IRL 3629

Initiation à l'apprentissage automatique en science des matériaux

4. Classification

J.-C. Crivello, LINK : jean-claude.crivello@cnrs.fr
C. Barreteau, ICMPE : celine.barreteau@cnrs.fr
S. Junier, ICMPE : sebastien.junier@cnrs.fr

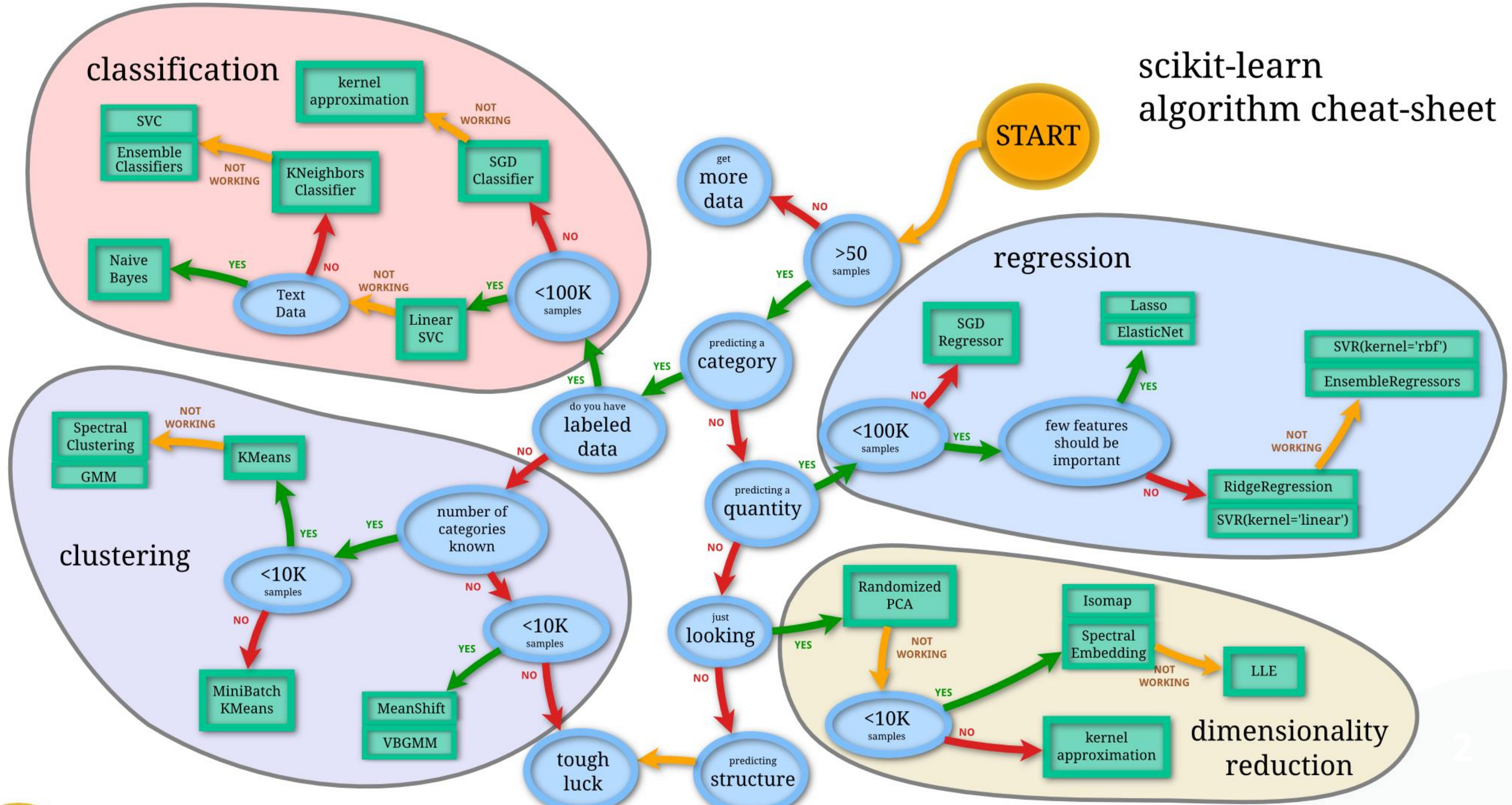
https://link.cnrs.fr/ML/

Several approaches of the data science

- Data-mining algorithms
- Optimization algorithms
- Machine learning algorithms (ML)
  - Reinforcement learning
  - Unsupervised learning (*clustering*)
  - Supervised learning

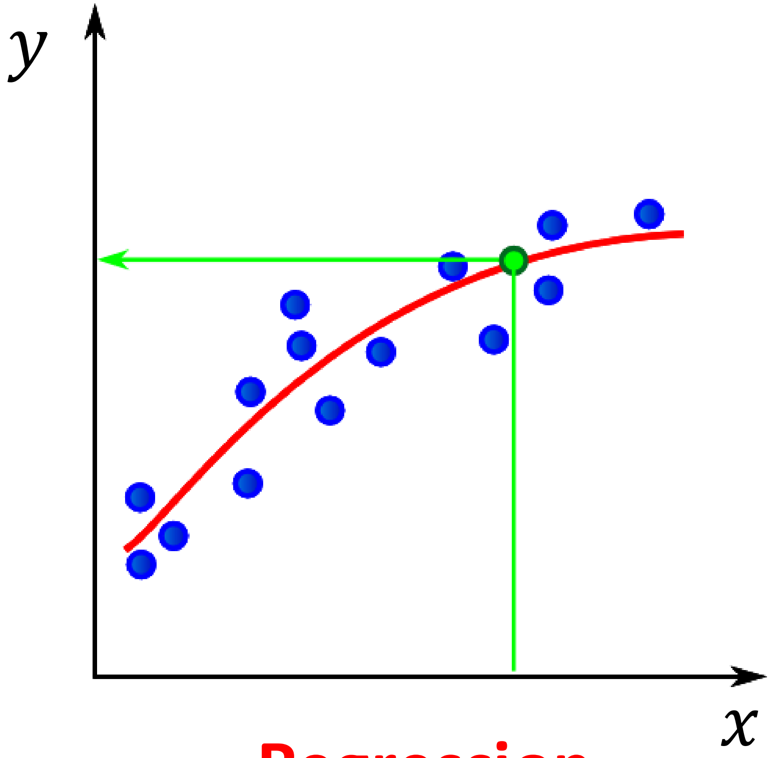
Choosing the right estimator?

https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

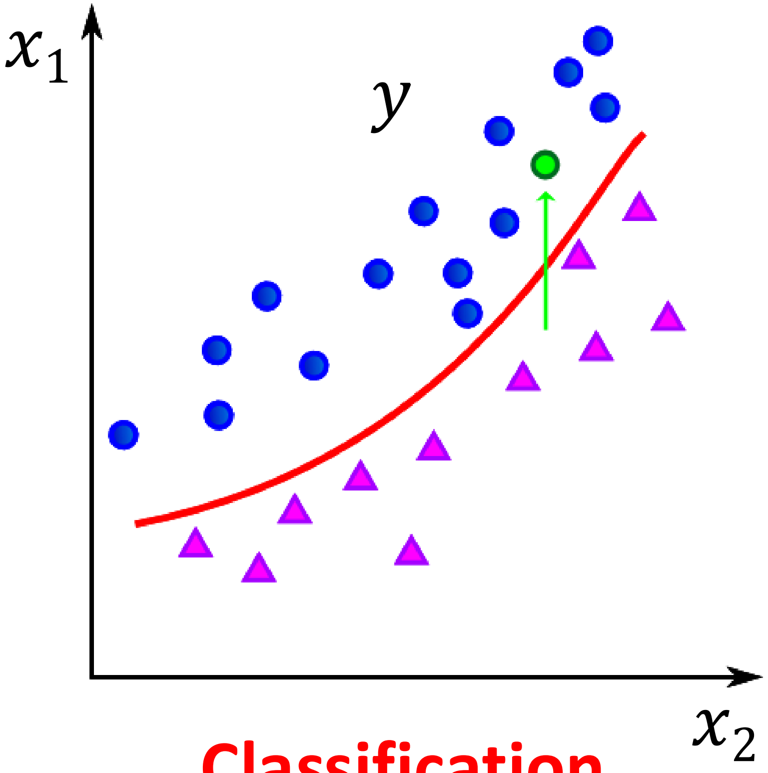


3. Supervised learning

The training dataset often consists of pairs of an input vector (or scalar) and the corresponding output vector (or scalar), the output of the function can be **regression** or **classification**.

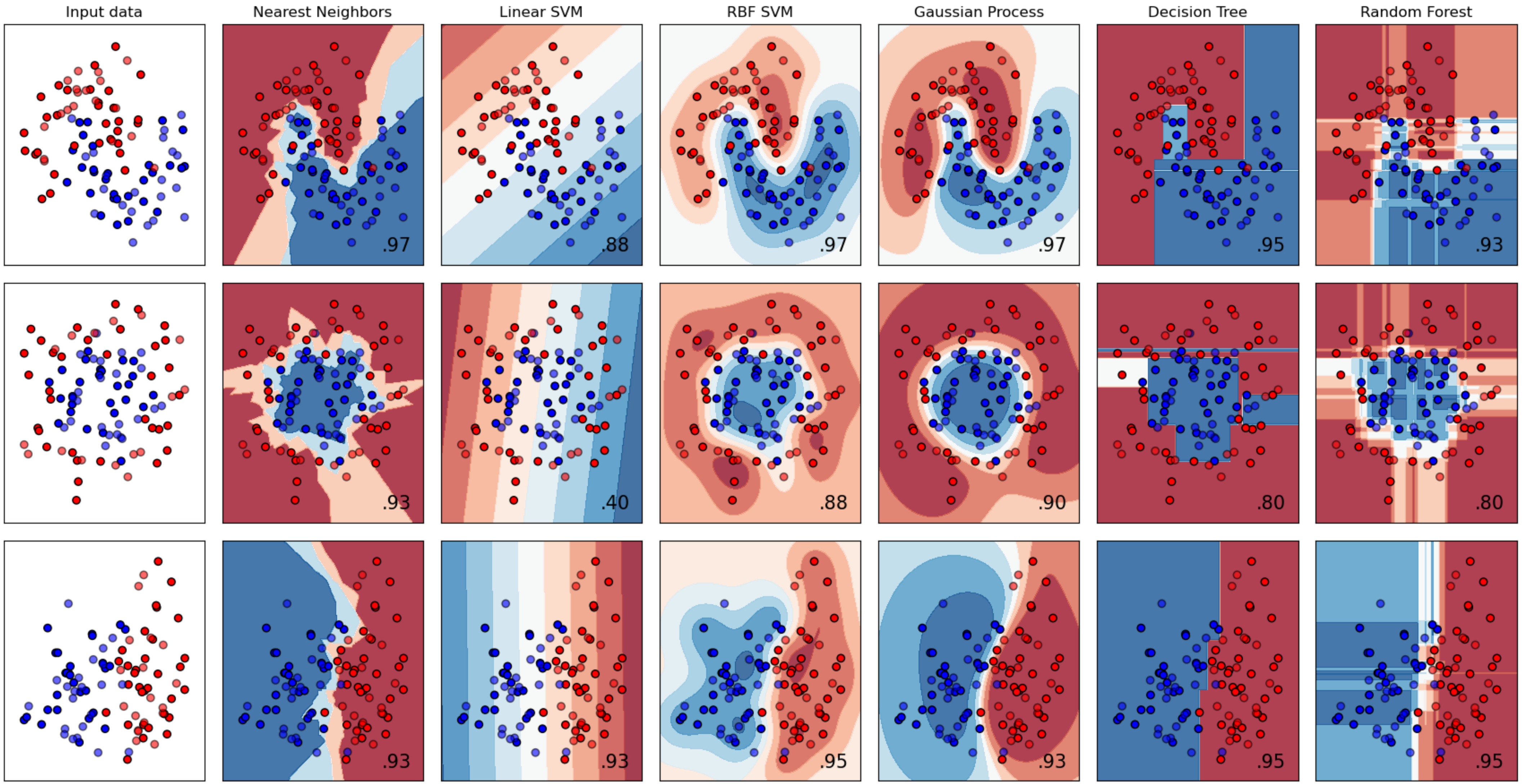


Regression

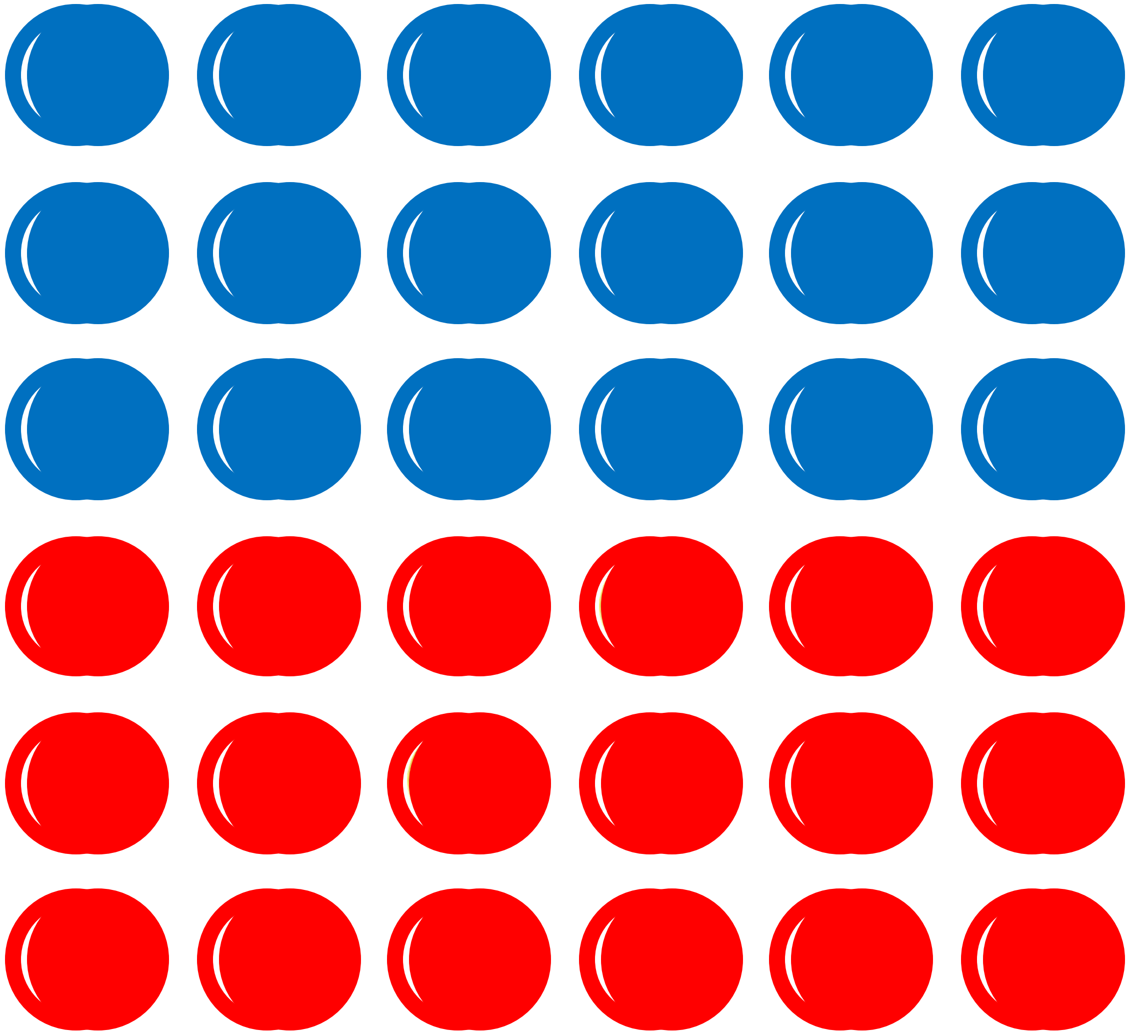


Classification

Binary classifier comparison



Naïve Bayes Classifier



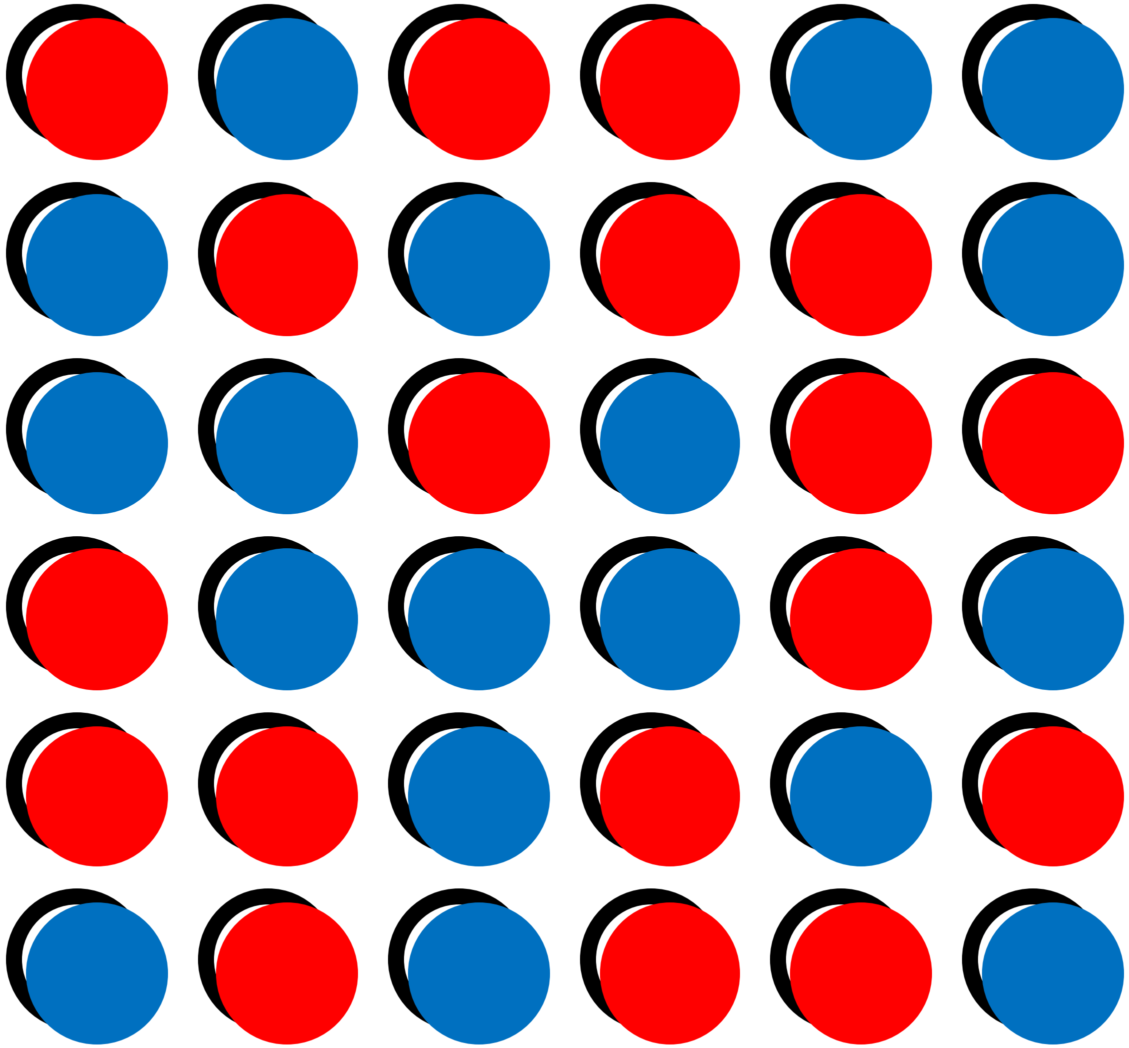
$p(A|B)$?

Ok if

$p(B|A)$

known

Naïve Bayes Classifier



$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A).P(A)}{P(B)}$$

Independence hypothesis

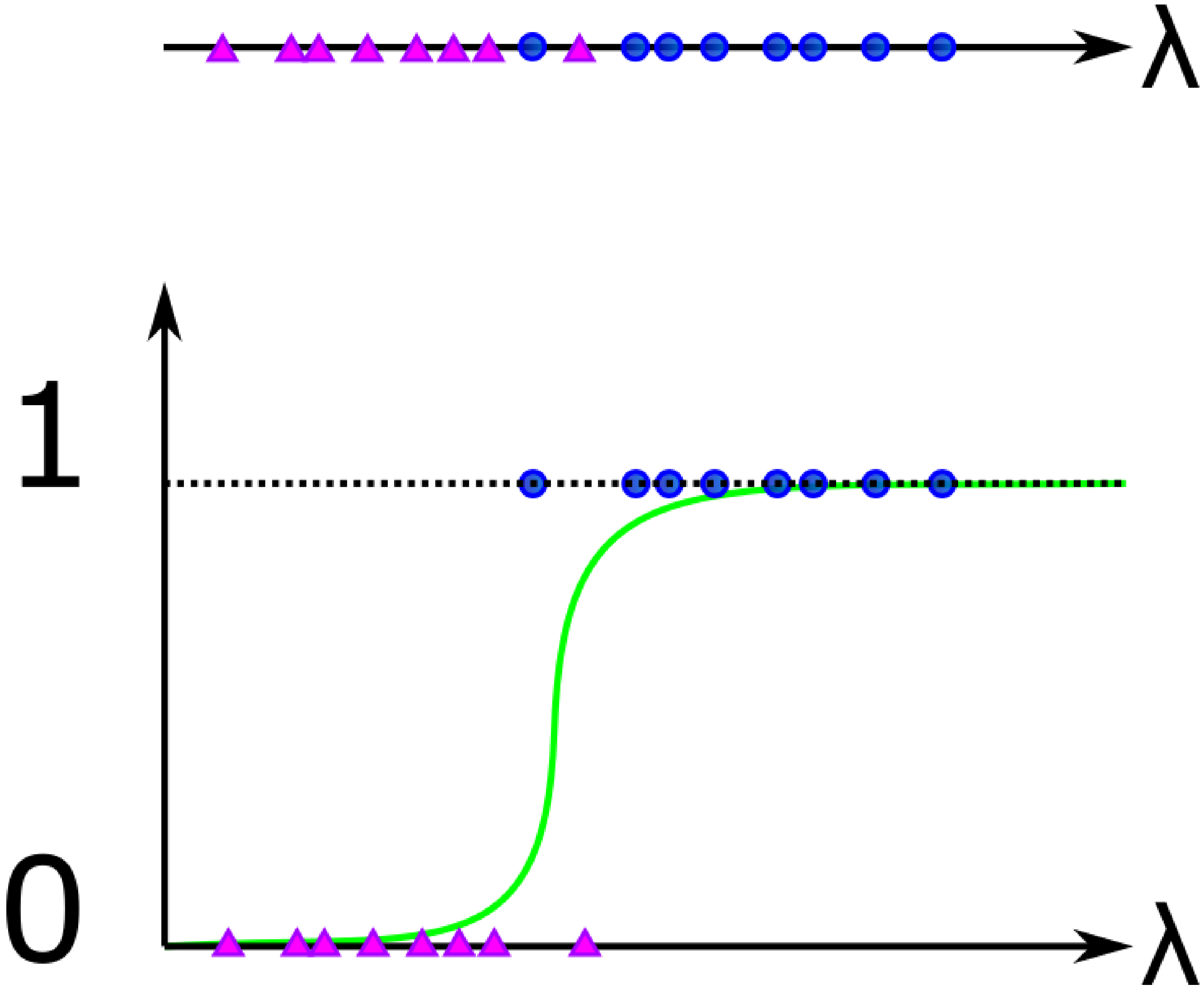
Strengths:

- Good performance
- Even with few data

Weaknesses:

- Not valid if condition of independence is not valid

Logistic regression



Is one of most used
Is a part of generalized linear model (GLM)

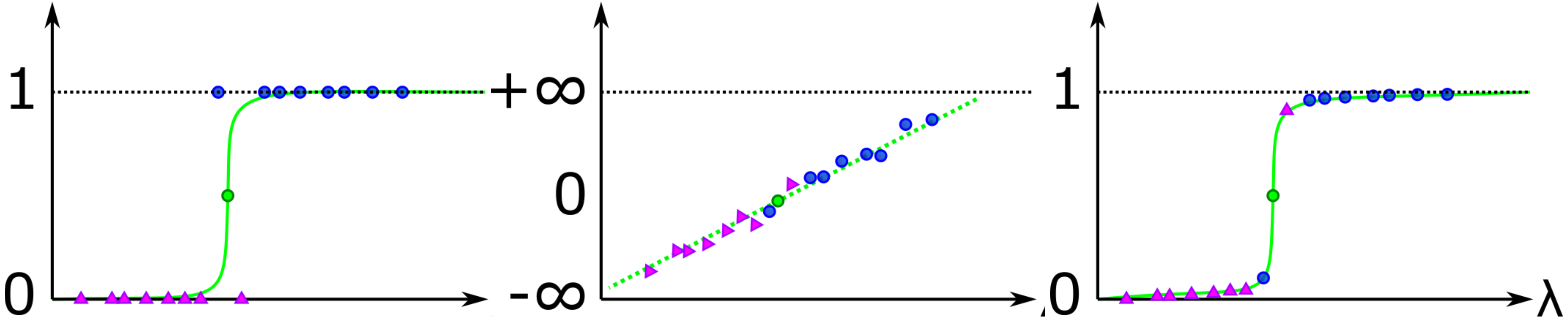
$$p(y=1|X)=\sigma(w^{\mathrm{T}}X+b)$$

Sigmoid function:

$\boldsymbol{\sigma}$**(z)=** $\frac{1}{1+e^{-z}}$

as a partition function

Logistic regression

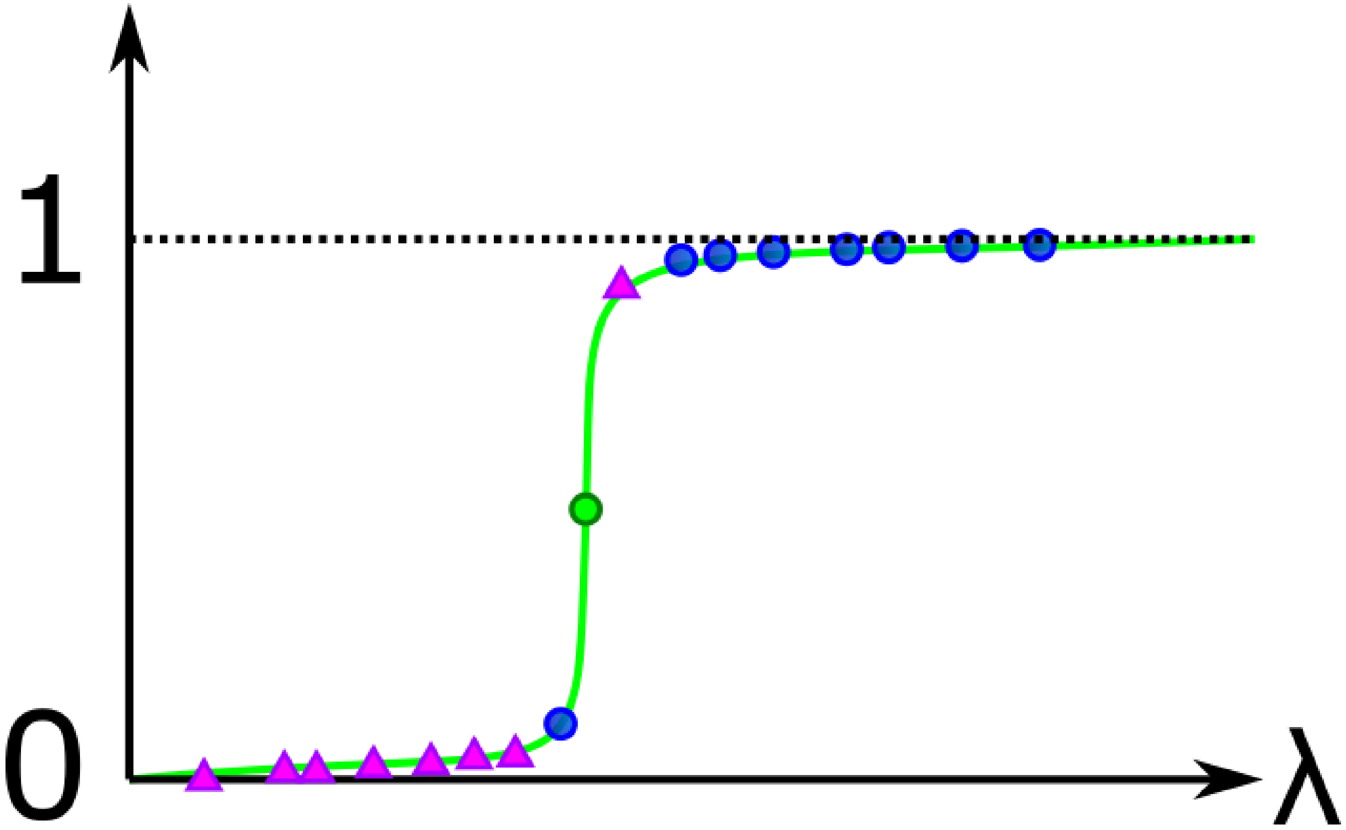


Log odd ratio : Log(p/1-p)

Coefficient ? Similar to a regression

Optimization by Maximum Likelihood estimation

Confusion matrix



		Predicted Values	
		●	▲
Actual Values	●	7	1
	▲	1	7

Performance evolution metrics

- **True Positives**: outcome correctly predicted as positive class
- **True Negatives**: outcome correctly predicted as negative class
- **False Positives**: outcome incorrectly predicted as positive class
- **False Negatives**: outcome incorrectly predicted as negative class

		Predicted Values	
		Positive	Negative
Actual Values	Positive	TP	FN
	Negative	FP	TN

$$\text{TP} = \sum_{i=1}^{n} 1_{y_i=1,\hat{y}_i=1}$$

$$\text{TN} = \sum_{i=1}^{n} 1_{y_i=-1,\hat{y}_i=-1}$$

$$\text{FN} = \sum_{i=1}^{n} 1_{y_i=1,\hat{y}_i=-1}$$

$$\text{FP} = \sum_{i=1}^{n} 1_{y_i=-1,\hat{y}_i=1}$$

Confusion matrix

		Predicted $\hat{y}$	
		1	- 1
True y	1	TP	FN
	-1	FP	TN

Performance evolution metrics

		Predicted Values	
		Positive	Negative
Actual Values	Positive	TP	FN
	Negative	FP	TN

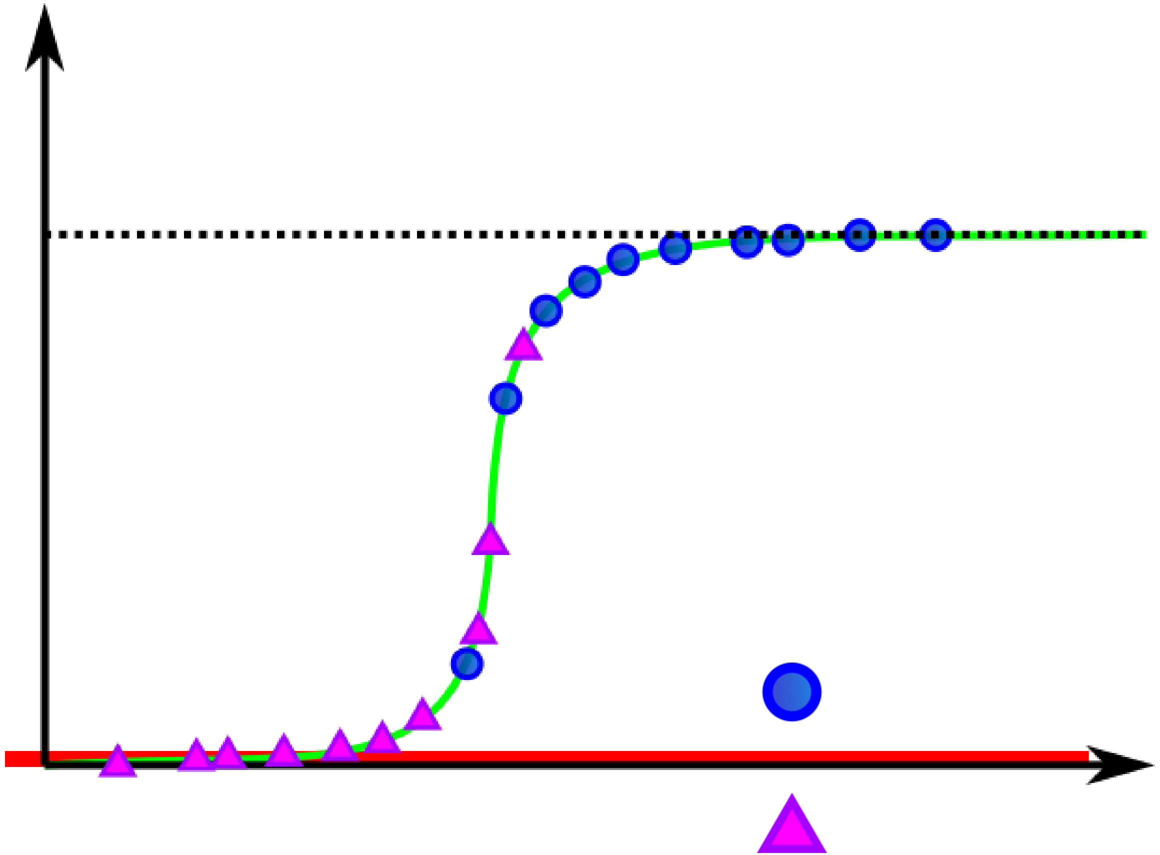
Performance Metric	*Formula*
Accuracy	(TP + TN) / (TP + TN + FP + FN)
Precision	TP / (TP + FP)
Recall (Sensitivity)	TP / (TP + FN)
Specificity	TN / (FP + TN)

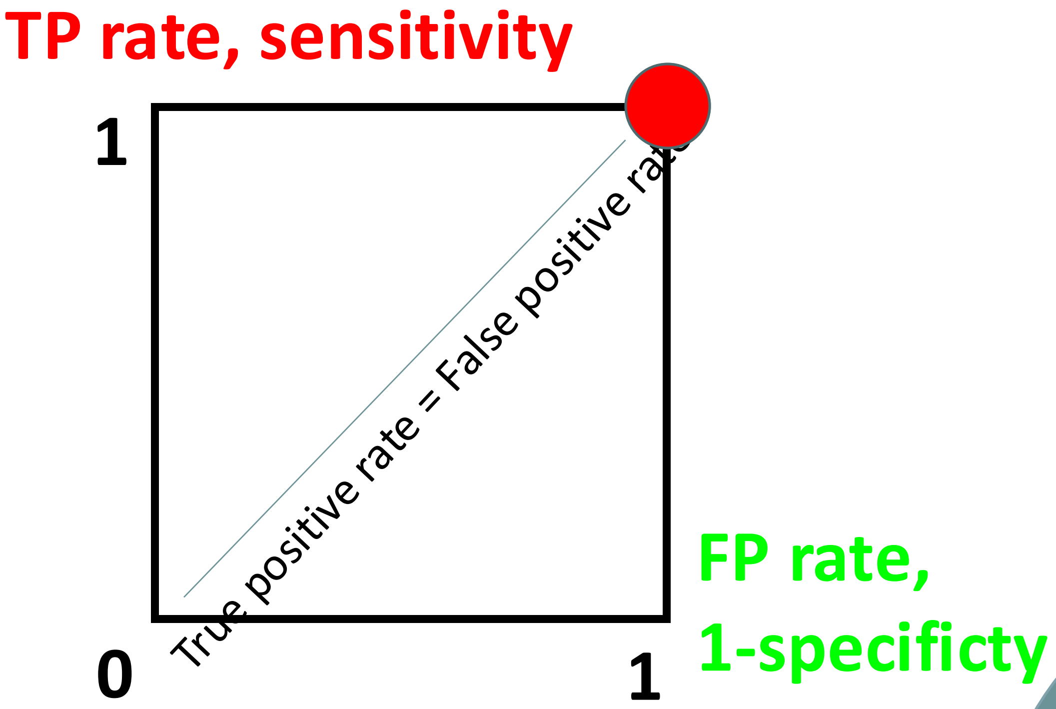
- **Accuracy (*exactitude*)**: test’s ability to correctly predict both classes
- **Precision (*précision*)**: test’s ability to correctly detect positive classes from all predicted positive classes
- **Recall (*sensitivité*):** test’s ability to correctly detect positive classes from all actual positive classes, true positive rate
- **Specificity**: true negative rate

Choice of the threshold (*seuil*)

Predicted Values

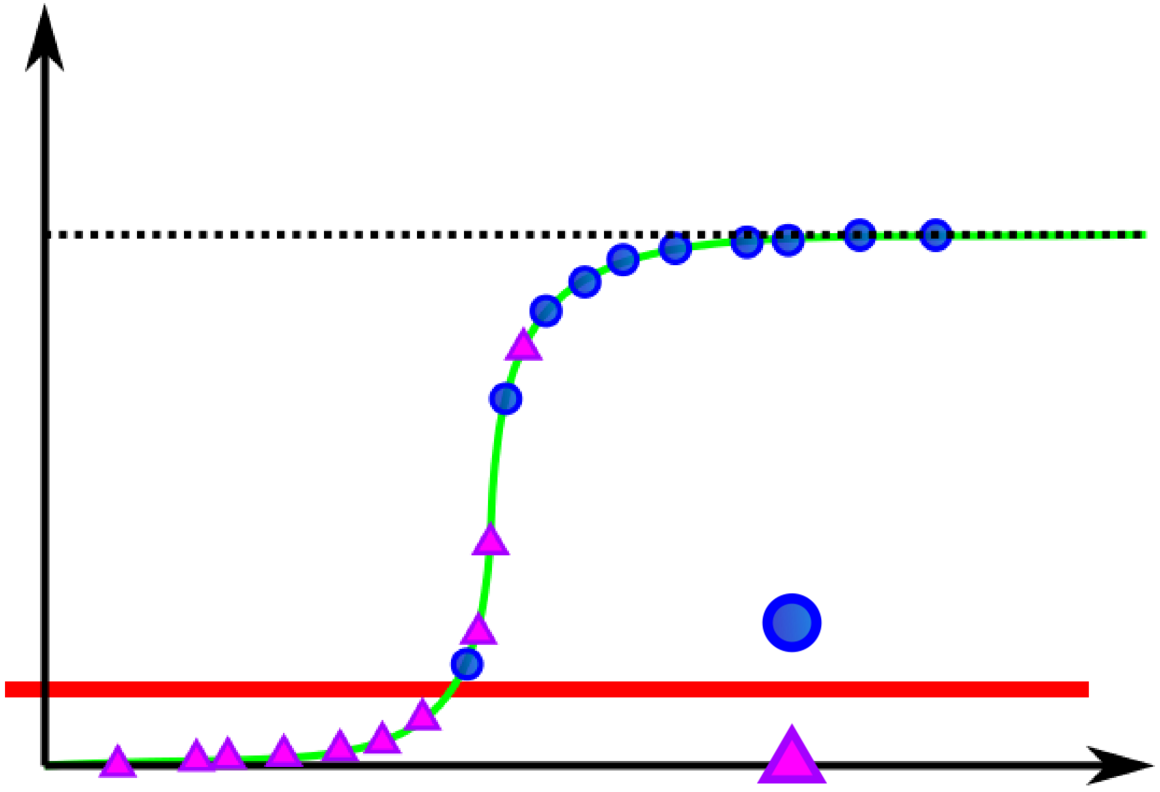
Actual Values		
	10	0
	10	0

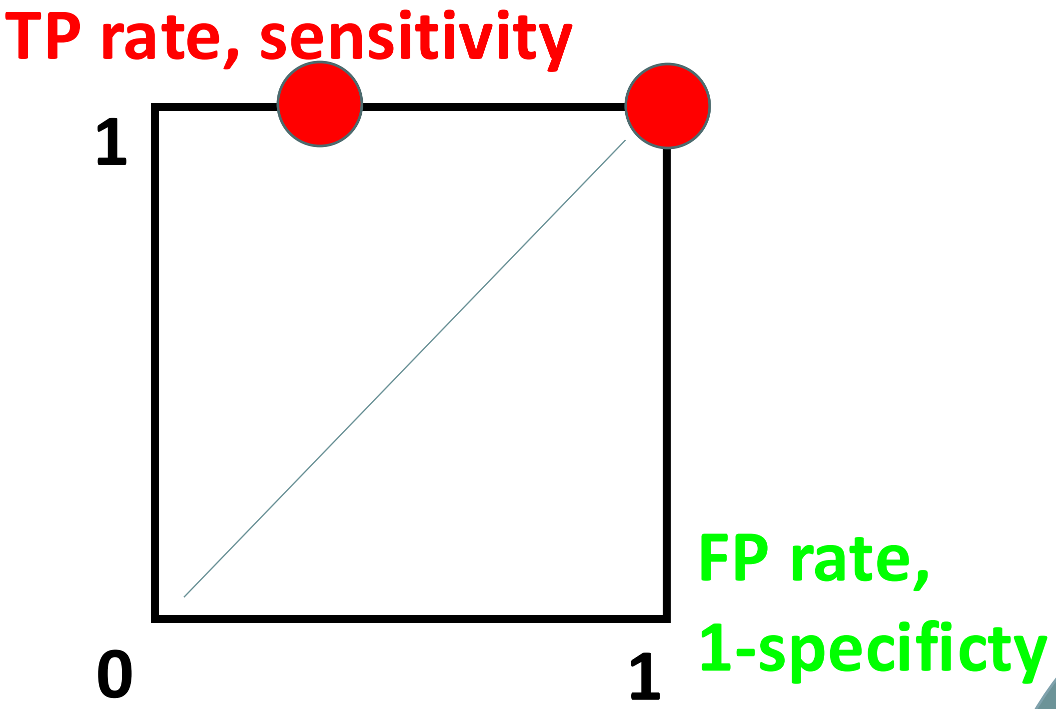




Choice of the threshold (*seuil*)

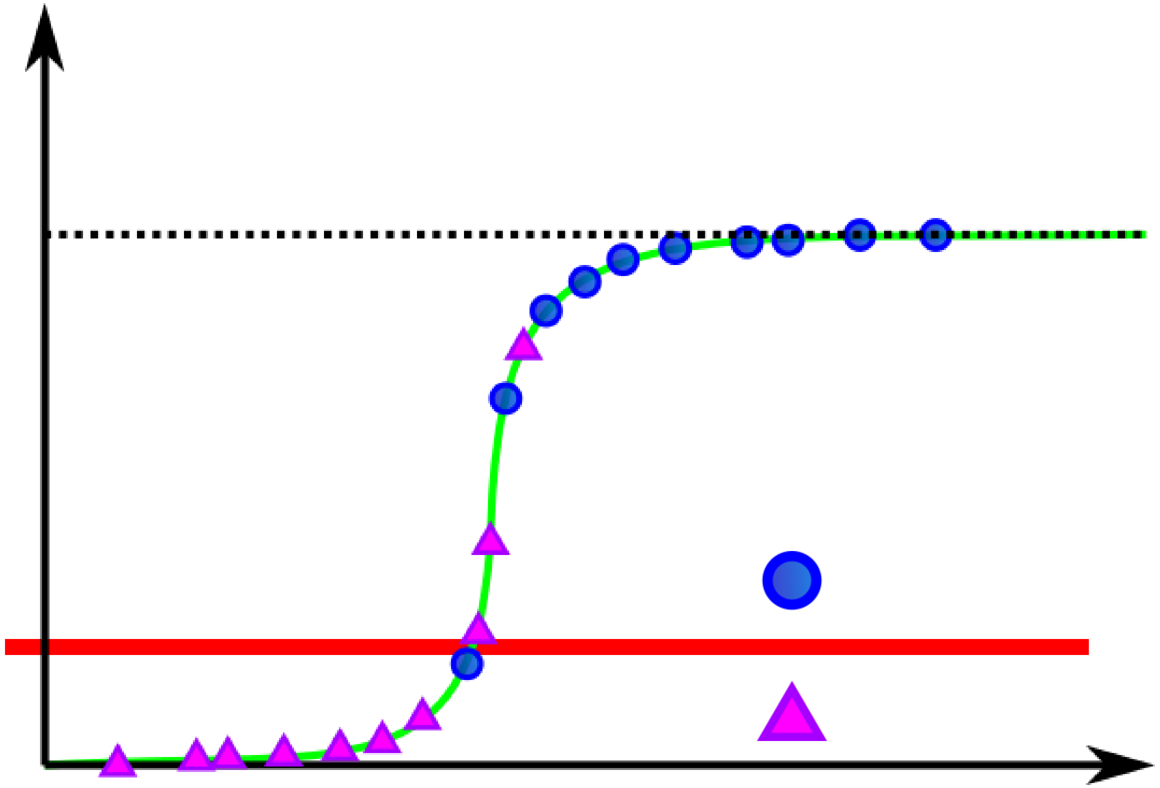
		Predicted Values ●	Predicted Values ▲
Actual Values	●	10	0
Actual Values	▲	3	7

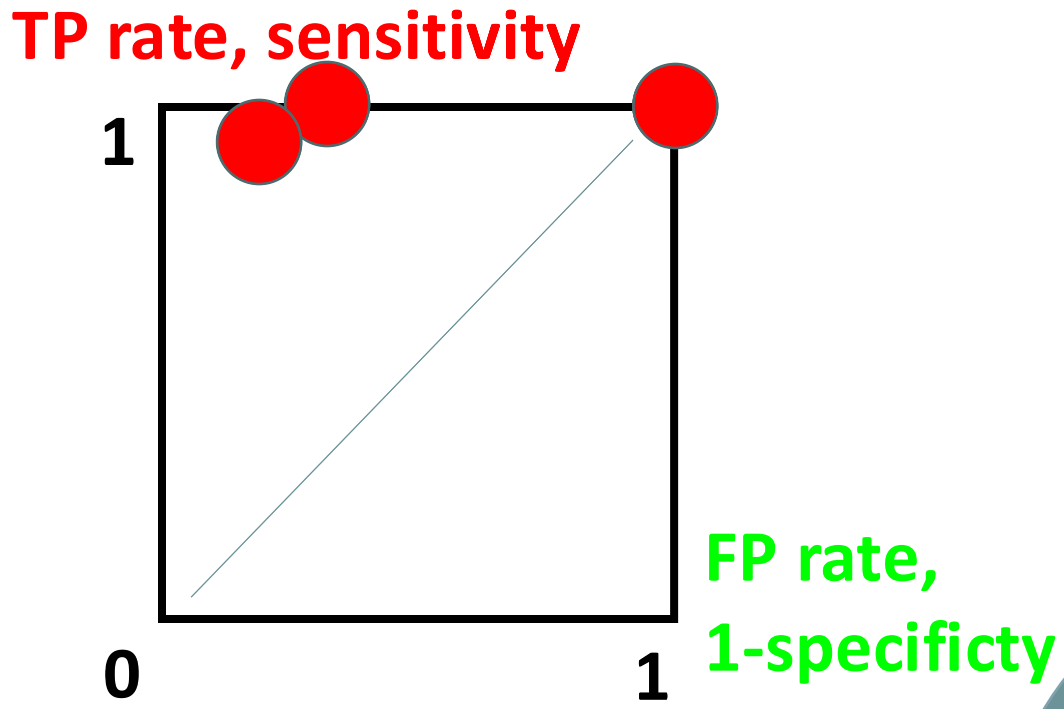




Choice of the threshold (*seuil*)

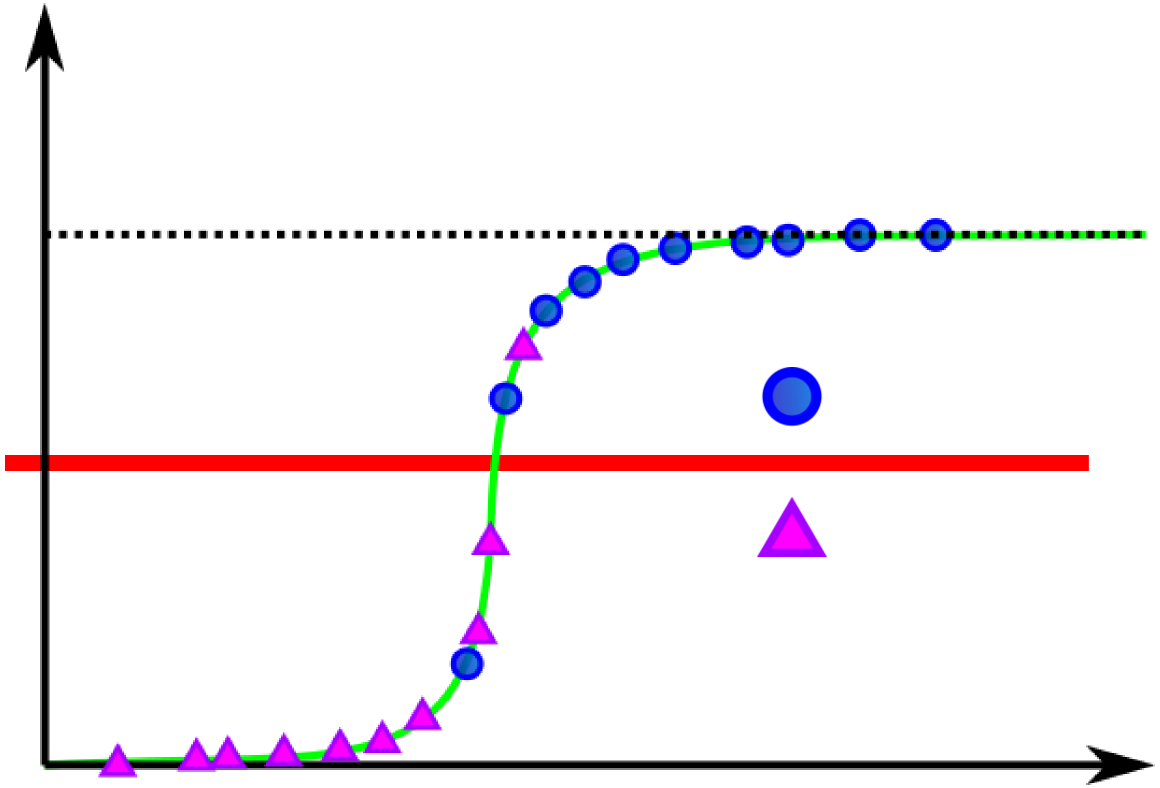
	Predicted Values	
Actual Values	9	1
	3	7

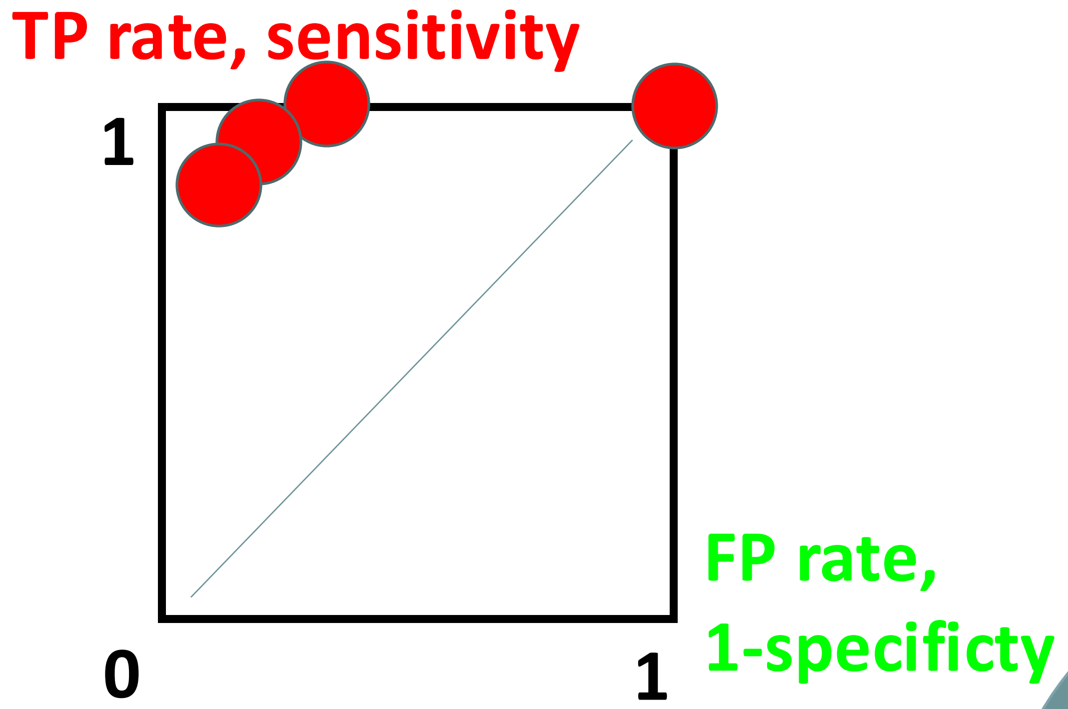




Choice of the **threshold** (*seuil*)

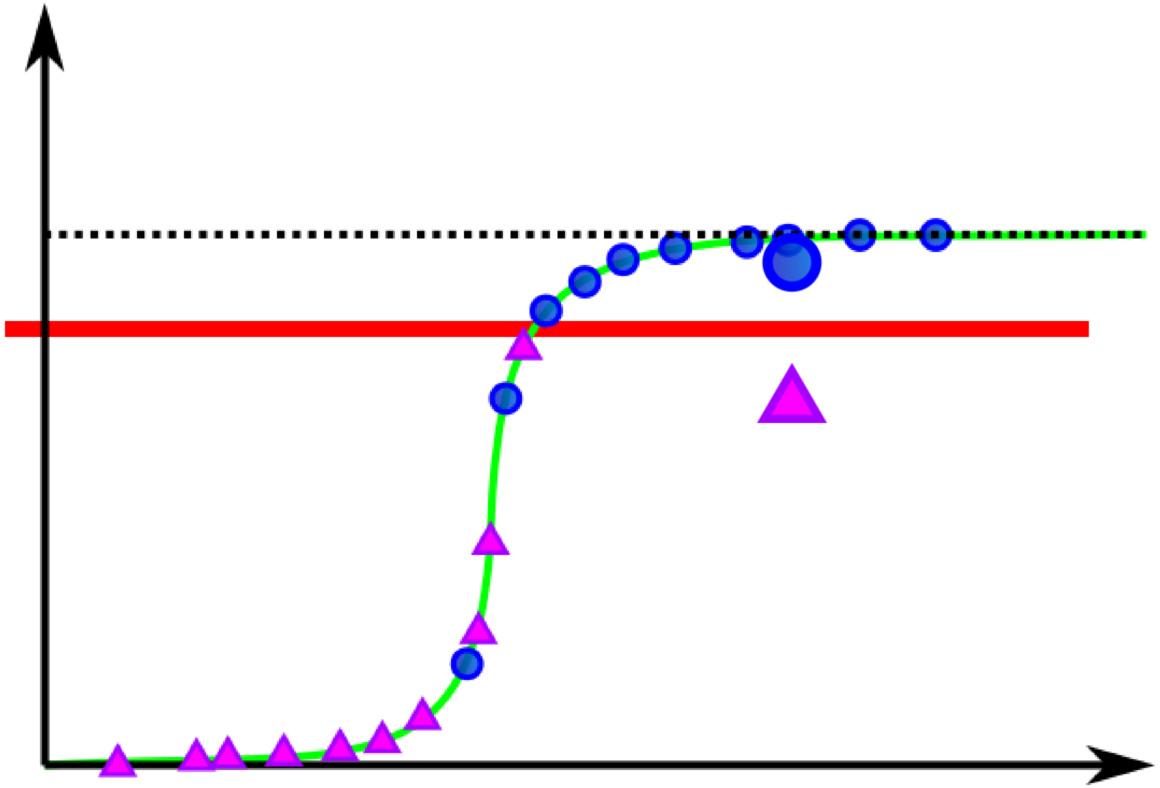
		Predicted Values ●	Predicted Values ▲
Actual Values	●	9	1
	▲	1	9

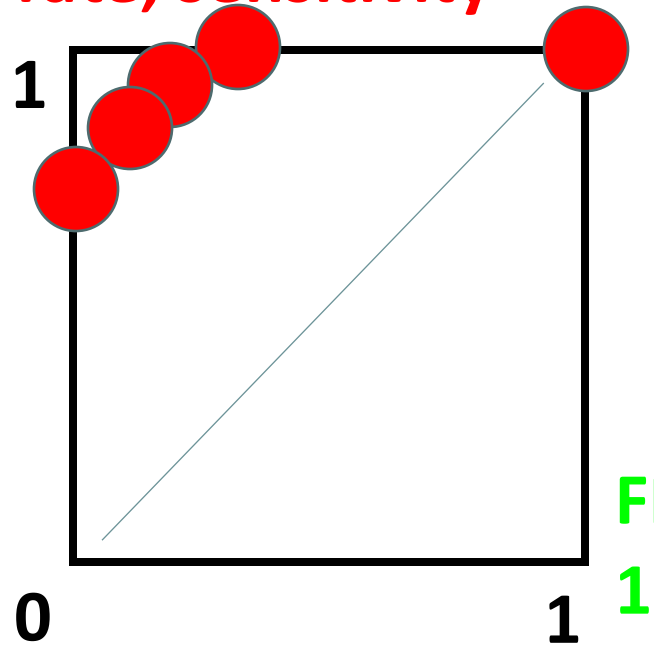




Choice of the threshold (*seuil*)

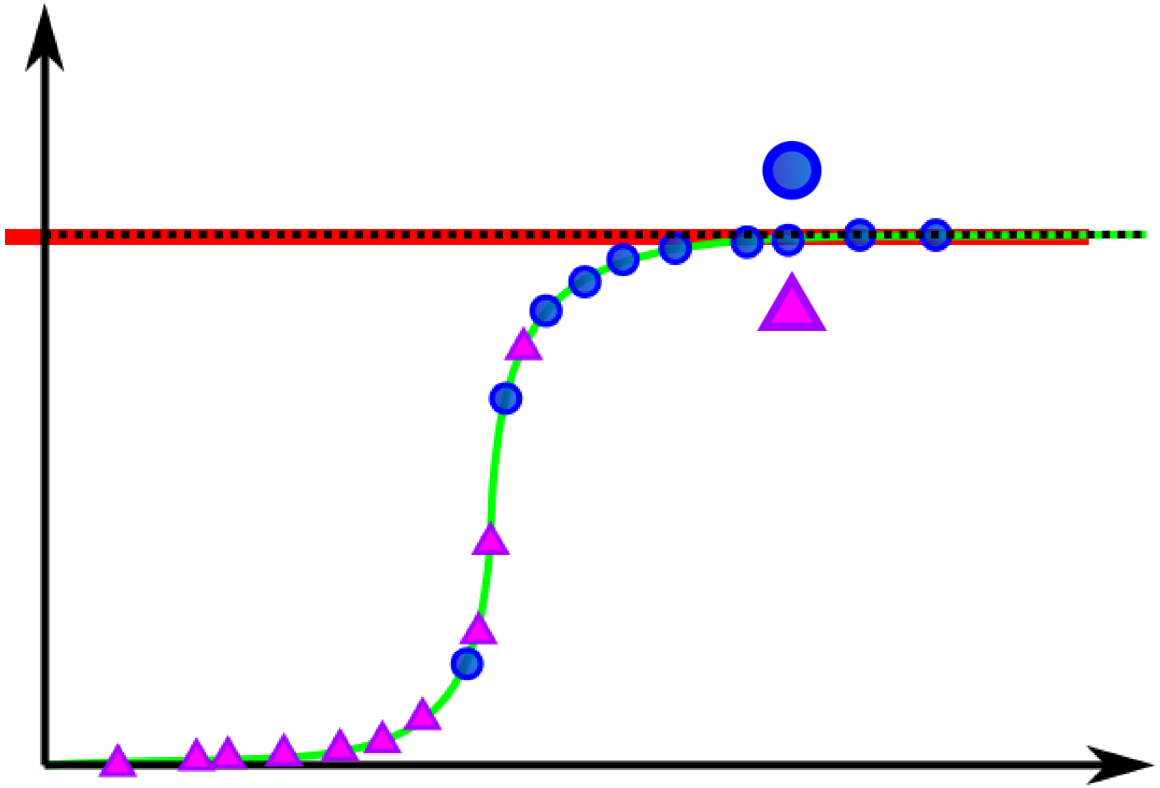
		Predicted Values: ●	Predicted Values: ▲
Actual Values	●	8	2
	▲	0	10

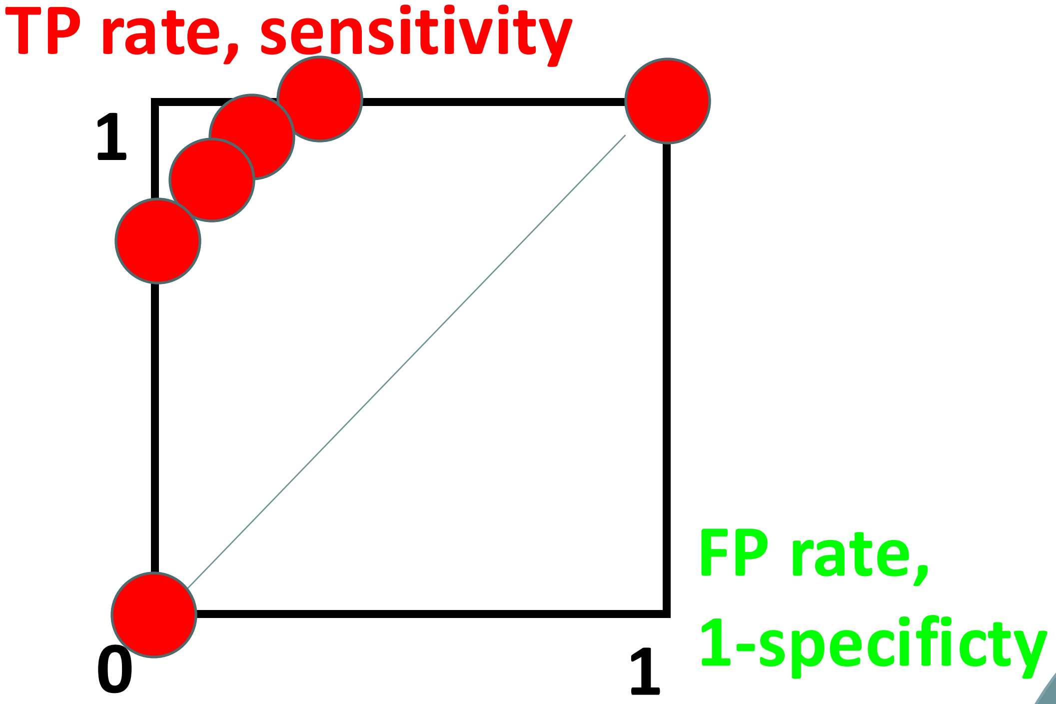




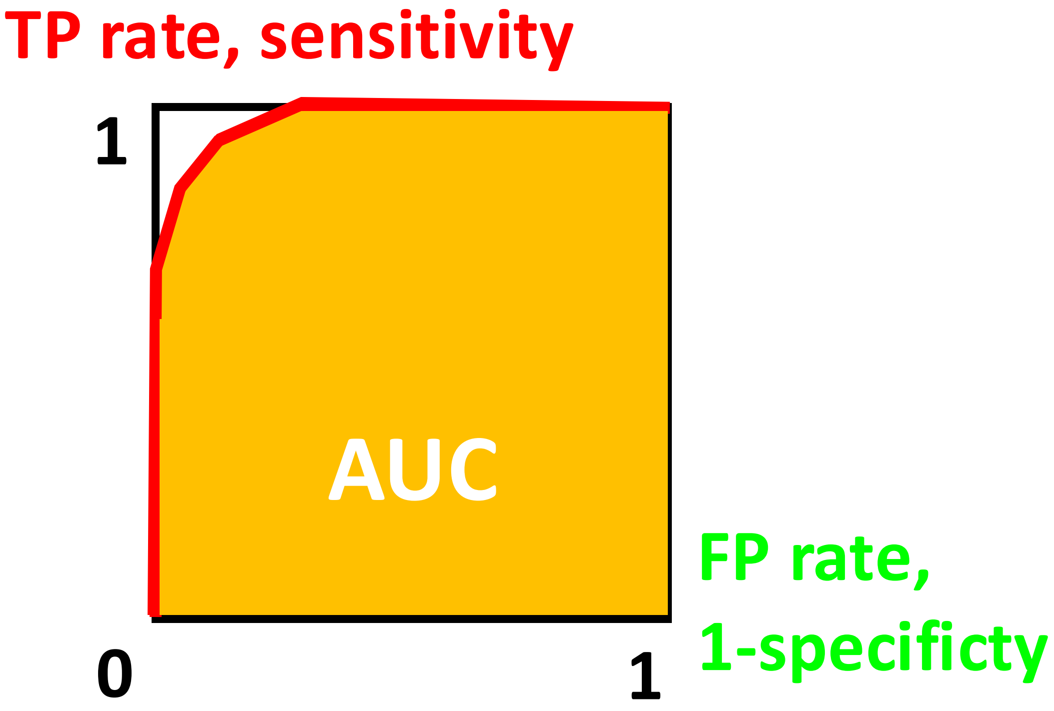
Choice of the **threshold** (*seuil*)

		Predicted Values ●	Predicted Values ▲
Actual Values	●	0	10
	▲	0	10

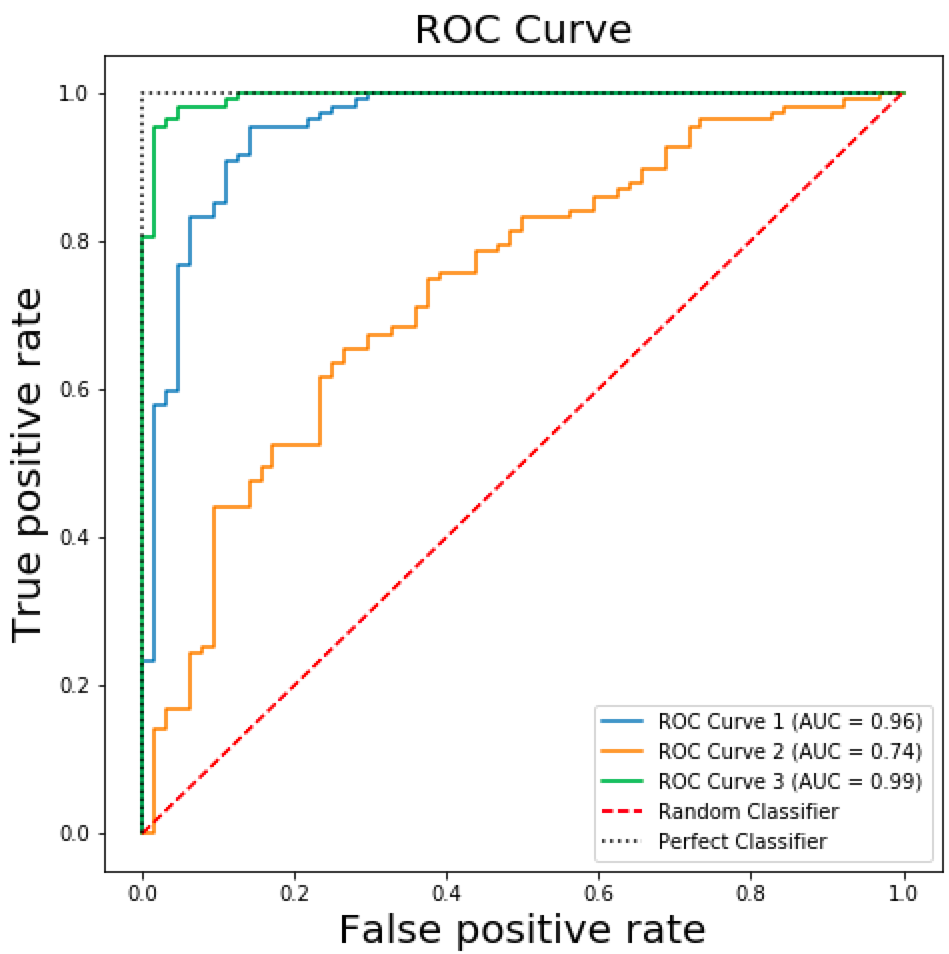




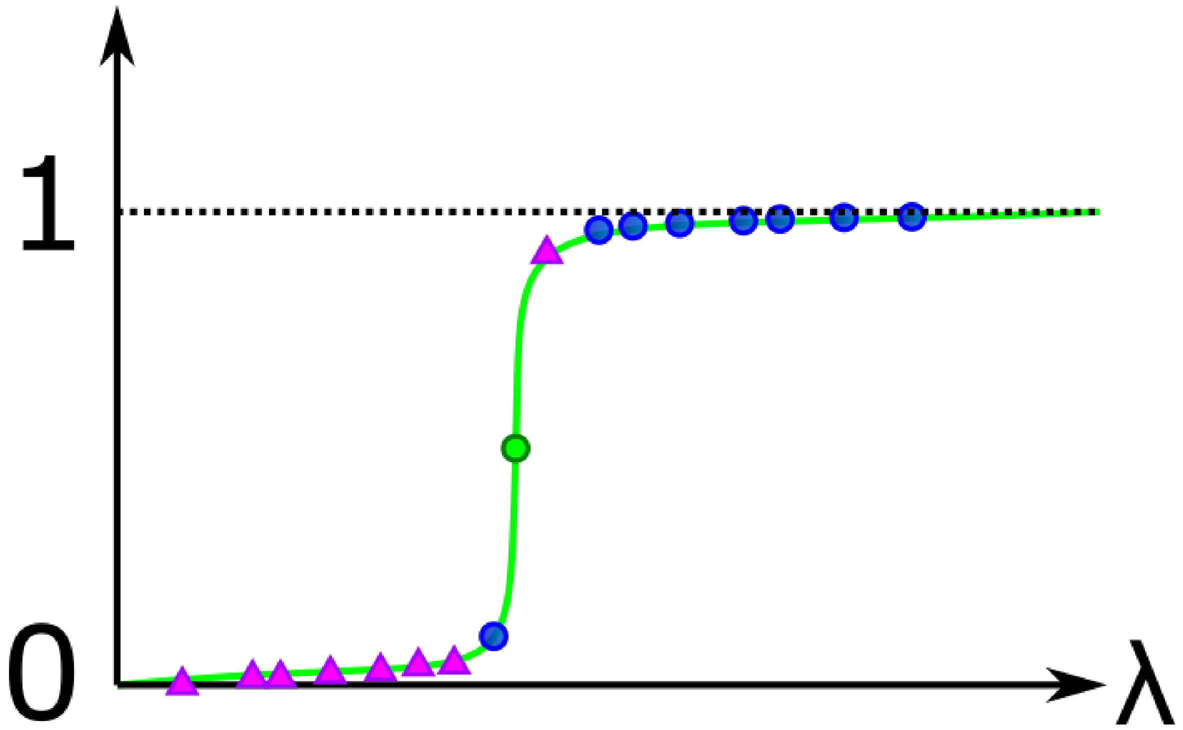
ROC graph (Receiver Operating Characteristic)



AUC score:
Area Under the ROC Curve



Logistic regression



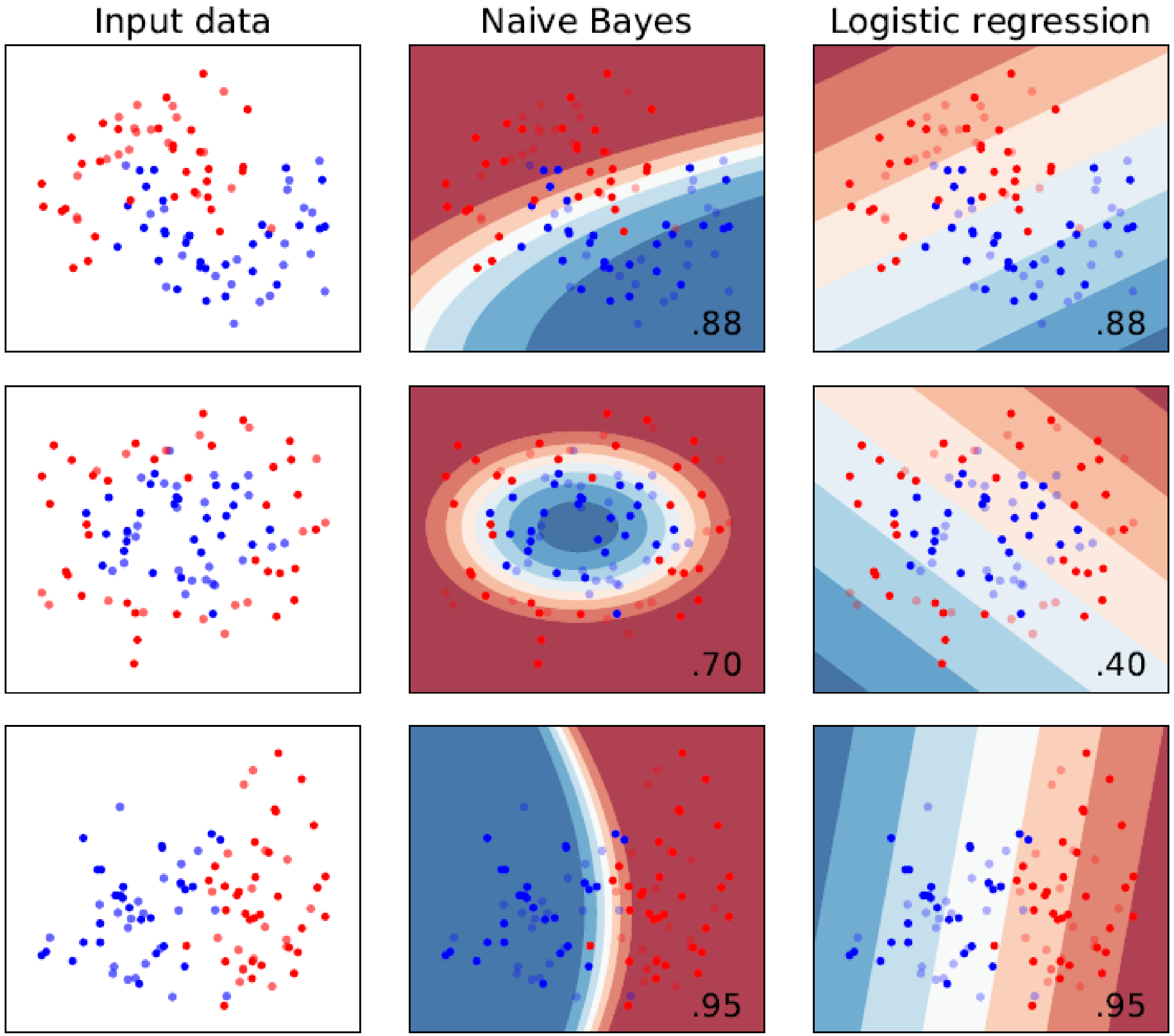
Strengths:

- Quick evaluation
- Easy interpretation of parameters

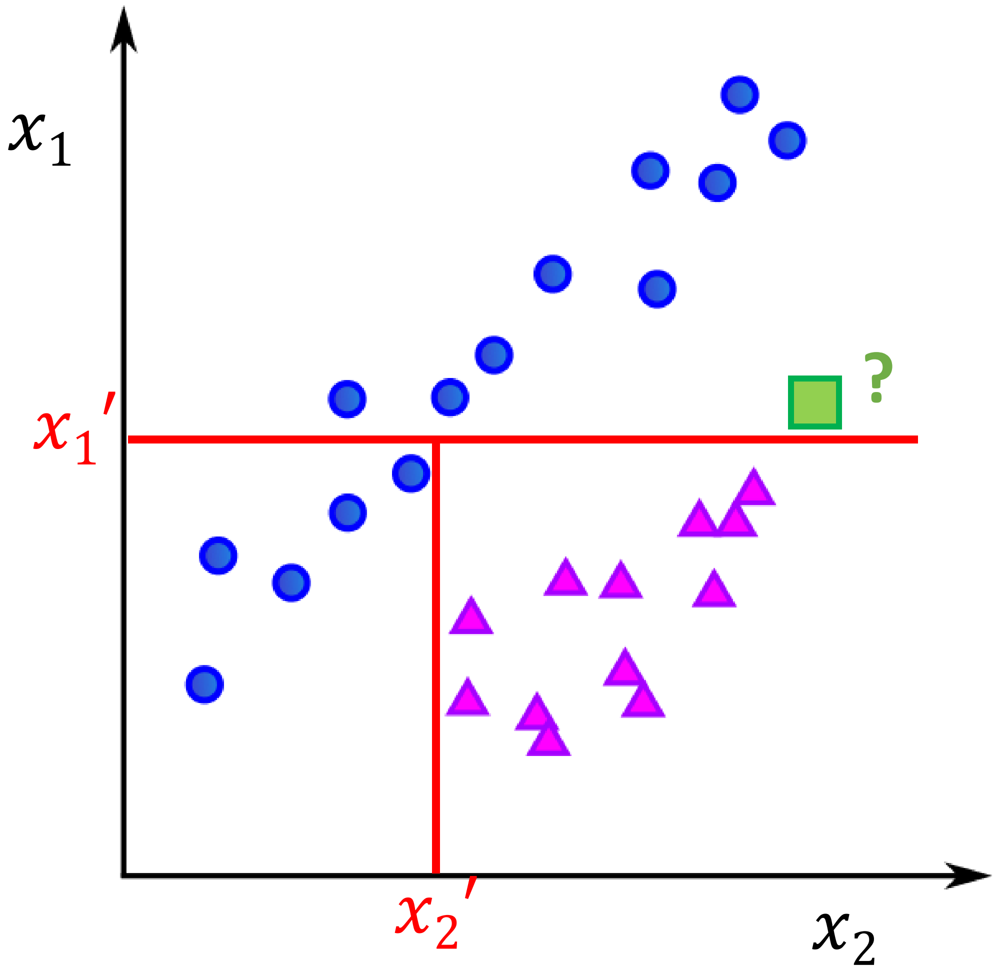
Weaknesses:

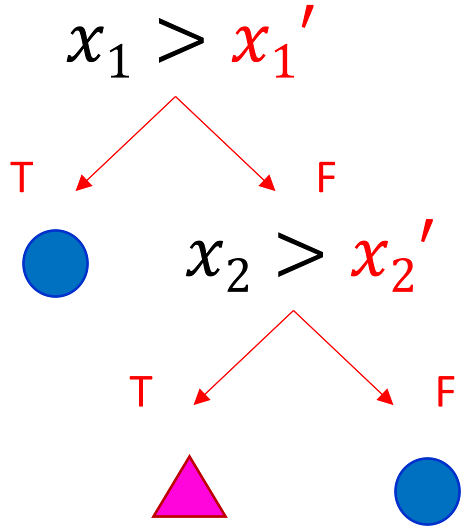
- Limited to binary classes
- Linearity hide interactions between variables
- Sensitive to outliers

Binary classifier comparison



Decision Tree

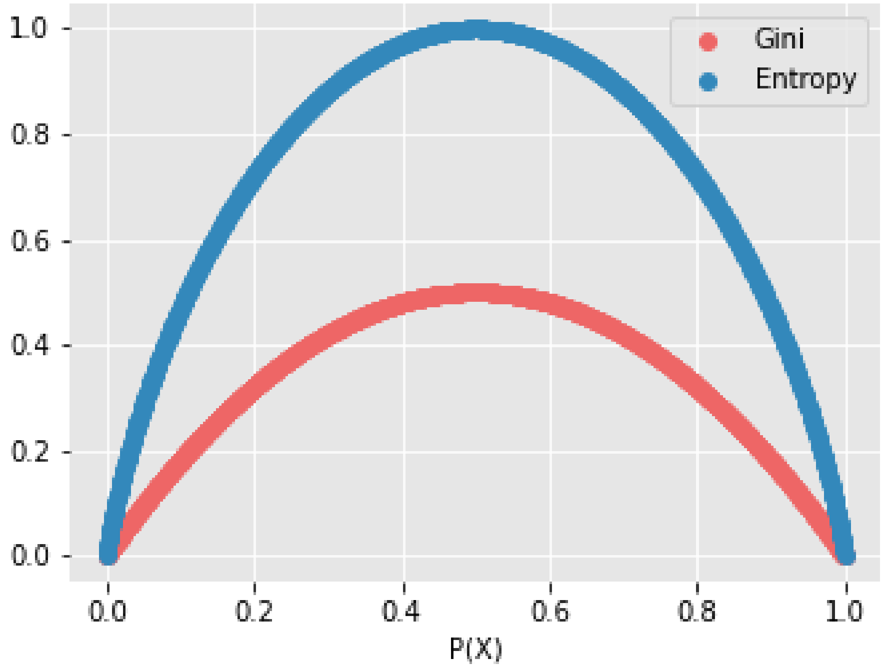




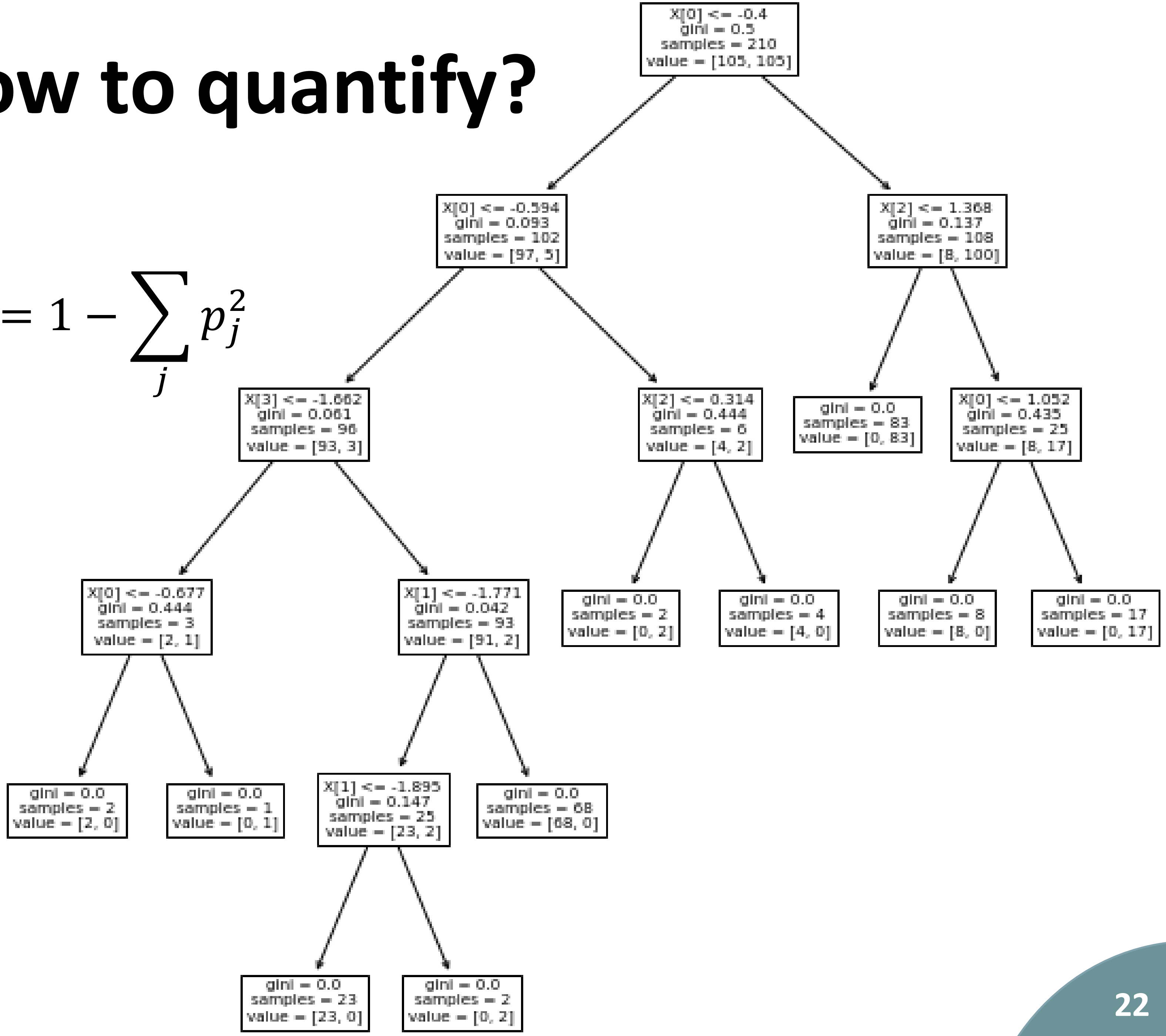
Decision Tree: How to quantify?

Classification in *j* class:

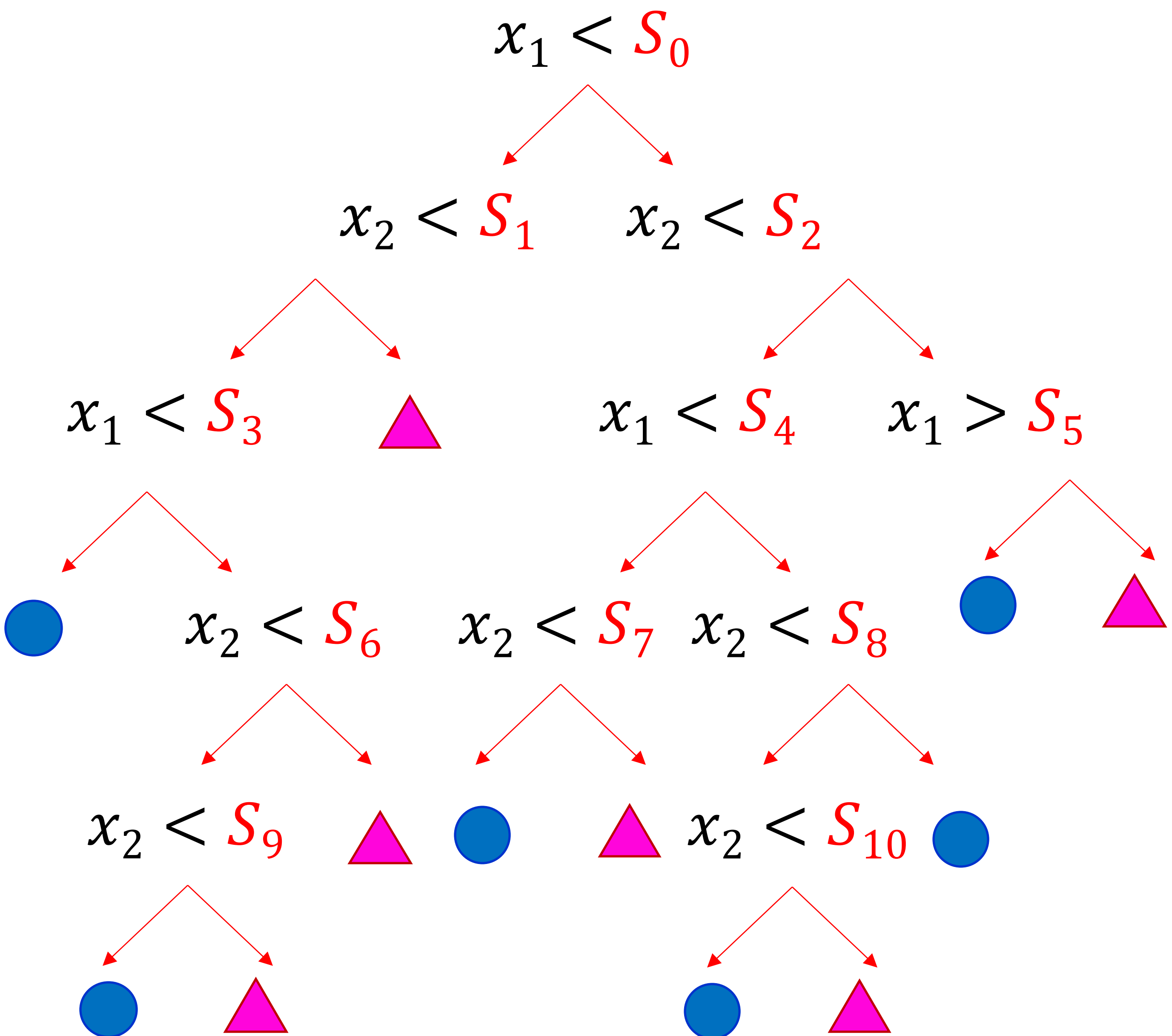
$$entropy = -\sum_j p_j \log_2 p_j \quad gini = 1 - \sum_j p_j^2$$



Regresssion: variance



Decision Tree



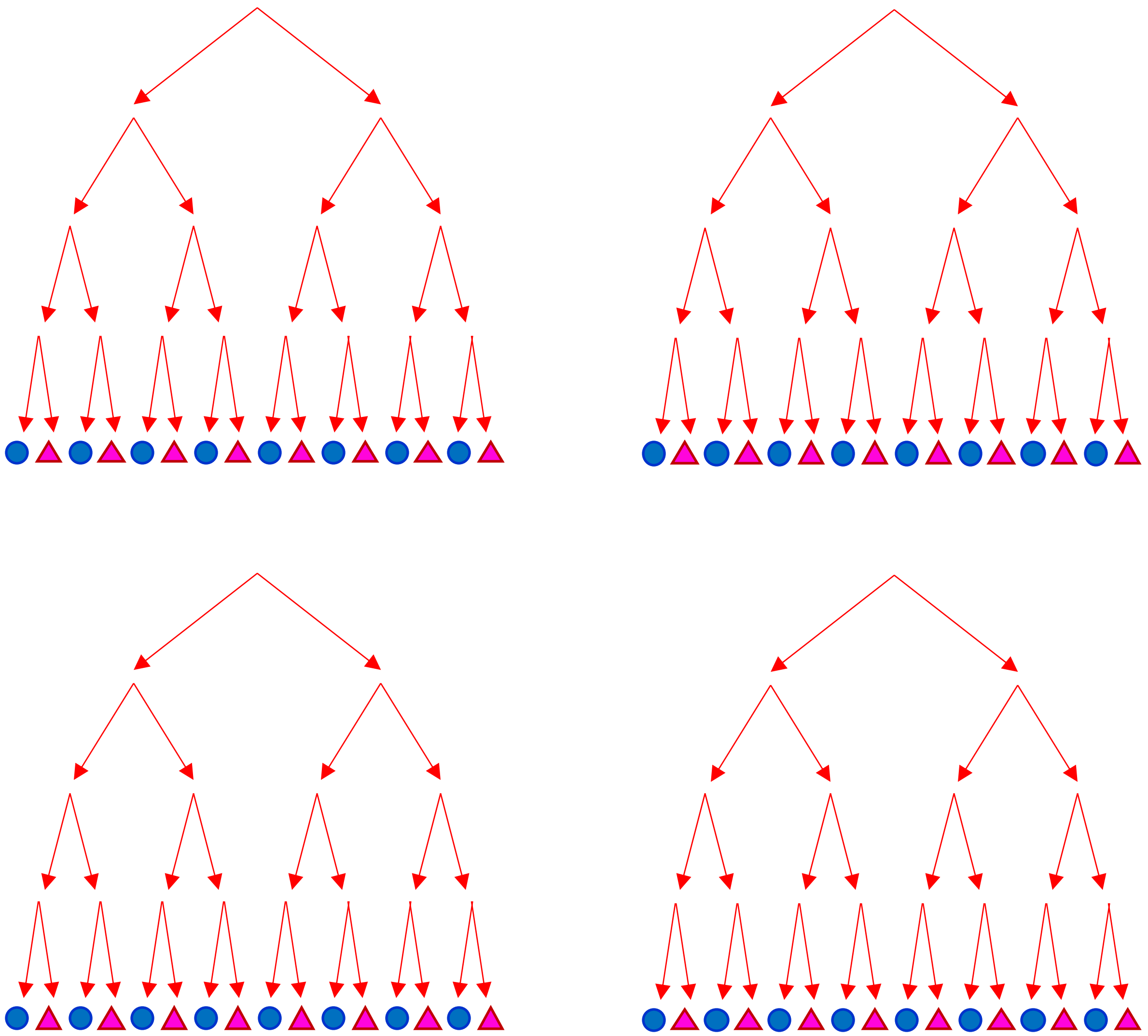
Strengths:

- visibility
- multiclasses

Weaknesses:

- No correlation
- Greedy approach, no regret
- Overfitting

Random forest



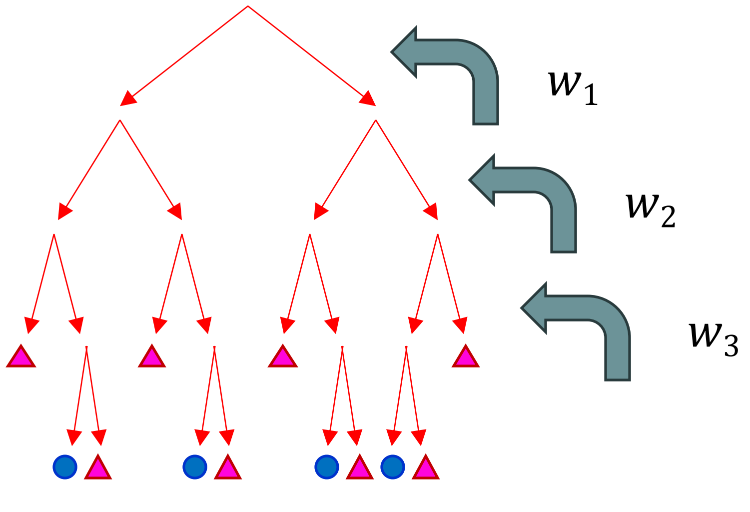
stochastic discrimination based on averaging multiple decision trees, trained on different parts

Strengths:
- Very accurate
- Overcomes the DT overfitt
- Efficient even if missing data

Weaknesses:
- Losing of the easy reading of DT (black box)
- Several parameters

sklearn.ensemble**.RandomForestClassifier**

Gradient boosting



Hyperparameters : $w_i = \dfrac{N_{\triangle}}{N_{\bullet} \times \mu}$

Strengths:

- Very efficient
- flexible

Weaknesses:

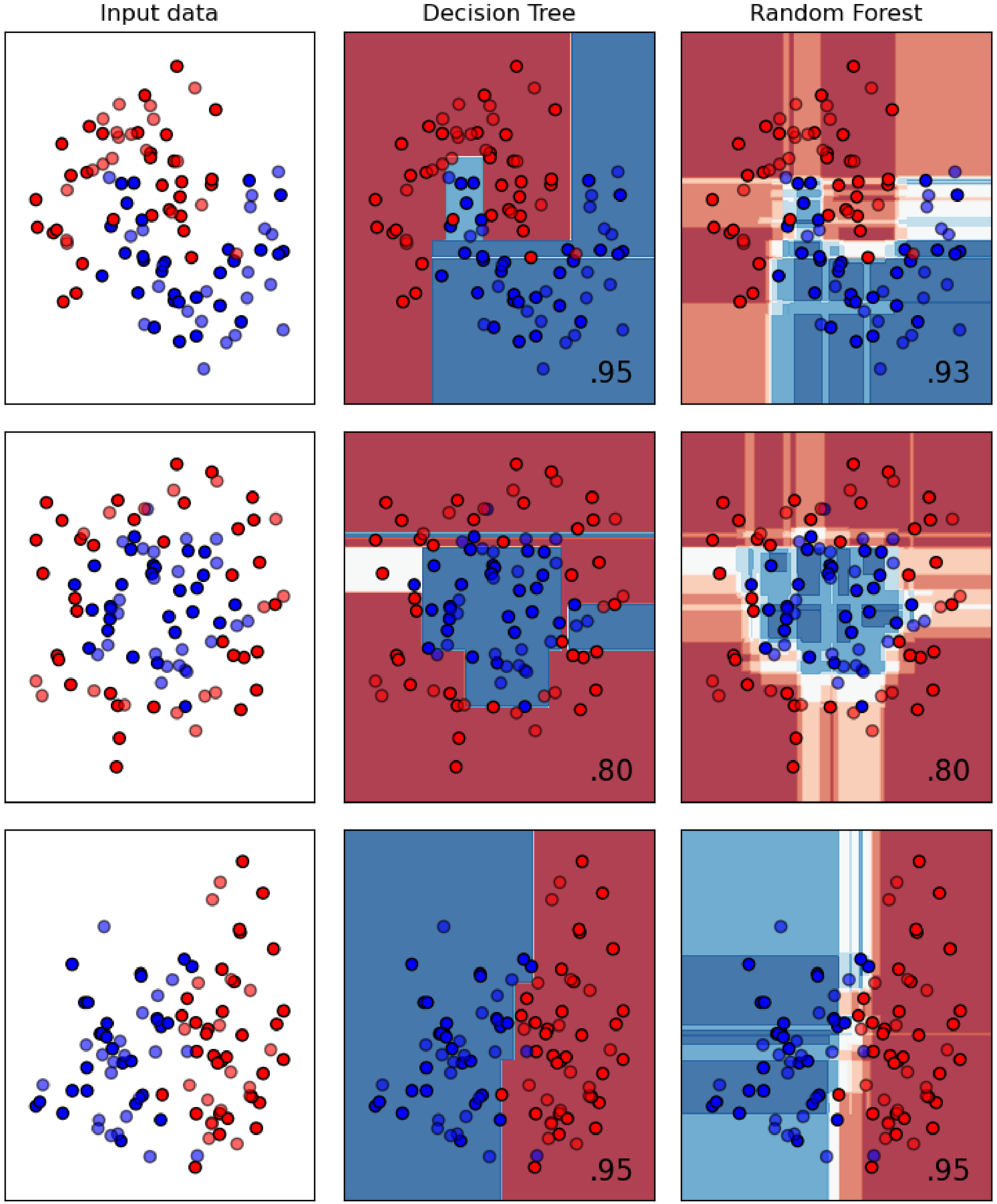
- Slow to train
- Sensitive to hyperparameters

Predicted Values (left: before; right: after)

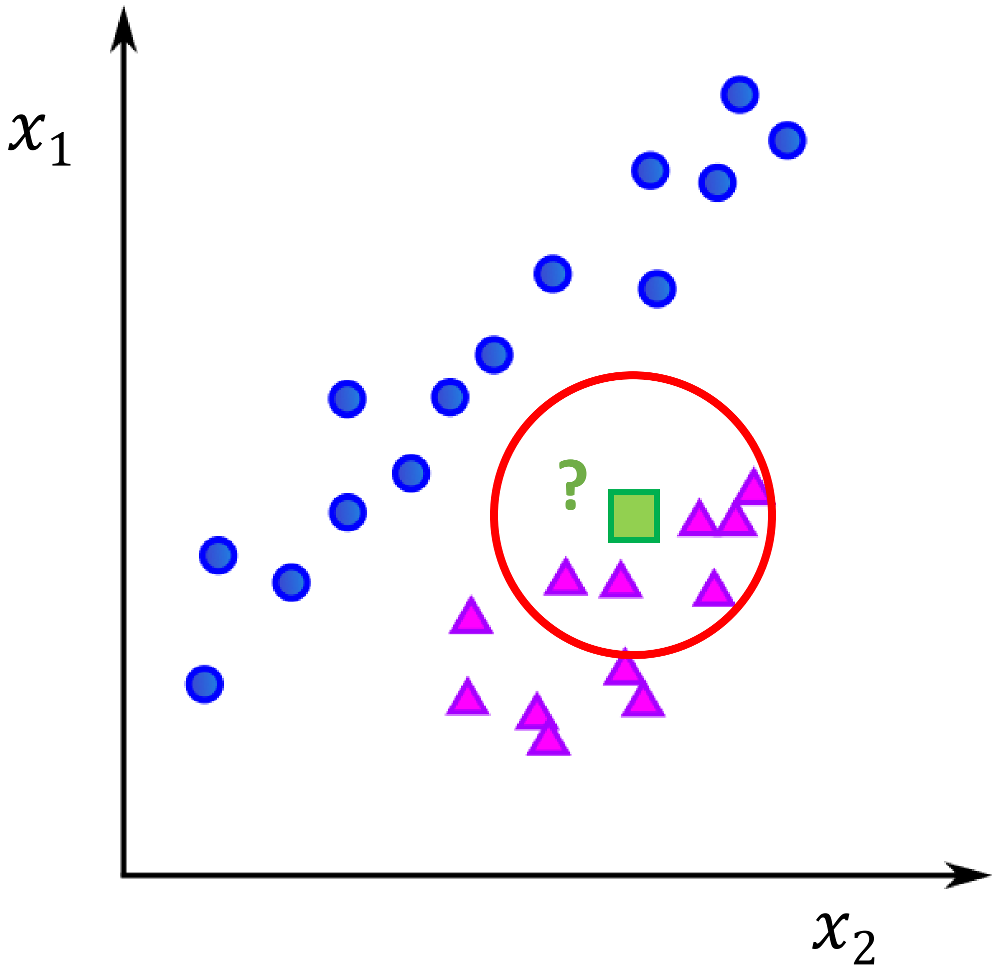
Actual Values	●	▲
●	2	2
▲	2	6

Actual Values	●	▲
●	4	4
▲	0	4

Tree VS Forest



K-nearest neighhors



Find portion of N in k cluster for which sum of squared distance is minimum

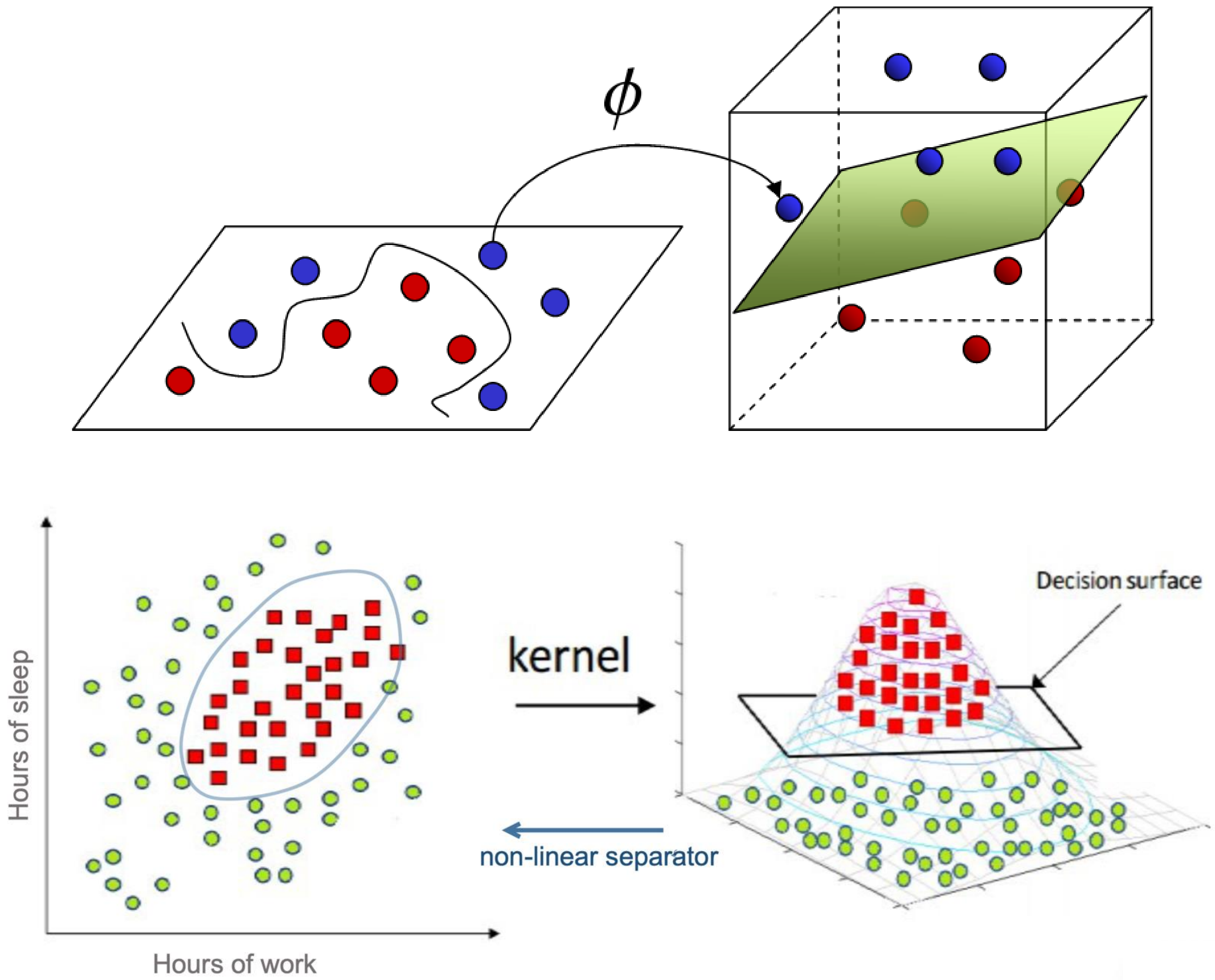
Strengths:

- Efficient, flexible
- Easy to understand

Weaknesses:

- Sensible to noise
- Too local
- Slow if many data
- Sensitive to outliers

Support Machine Vector (SVM)



Separation by a linear plan : find the hyper space (higher degree) easy to separate data

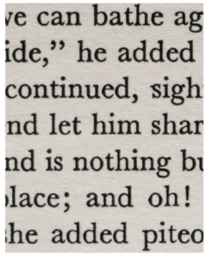
Strengths:

- Large dimension data
- Faster than NN

Weaknesses:

- Complexity in N^3
- Less efficient than RF
- Interpretability

Classification is a regression ?


ve can bathe ag
ide," he added
continued, sigh
nd let him shar
nd is nothing bu
place; and oh!
he added piteo

Image

we can bathe ag
ide," he added
continued, sigh
nd let him shar
nd is nothing bu
place; and oh!
she added piteo

Text

$$\begin{pmatrix} 8.167 \\ 1.492 \\ 2.782 \\ 4.253 \\ \ldots \\ 2.360 \\ 0.150 \\ 1.974 \end{pmatrix}$$

Letter occurrence

$$\begin{pmatrix} 0\ 0\ 0 \ldots 1\ 0\ 0\ 0\ 0 \\ 0\ 0\ 0 \ldots 0\ 0\ 0\ 0\ 1 \\ 0\ 0\ 0 \ldots 0\ 0\ 0\ 0\ 0 \\ 0\ 0\ 0 \ldots 0\ 0\ 0\ 0\ 0 \\ \ldots \\ 0\ 0\ 0 \ldots 0\ 1\ 0\ 0\ 0 \\ 0\ 1\ 0 \ldots 0\ 0\ 0\ 0\ 0 \\ 0\ 0\ 0 \ldots 0\ 0\ 0\ 1\ 0 \end{pmatrix}$$

One hot encoding

$$\begin{pmatrix} 0{,}818 & 0{,}069 & \ldots & 0{,}576 & 0{,}773 \\ 0{,}415 & 0{,}952 & \ldots & 0{,}539 & 0{,}835 \\ 0{,}845 & 0{,}572 & \ldots & 0{,}002 & 0{,}775 \\ 0{,}562 & 0{,}805 & \ldots & 0{,}651 & 0{,}823 \\ \ldots & & & & \\ 0{,}332 & 0{,}348 & \ldots & 0{,}608 & 0{,}808 \\ 0{,}908 & 0{,}968 & \ldots & 0{,}701 & 0{,}689 \\ 0{,}424 & 0{,}180 & \ldots & 0{,}587 & 0{,}129 \end{pmatrix}$$

Embedding

Scalar
1.34 kg

Characteristics

Vector
H_2O : 53.6
CL^- : 0.546
Na^+ : 0.469

Composition

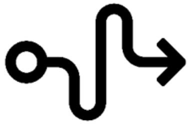
Series
Sampling

Audio

Matrix/Tensor
[pixels]

Images

Series
[positions]



Trajectory

Series
[tokens]

Text