y, high.school,.. unknown
Bank	default	cat	The client has credit in default: no,yes,unknown
	housing	cat	The client has a housing loan contract: no,yes,unknown
	loan	cat	The client has a personal loan: no,yes,unknown
Campain	contact	cat	Communication type: cellular,telephone
	month	cat	Last month contacted: jan, feb ,..., dec
	day_of_week	cat	Last contact day : mon, tue,..., fri
	duration	num	Last contact duration (in seconds)
	campain	num	Number of contacts performed during this campaign
	pdays	num	Number of days that passed by after last contact
	previous	num	Number of contacts performed before this campaign
	poutcome	cat	Outcome of the previous marketing campaign: failure,nonexistent,success
Economical	emp.var.rate	num	Employment variation rate in the last quarter
	cons.price.idx	num	Consumer price index in the last month
	cons.conf.idx	num	Monthly consumer confidence index
	euribor3m	num	Dayly Euro Interbank Offered Rate
	nr.employed	num	Number of employees in the last quarter
Target	**success**	**target**	**0: no, 1: yes**

[1] A data-driven approach to predict the success of bank telemarketing. S. Moroa, P. Cortez, P. Rita.Decision Support Systems, 62:22-31, 2014.

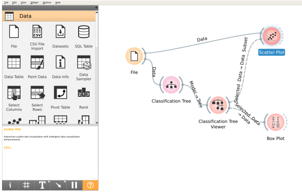
Workflow



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
3. Predictions

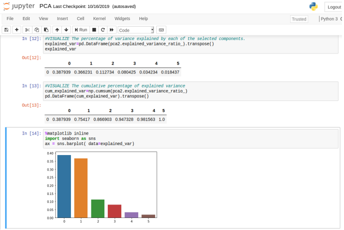
Programming Tools

1. Orange
https://orange.biolab.si/



- Intuitive interface
- Fast development

2. Jupyter-Notebook (Anaconda)
https://www.anaconda.com/



- Advanced functions
- Customization

A library featuring various ML algorithms designed to inter-operate with the Python numerical and scientific libraries e.g. NumPy, Pandas.
https://scikit-learn.org/stable/

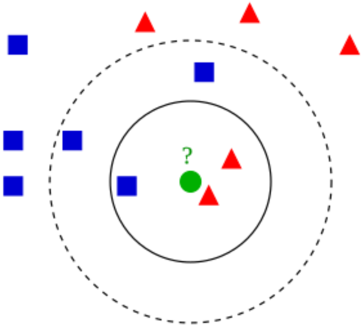
Data preparation

1. Data validation
   - Incomplete data (drop, replace)
   - Noisy data (Outliers)
2. Data transformation
   - Standardization
   - Discretization
   - Dummy variables
   - Feature construction
3. Data reduction
   - Sampling
   - Discretization
   - PCA: Principal Component Analysis

Data Exploration

1. Uni-variate
   - Histogram
   - Box-plot
2. Bi-variate
   - Scatter
   - Box-plot (by class)
3. Categorical
   - Contingency matrix

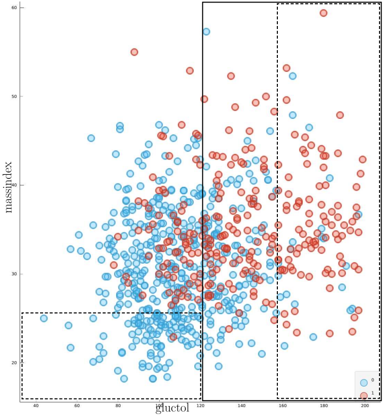
KNN K-nearest Neighbours

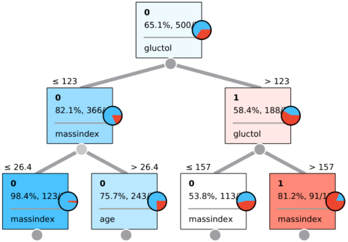


Main Parameters

- k : number of neighbours
- neighbour weights
- distances

Decision tree





Main Parameters

- impurity measure: "gini", "entropy"
- max_depth
- min_samples_split: minimum number of samples to split an internal node
- min_sample_leaf: minimum number of samples required to be at a leaf node

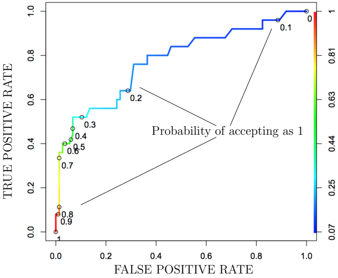
Quality measures

Prediction outcome

Actual value		**0**	**1**
	0	True Negative	False Positive
	1	False Negative	True Positive

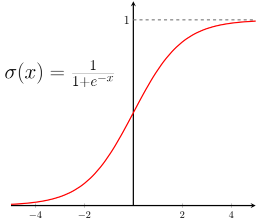
- Precision $= \frac{TP}{TP+FP}$
  "proportion of true positives among ***positive predictions****"*
- False Positive rate$=\frac{FP}{FP+TN}$
  "proportion of false positives among ***actual negatives****"*
- Recall (True Positive rate)$=\frac{TP}{FN+TP}$
  "proportion of true positives among ***actual positive****"*
- F-score$=\frac{2}{\frac{1}{\text{Precision}}+\frac{1}{\text{Recall}}}$

ROC curve & AUC



- If we accepting even with small probability then $TPR = FPR = 1$
- If we accepting just with high probability then $TPR = FPR = 0$
- The perfect classificator is the the point $(0, 1)$
- $AUC \in [0.5, 1]$ area under the curve is a quality measure of our algorithm.

Logistic regression



$$\log \frac{P(y=1|x)}{P(y=0|x)} = w_0 + w_1 x_1 + \cdots + w_n x_n$$

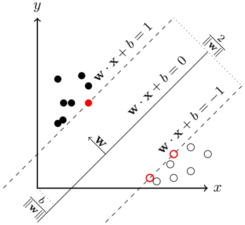
$$P(y=0|x) = \frac{1}{1+e^{w^\top x}}$$

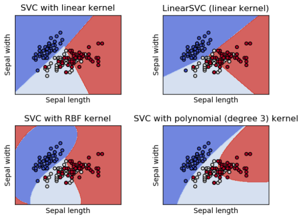
$$\min_w \underbrace{\frac{1}{2}||w||^2}_{\text{regularization}} + C \sum_{i=1}^{n} \log(1+\exp(-y_i(w^T X_i)))$$

Main Parameters

- C: Inverse of regularization strength
- Resolution algorithm parameters:
  - solver: lbfgs, newton-cg, liblinear, sag, saga.
  - tol: Tolerance for stopping criteria.
  - max_iter: max. number of iterations int
  - n_jobs: Number of CPU cores

Support Vector Machine - SVM





$$\min_{w,b,d} \quad \frac{1}{2}||w||^2 + C\sum_{i=1}^{m} d_i$$

$$\text{subject to } y_i(w^T \underbrace{\phi(x_i)}_{\text{kernel}} - b) \geq 1 - d_i,$$

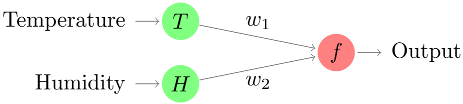
$$d_i \geq 0$$

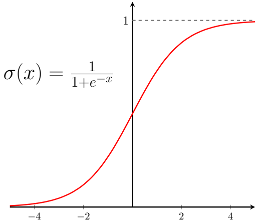
Main Parameters

- C: Inverse of regularization strength
- kernel:
  - linear: $x'x$
  - poly: $(\gamma x'x + r)^d$
  - rbf: $exp(-\gamma||x - x'||^2)$
  - sigmoid: $\tanh(\gamma x'x + r)$
- degree(d), gamma(γ), coef0(r)
- Resolution algorithm parameters

Multi-Layer Perceptron - small example

Temp. [C]	20	31	15	18	21
Humidity [%]	40	36	23	45	30
Prob. Rain	0.70	0.52	0.55	0.73	0.60





$$f(T, H, w_1, w_2) = \underbrace{\sigma}_{\text{activation function}}(w_1 \cdot T + w_2 \cdot H)$$

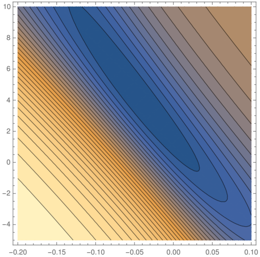
$$\max_{w_1, w_2} \sum_{i=1}^{5} [P_i(\text{rain}) - f(T_i, H_i, w_1, w_2)]^2$$

For a classification problem we can use the Likelihood as cost function.

Multi-Layer Perceptron - small example

$$\min_{w_1, w_2} \sum_{i=1}^{5} [P_i(\text{rain}) - f(T_i, H_i, w_1, w_2)]^2$$

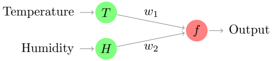
$$= \min \left[0.7 - 1/(1 + e^{-(w_1 \cdot 20 + w_2 \cdot 0.4)})\right]^2 + \left[0.52 - 1/(1 + e^{-(31 \cdot w_1 + w_2 \cdot 0.36)})\right]^2 + \cdots$$

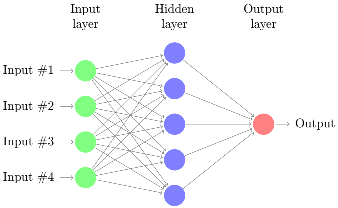


$$(w_1^*, w_2^*) = (-0.044, 4.147)$$

Temp. [C]	20	31	15	18	21
Humidity [%]	40	36	23	45	30
Prob. Rain	0.70	0.52	0.55	0.73	0.60
Predicted	0.70	0.56	0.58	0.75	0.60
Error	**0.0**	**-0.04**	**-0.03**	**-0.02**	**0.0**

Multi-Layer Perceptron





Main Parameters

- hidden_layer_sizes: $(n_1, n_2, \ldots, n_L)$
- activation: identity, logistic, tanh, relu
- alpha regularization term parameter
- Resolution algorithm parameters: solver, tol, batch_size, learning_rate, max_iter.

Ensemble: Random Tree