BABD

INTERNATIONAL MASTER IN BUSINESS ANALYTICS AND BIG DATA

Classification - 01

POLITECNICO DI MILANO
GRADUATE SCHOOL
OF BUSINESS

Cefriel
POLITECNICO DI MILANO

IBM

FT Executive Education Ranking 2019

FT European Business Schools Ranking 2018

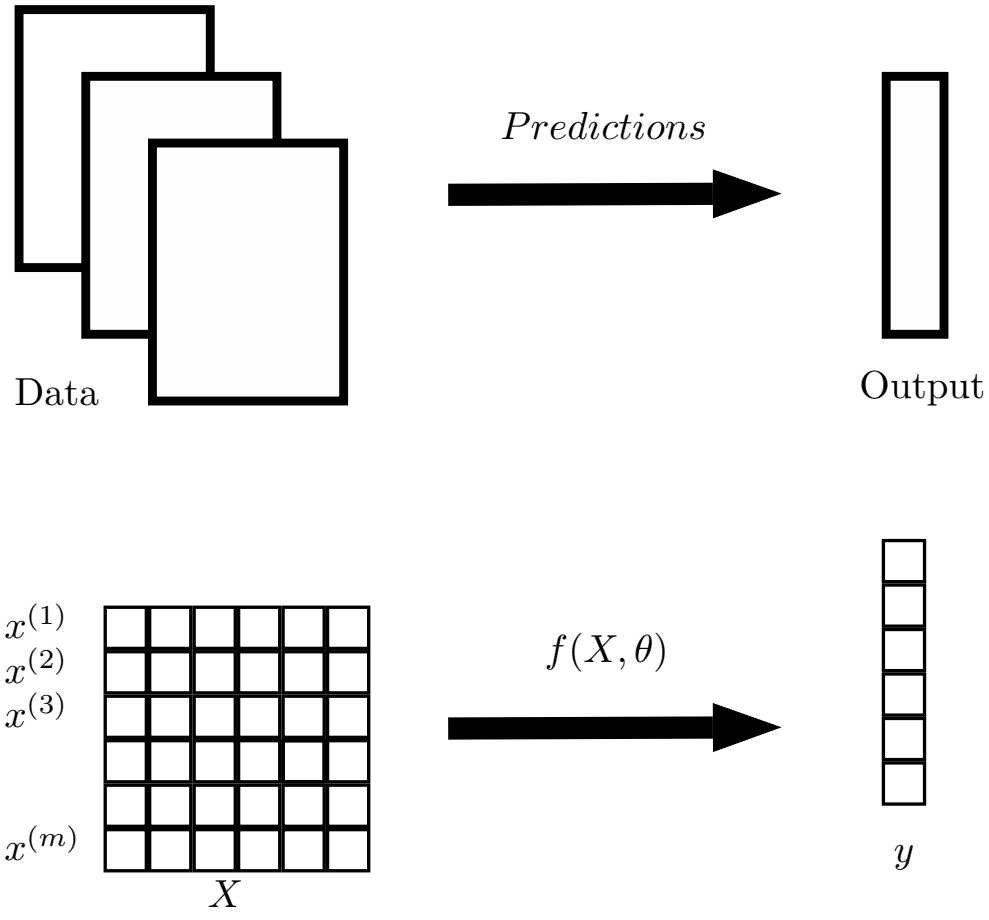
Bloomberg Businessweek 2018 BEST B-SCHOOLS

EFMD EQUIS ACCREDITED

ASSOCIATION AMBA ACCREDITED

EFMD GLOBAL NETWORK EOCCS CERTIFIED

Supervised Learning



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 employed citizens in the last quarter (thousands)
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.

Classification problem

attributes

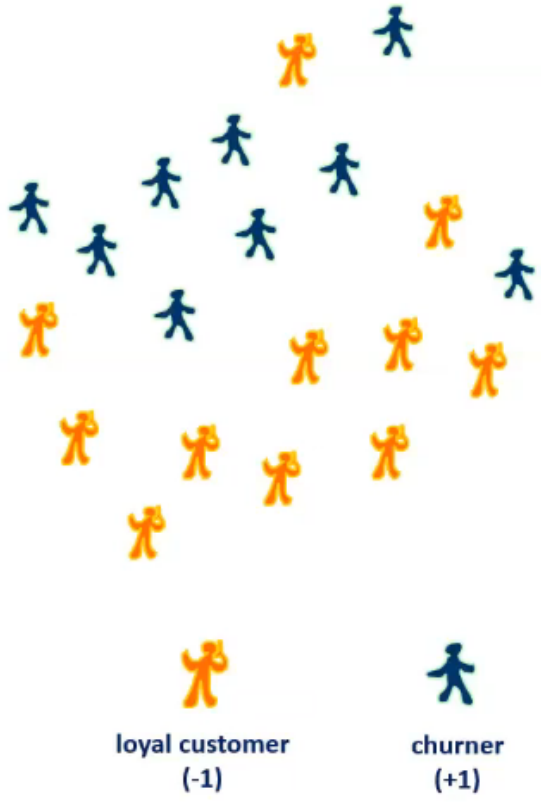
customers

Area	Pothers	Pmob	...	NumSMS	Class
2	0.14	0.59	...	18	1
3	0.26	0.35	...	9	-1
1	0.37	0.23	...	1	1
⋮	⋮	⋮	⋮	⋮	⋮
4	0.41	0.27	...	64	-1

← past data →

Area	Pothers	Pmob	...	NumSMS	Class
1	0.27	0.67	...	36	?
4	0.44	0.22	...	50	?
4	0.31	0.47	...	14	?
⋮	⋮	⋮	⋮	⋮	⋮
2	0.31	0.14	...	49	?

← future data →

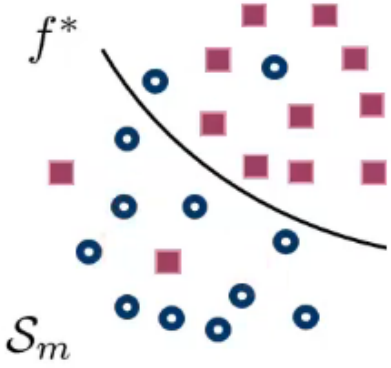


Classification formulation

$\mathcal{S}_m = \{(\mathbf{x}_i, y_i),\ i \in \mathcal{M}\}$: **training set, where** $\mathbf{x}_i \in \Re^n$ **and** $y_i \in \mathcal{D}$

$\mathcal{H}$ **denotes a set of functions** $f(\mathbf{x}) : \Re^n \mapsto \mathcal{D}$

Classification problem: define a hypotheses space $\mathcal{H}$ **and a function** $f^* \in \mathcal{H}$ **which optimally describes the relationship between** $\mathbf{x}_i$ **and** y_i



new patient

$\Rightarrow$ $f^* \in \mathcal{H}$ $\Rightarrow$ $\overline{\mathbf{x}}$ $\Rightarrow$ $f(\overline{\mathbf{x}}) = -1$

healthy !

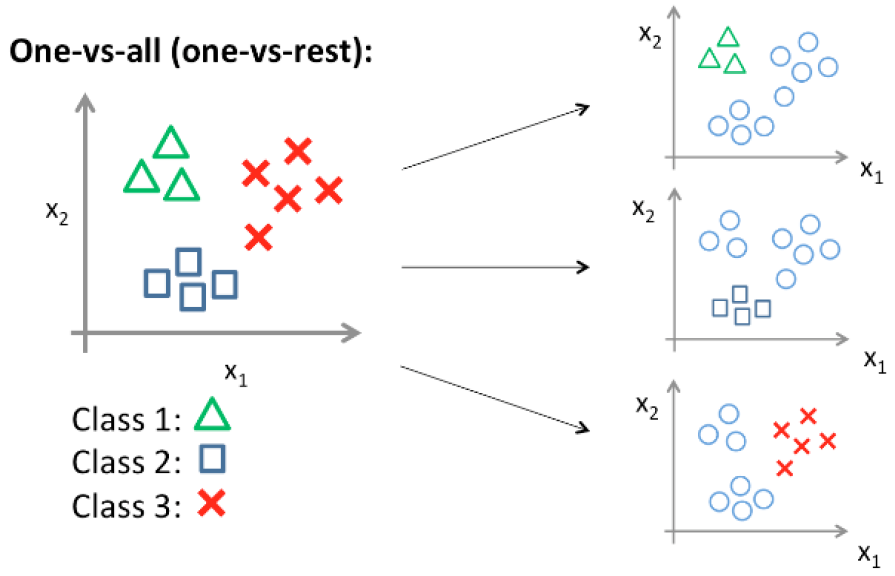
learning prediction

healthy patient (-1)

ill patient (+1)

Multi-class classification

1. **One-vs-Rest** We perform $|H|$ different binary classifications: one for every class.



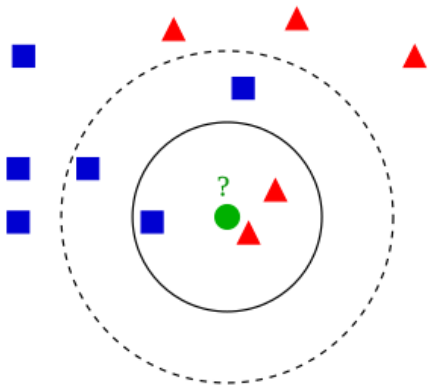
[1]

   We decide based on a majority vote.

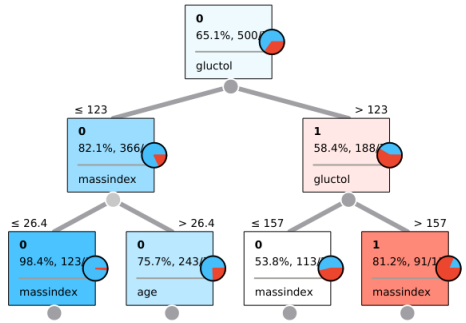
2. **One-vs-One** We perform $|H|(|H-1|)/2$ binary classifications: one for every pair of classes. We decide based on a majority vote.

[1] Image via www.cc.gatech.edu

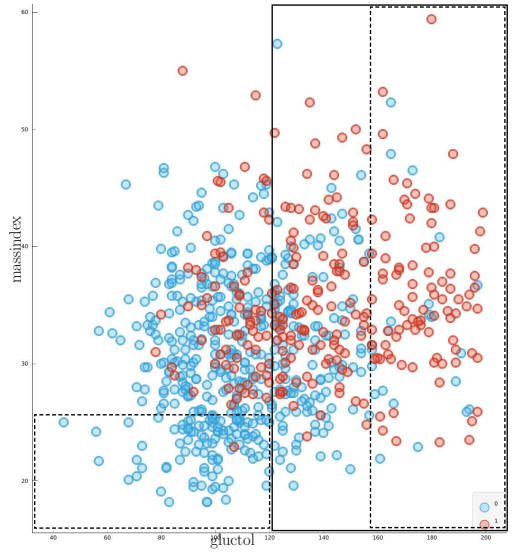
Classification Models

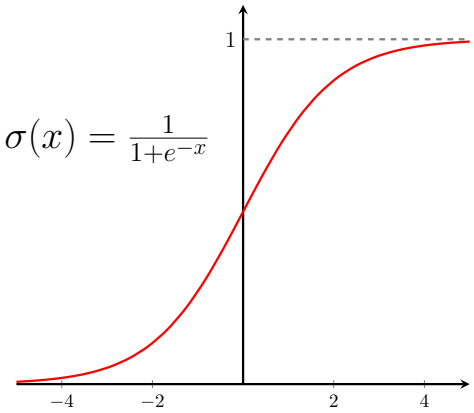


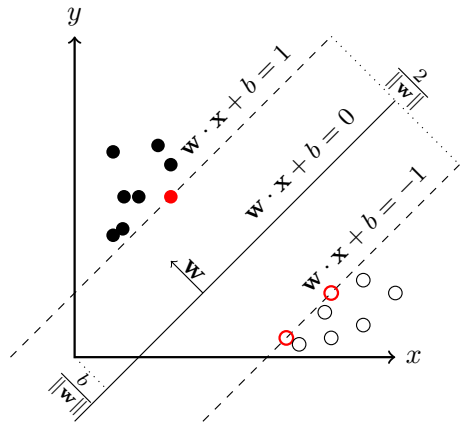
KNN

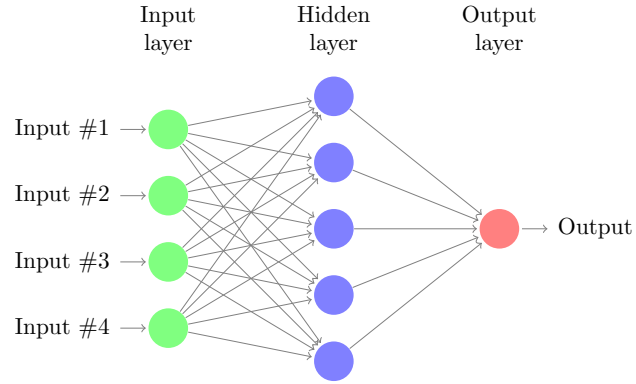


Classification Tree









$$\min_w \underbrace{\frac{1}{2}||w||^2}_{\text{regularization}} + C \underbrace{L(y,x)}_{\text{likelihood}}$$

BABD

Logistic Regression | SVM | Neural Networks

Classification Models

- Heuristics Methods
  - Nearest Neighbours
  - Classification Trees
- Probabilistic Methods
  - Bayesian Methods
- Regression Methods
  - Logistic regression
- Separation Methods
  - Support vector machine
  - Perceptron
  - Neural Networks

Evaluation Dimensions

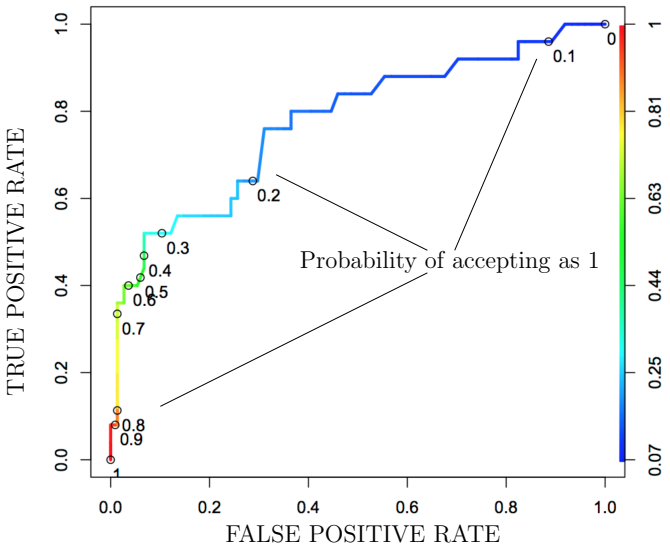
- Prediction accuracy
- Speed
- Robustness
- Scalability
- Intepretability
- Rules effectiveness

Classification - Quality measures - Confusion Matrix

Actual value \ Prediction outcome	0	1
0	True Negative	False Positive
1	False Negative	True Positive

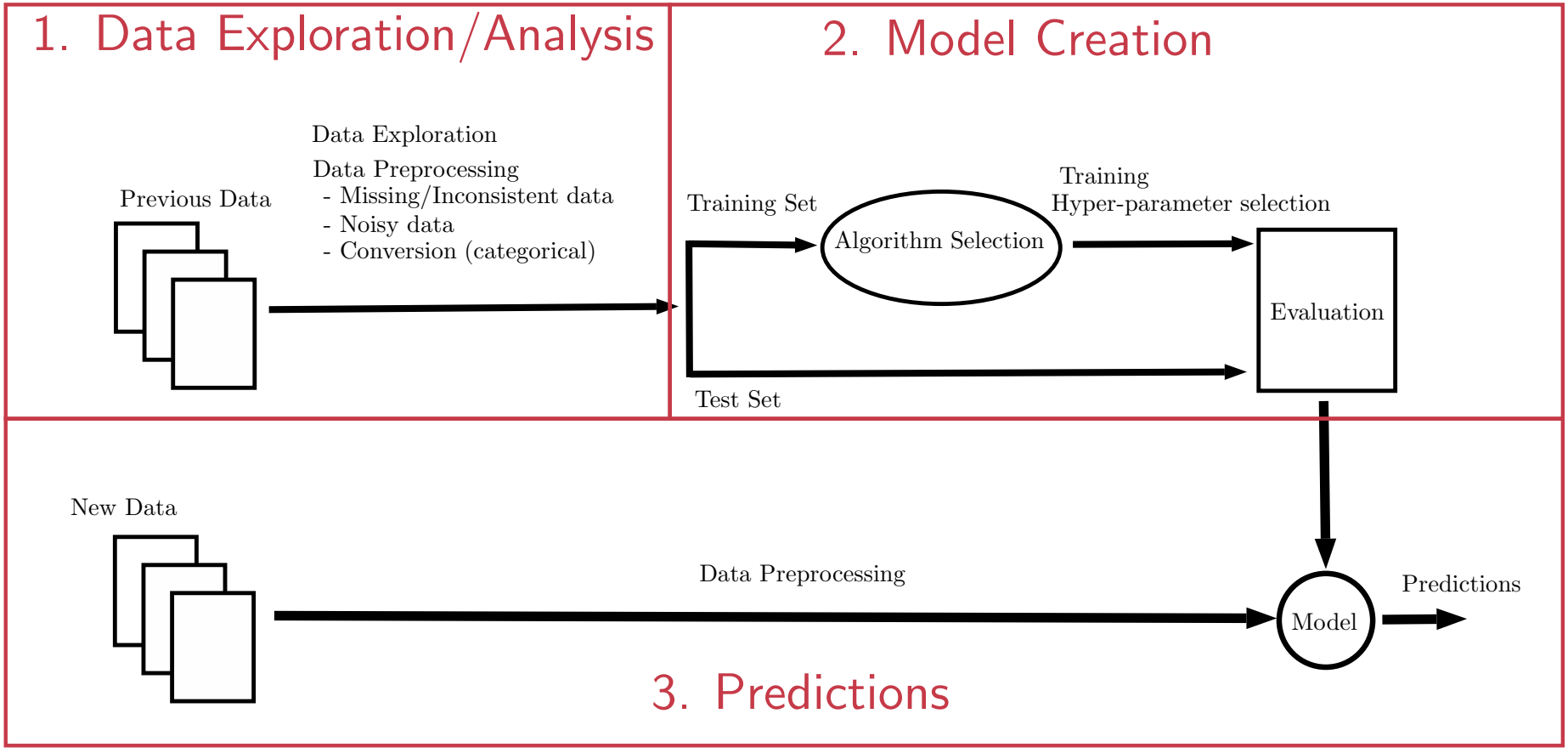
- Precision $= \frac{TP}{TP+FP}$
  "proportion of true positives among **positive predictions**"
- False Positive rate$=\frac{FP}{FP+TN}$
  "proportion of false positives among **actual negatives**"
- Recall (True Positive rate)$=\frac{TP}{FN+TP}$
  "proportion of true positives among **actual positive**"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

Classification - Quality measures - ROC curve & AUC

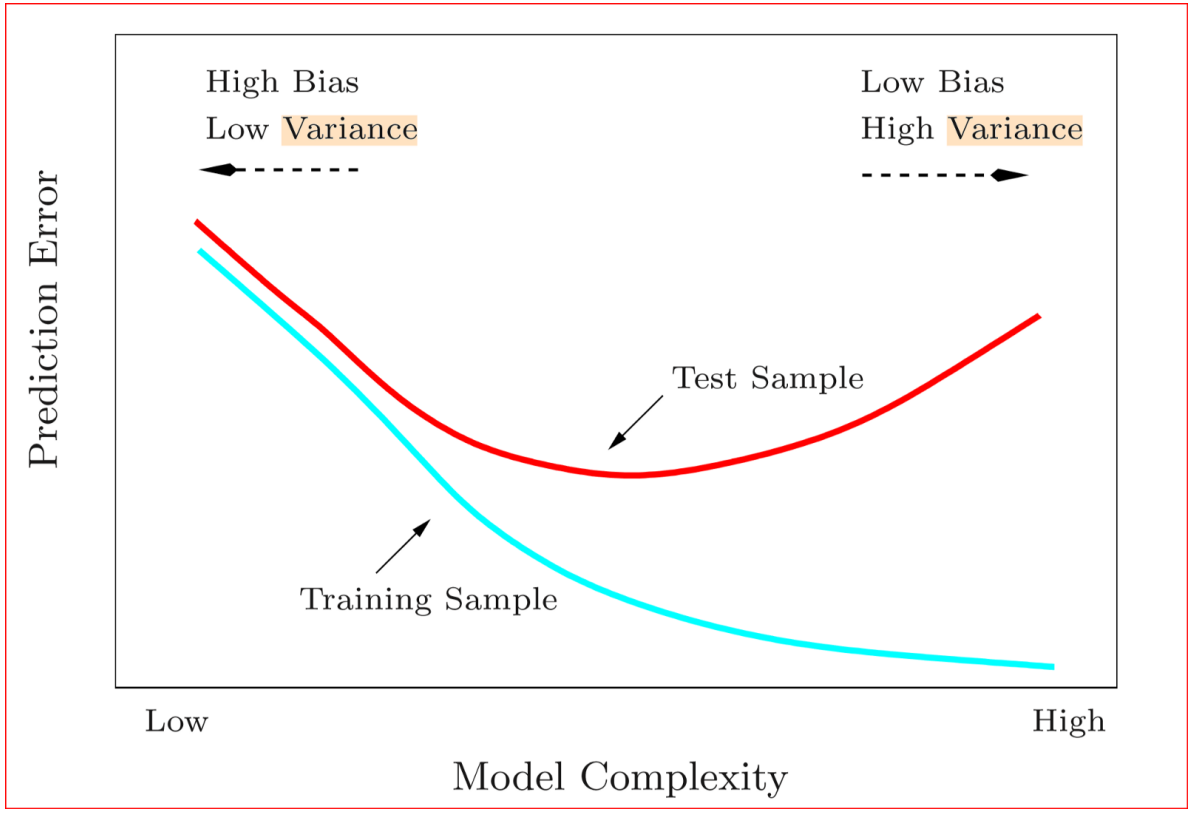


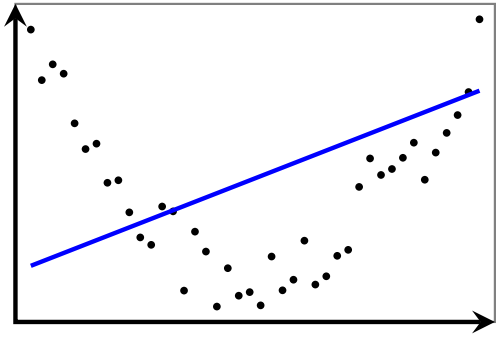
- If we accepting even with small probability then $TPR = FPR = 1$
- If we accepting just with high probability then $TPR = FPR = 0$
- The perfect classificator is the the point $(0, 1)$
- $AUC \in [0.5, 1]$ area under the curve is a quality measure of our algorithm.

Workflow

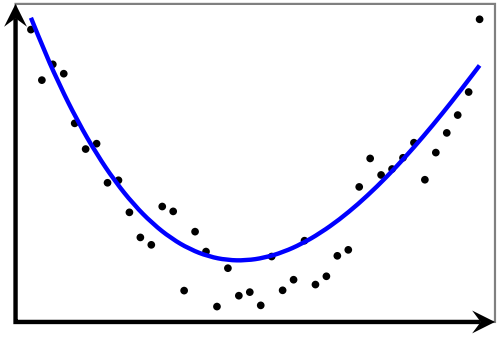

1. Data Exploration/Analysis
2. Model Creation
Data Exploration
Data Preprocessing
- Missing/Inconsistent data
- Noisy data
- Conversion (categorical)
Previous Data
Training Set
Algorithm Selection
Training
Hyper-parameter selection
Evaluation
Test Set
New Data
Data Preprocessing
Model
Predictions
3.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

Under/Over-fitting

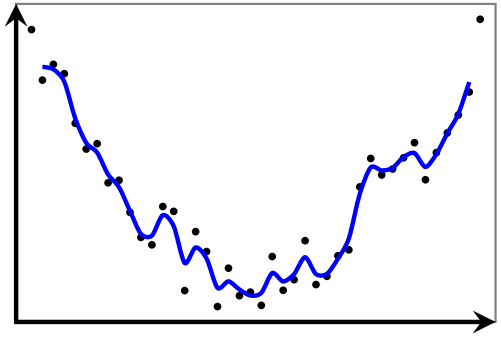




Underfitting

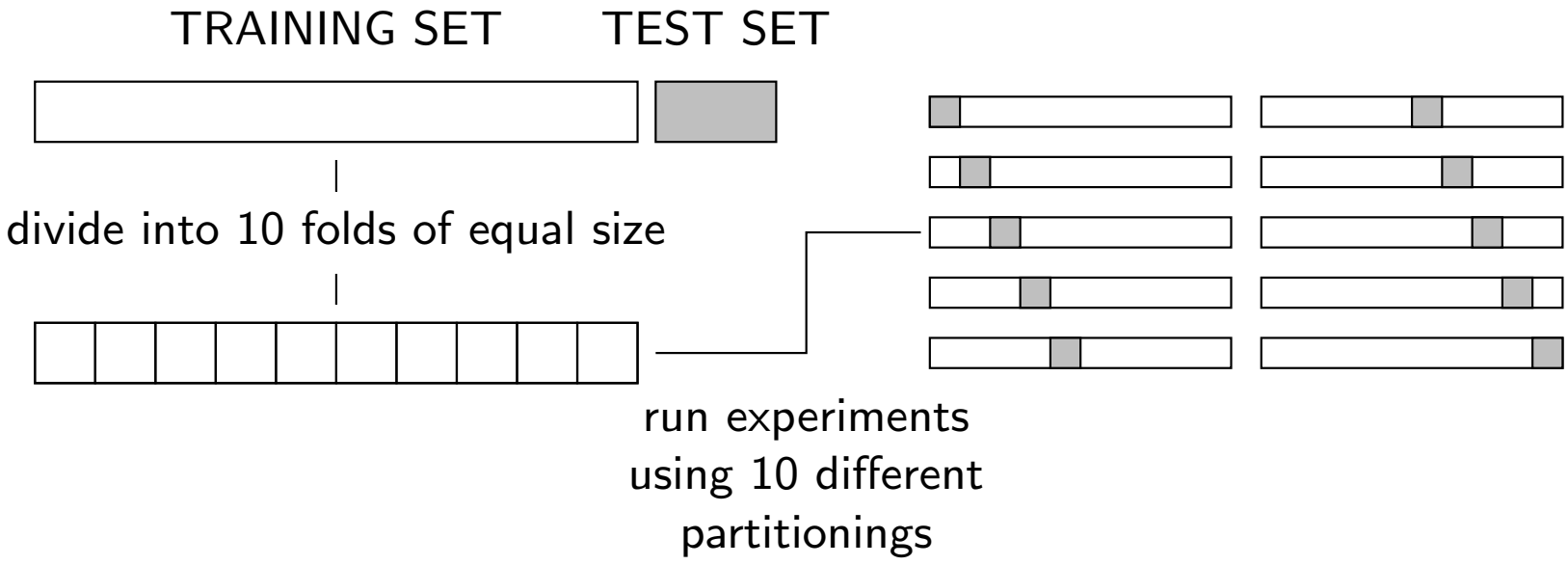


Balance

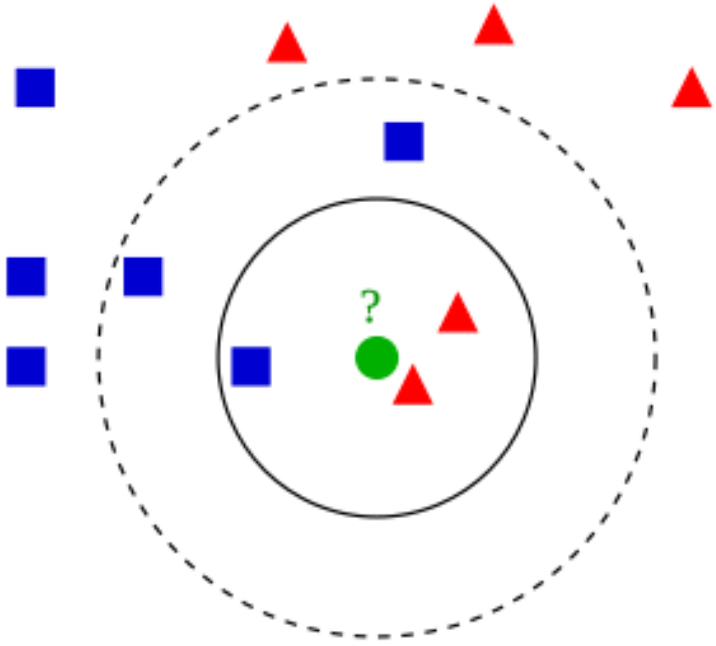


Overfitting

Cross validation



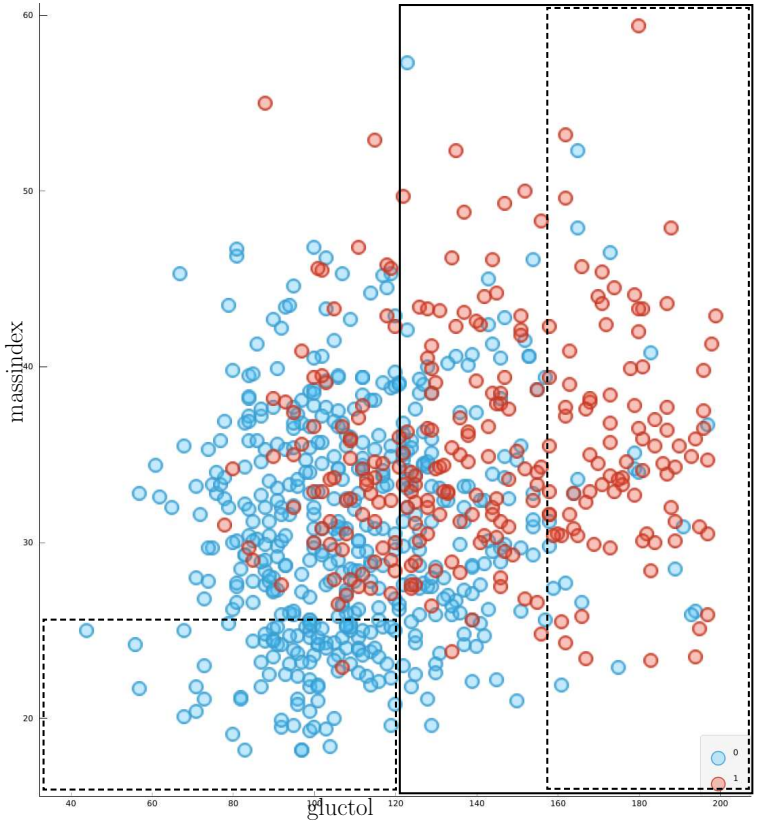
KNN K-nearest Neighbours

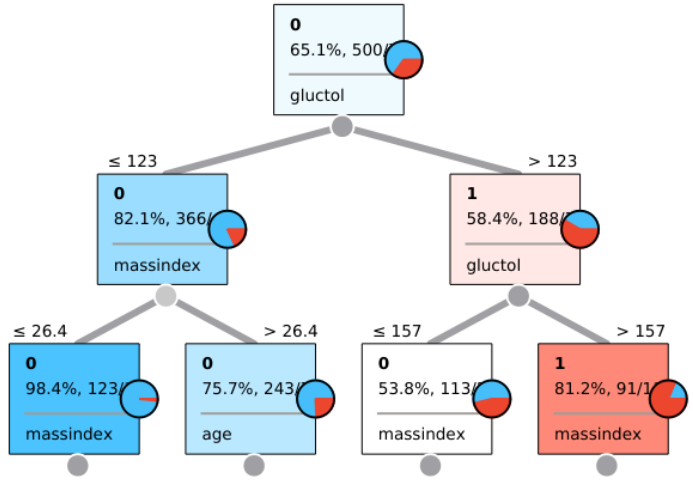


Main Parameters

- k : number of neighbours
- neighbour weights
- distances

Classification tree

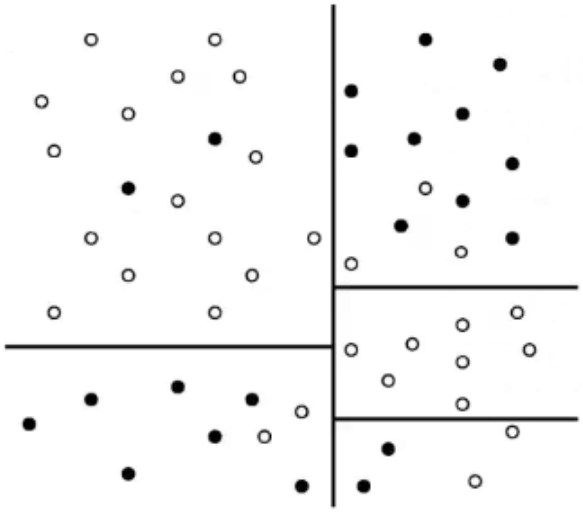




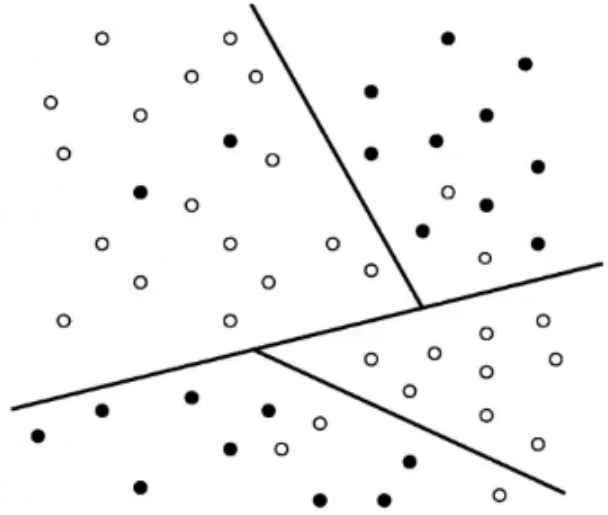
Tree types

- Binary tree (zero/two descendants)
- General trees
- Uni-variate tree ($X_j < b$)
- Multi-variate tree ($\sum_{j=1}^{n} w_j x_j = b$)

Classification tree

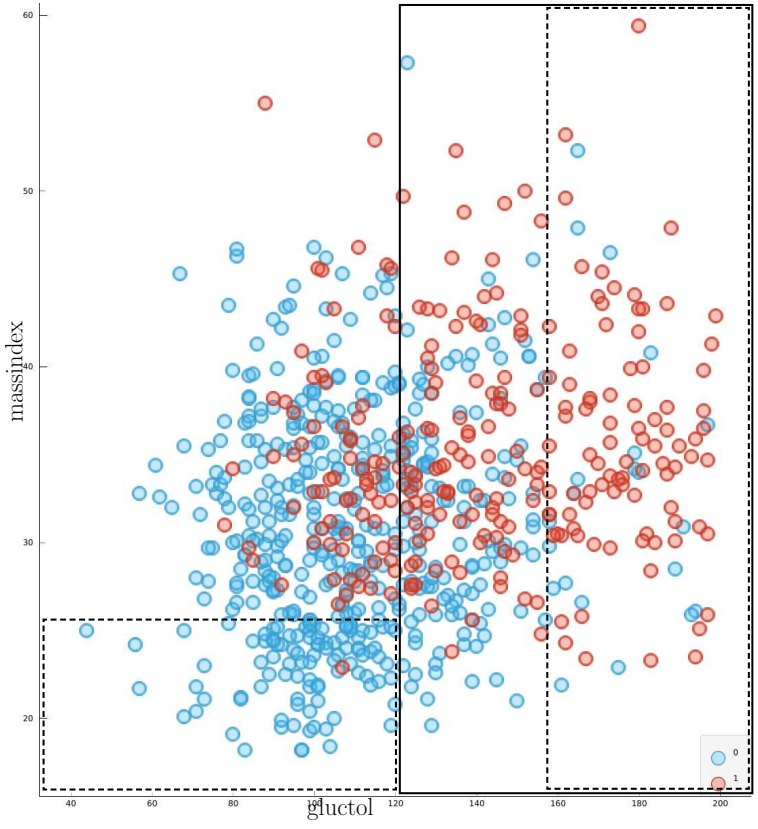


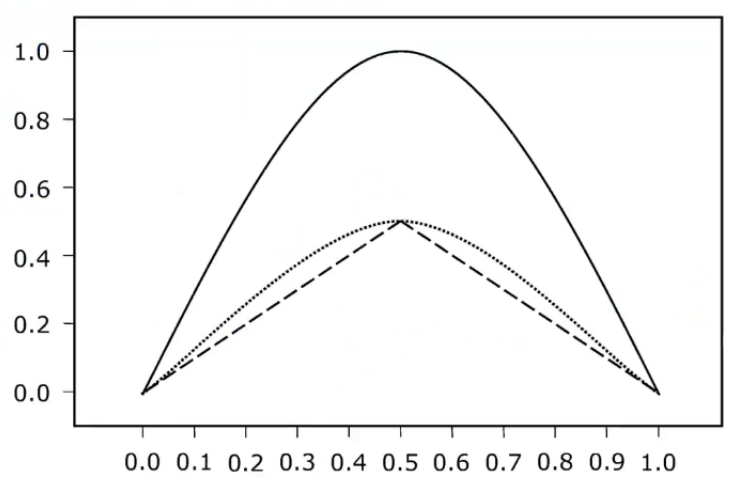
classification by an
axis parallel tree



classification by an
oblique tree

Classification tree





Split criteria

- **Gini index:** $1 - \sum_{h=1}^{H} f_h^2$
- **Entropy index:** $-\sum_{h=1}^{H} f_h \log_2 f_h$
- **Miss-classification index:** $1 - \max_h f_h$

Classification tree

- **Impurity of a splitting rule**

$$I(q_1, q_2, \ldots, q_K) = \sum_{k=1}^{K} \frac{Q_k}{Q} I(q_k).$$

- **At each node select the rule minimizing the impurity or, equivalently, maximizing the information gain**

$$\begin{aligned}\Delta(q, q_1, q_2, \ldots, q_K) &= I(q) - I(q_1, q_2, \ldots, q_K) \\ &= I(q) - \sum_{k=1}^{K} \frac{Q_k}{Q} I(q_k).\end{aligned}$$