

# 2

## L'algorithme K-NN

- Fonctionnement
- Critères de qualité



FACULTÉ DE  
GESTION,  
ÉCONOMIE  
& SCIENCES



# Algorithme kNN : sélection de la classe

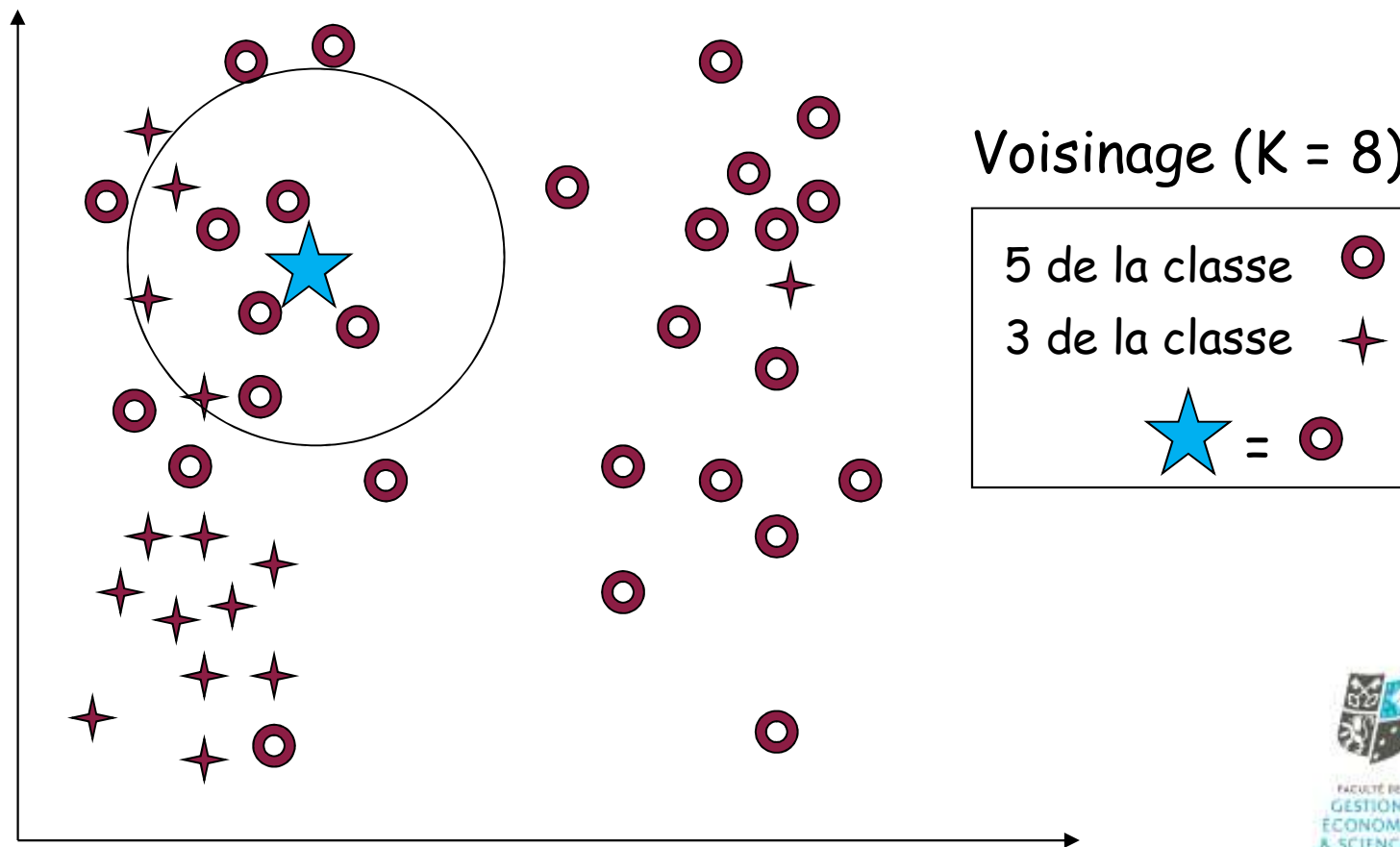
**Solution simple** : rechercher le cas le plus proche et prendre la même décision (Méthode 1-NN).

## Combinaison des k classes :







- Heuristique :  $k = \text{nombre d'attributs} + 1$
- Vote majoritaire : prendre la classe majoritaire.
- Vote majoritaire pondéré : chaque classe est pondérée. Le poids de  $c(x_i)$  est inversement proportionnel à la distance  $d(y, x_i)$ .

**Confiance** : Définir une confiance dans la classe attribuée = rapport entre les votes gagnants et le total des votes.

# Illustration K-NN









## Retour sur KNN : Exemple (1)

Customer	Age	Income	No. credit cards	Loyal
John 	35	35K	3	No
Rachel 	22	50K	2	Yes
Hannah 	63	200K	1	No
Tom 	59	170K	1	No
Nellie 	25	40K	4	Yes
David 	37	50K	2	?

## Retour sur KNN : Exemple (2)

$K = 3$

Customer	Age	Income	No. credit cards	Loyal
John 	35	35K	3	No
Rachel 	22	50K	2	Yes
Hannah 	63	200K	1	No
Tom 	59	170K	1	No
Nellie 	25	40K	4	Yes
David 	37	50K	2	Yes

Distance from David
$\text{sqrt} [(35-37)^2 + (35-50)^2 + (3-2)^2] = 15.16$
$\text{sqrt} [(22-37)^2 + (50-50)^2 + (2-2)^2] = 15$
$\text{sqrt} [(63-37)^2 + (200-50)^2 + (1-2)^2] = 152.23$
$\text{sqrt} [(59-37)^2 + (170-50)^2 + (1-2)^2] = 122$
$\text{sqrt} [(25-37)^2 + (40-50)^2 + (4-2)^2] = 15.74$

# Comment évaluer un algorithme de classification supervisée ?

On part d'un groupe d'individus (cohorte si recherche médicale) dont le statut est connu

Individu	Ivre ?
1	O
2	N
3	N
4	O
5	O
6	N

On passe l'algorithme sur chacun des individus



# Evaluation : exemple de mesure : l'exactitude

Classification accuracy  
(CA) = % de bien classés

individu	Ivre?	T1	T2
1	O		
2	N		
3	N		
4	O		
5	O		
6	N		

- **Test 1**
  - CA : 5/6
- **Test 2**
  - CA : 5/6



# Evaluation : exemple de mesure : l'exactitude

Classification accuracy  
(CA) = % de bien classés

individu	Ivre?	T1	T2
1	O		
2	N		
3	N		
4	O		
5	O		
6	N		

- **Test 1**
  - CA : 5/6
- **Test 2**
  - CA : 5/6





# Evaluation : matrice de confusion

Plus de 50 mesures différentes  
Basées sur le comptage des :

- Vrais positifs (VP)
  - Ivre & test positif
- Vrais négatifs (VN)
  - Sobre & test négatif
- Faux positifs (FP)
  - Sobre & test positif
- Faux négatifs (FN)
  - Ivre & test négatif

Ex: **Classification Accuracy (CA)** = 
$$\frac{VP + VN}{VP + VN + FP + FN}$$

Représentation :

- C : l'individu a la condition
- P : le test est positif

	P	$\bar{P}$
C	VP	FP
$\bar{C}$	FN	VN

# Principales mesures en classification supervisée

0 ←————→ 1  
(performances minimum) (performances maximum)

- **CA** (Classification Accuracy)

- % d'individus correctement classés

$$\frac{VP + FP}{VP + FP + VN + FN}$$

- **Confiance** (*precision*)

- Capacité à ne pas se tromper lorsque l'on trouve la cible
- % d'individus détectés qui ont effectivement la cible

$$\frac{VP}{VP + FP}$$

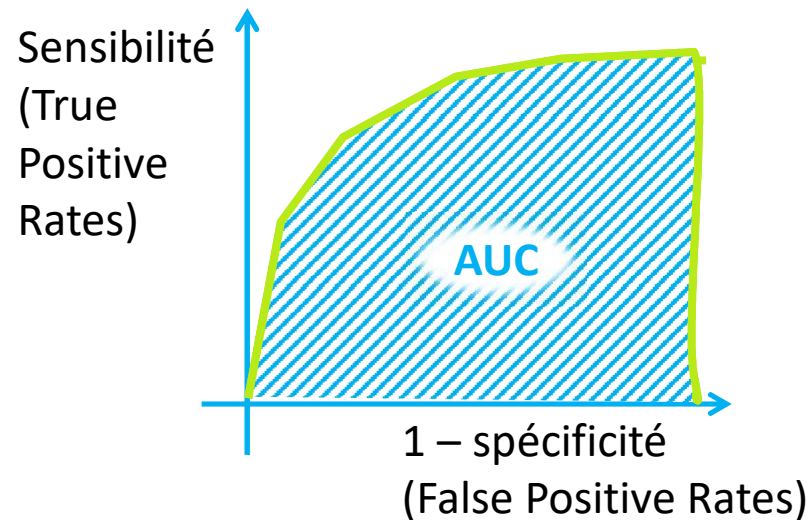
- **Sensibilité** (*recall* / rappel)

- Capacité à détecter la cible
- % d'individus avec la condition identifiés par le test

$$\frac{VP}{VP + FN}$$

# Principales mesures en classification supervisée

- F1-mesure
  - Moyenne harmonique de la Confiance et de la Sensibilité
$$2 \times \frac{\text{confiance} \times \text{sensibilité}}{\text{confiance} + \text{sensibilité}}$$
- AUC (Area Under ROC curve)



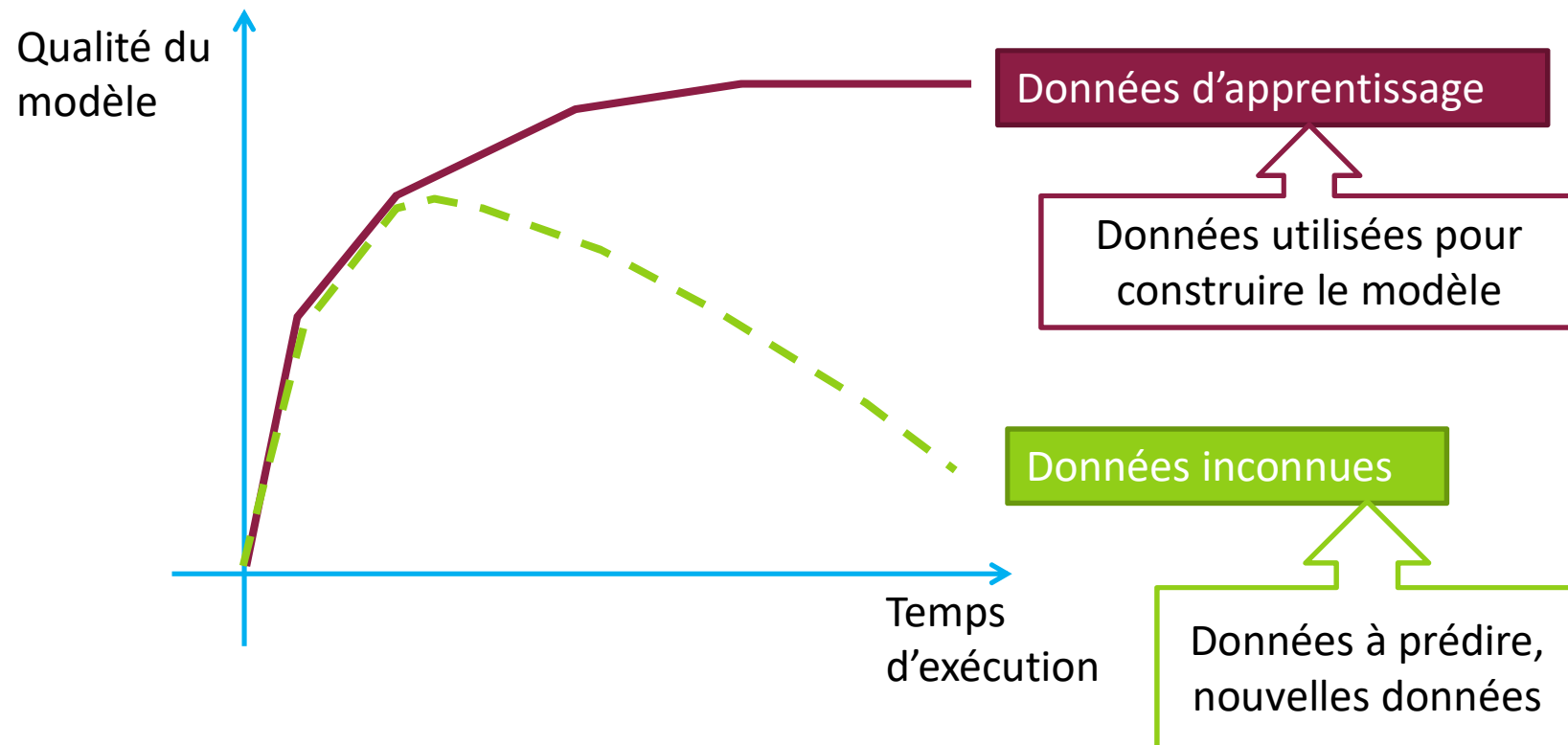
# C'est à vous : exemple d'évaluation

T1 et T2 détectent si la personne est ivre

- T1 ou T2 pour faire du préventif ? (ex: Ethylotest avant de prendre le volant) ?
- T1 ou T2 pour faire du curatif ? (ex: Emprisonnement si ivresse détectée au volant)
- T1 ou T2 pour un usage « polyvalent » ?

id	Ivre?	T1	T2	T1								T2							
				VP	FP	FN	VN	CA	Se	Cf	F1	VP	FP	FN	VN	CA	Se	Cf	F1
1	O	O	O	1	0	0	0					1	0	0	0				
2	N	O	N	0	1	0	0					0	0	0	1				
3	N	N	N	0	0	0	1					0	0	0	1				
4	O	O	O	1	0	0	0					1	0	0	0				
5	O	O	N	1	0	0	0					0	0	1	0				
6	N	N	N	0	0	0	1					0	0	0	1				
Scores				3	1	0	2	0,83	1	0,75	0,86	2	0	1	3	0,83	0,67	10,74	

# Le sur-apprentissage



Un algorithme d'apprentissage qui tourne trop longtemps « apprend par cœur » les données

-> performances mauvaises sur nouvelles données

# Le sur-apprentissage

Il est important d'évaluer les modèles obtenus sur des données différentes

	F	M P	C	H	Flu?	R1	R2
P1	Y	Y	Y		Y	Y	Y
P2	Y		Y		N	Y	N
P3				Y	N	N	N
P4	Y	Y	Y	Y	Y	Y	Y
P5	Y	Y		Y	Y	Y	N
P6		Y			N	N	N
N1	Y	Y			Y	Y	N
N2	Y		Y	Y	Y	Y	N
N3			Y		N	N	N

Données  
apprentissage

Données  
test

R1: Fever → Flu

CA apprentissage: 5/6

CA test: 3/3

R2: F & MP & C → Flu

CA apprentissage : 5/6

CA test: **0/3**