

# 5 - Bayésien naïf et K plus proches voisins

Jean Massardi - Été 2022

# Plan

1 - Réseaux Bayésiens

2 - K plus proches voisins

# Plan

## **1 - Réseaux Bayésiens**

## 2 - K plus proches voisins

# Probabilité

Soit X et Y deux variables aléatoire

$P(X, Y)$       Probabilité jointe

$P(X|Y)$       Probabilité conditionnelle

$$P(X, Y) = P(X|Y) * P(Y) = P(Y|X) * P(X)$$


# Théorème de Bayes

*same*

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

# Théorème de Bayes

Dans le cas qui nous intéresse ici on a une classe  $Y$  et  $n$  nombreux attributs  $X_i$

$$P(Y|X_1, X_2, \dots) = \frac{P(X_1, X_2, \dots | Y)P(Y)}{P(X_1, X_2, \dots)}$$

Problème : comment on calcul les termes de la fraction ?

# Classification Bayésienne

Dans le cas qui nous intéresse ici on a une classe  $Y$  et de nombreux attributs  $X_i$

$$P(Y|X_1, X_2, \dots) = \frac{P(X_1, X_2, \dots | Y)P(Y)}{P(X_1, X_2, \dots)}$$

Pour déterminer la classe on utilise le principe de **maximum de vraisemblance**, i.e.

on part du principe que la classe d'appartenance  $Y$  de l'objet est celle qui maximise la probabilité ci dessus

# Classification Bayésienne

$$P(Y|X_1, X_2, \dots) = \frac{P(X_1, X_2, \dots | Y)P(Y)}{P(X_1, X_2, \dots)}$$

Problème : comment on obtient les termes de la fraction de <sup>droite</sup>~~gauche~~ ?



# Classification Bayésienne

	$x_1$	$x_2$	$x_3$	$y$
Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

$P(y)$  non = 7/10  
oui = 3/10

# Classification Bayésienne

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	MarSt	TaxInc	Class
No	Married	135K	?

# Classification Bayésienne naïve

**Solution** : on considère les variables indépendantes entre elles (hypothèse naïve)

Avec cette hypothèse on obtient l'équation suivante :

$$P(Y|X_1, X_2, \dots) = \frac{P(X_1|Y)P(X_2|Y)\dots P(Y)}{P(X_1), P(X_2)\dots}$$

# Estimation des probabilités

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	<u>Single</u>	125K	<u>No</u>
2	No	Married	100K	No
3	No	<u>Single</u>	70K	<u>No</u>
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	<u>Single</u>	85K	<del>Yes</del>
9	No	Married	75K	No
10	No	<u>Single</u>	90K	<del>Yes</del>

Dans le cas des attributs discret, on compte et on divise

$$P(Y = No) = \frac{(n_{no})}{n} \quad 7/10$$

$$P(Single|No) = \frac{(n_{single} \wedge n_{no})}{n_{no}} \quad \frac{2}{7}$$

# Estimation des probabilités (variables continues)

★ fonction de probabilité

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Pour les variables continues 2 options :

- On discrétise (voir le cours numéro 4)
- On projette sur une fonction de probabilité

# Exemple

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	MarSt	TaxInc	Class
No	Married	135K	?

$$\left. \begin{array}{l} P(y): \text{swi} \\ P(y): \text{non} \end{array} \right\} P(\text{oui} | x) = \text{photo}$$
$$P(\text{non} | x) = \text{photo}$$

# Discussion

**Avantages :**

**Limites :**

# Plan

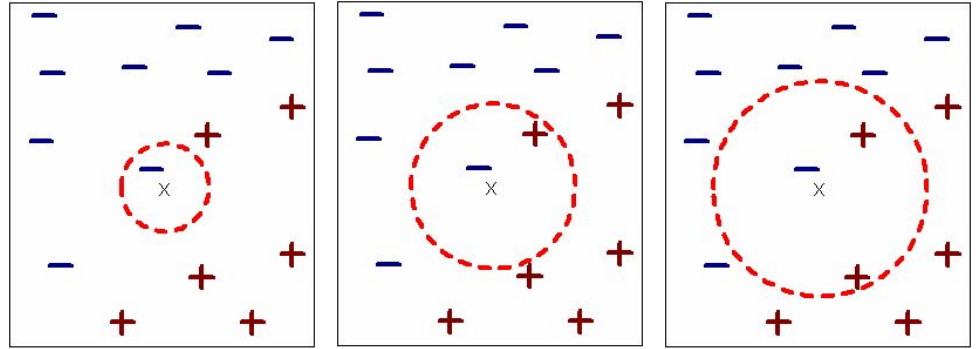
1 - Réseaux Bayésiens

**2 - K plus proches voisins**



# Principes

On cherche à connaître la classe de  $k$  voisins proches et on attribue une classe à l'objet par vote majoritaire.









# Distances

La fonction de distance la plus souvent utilisée est la distance euclidienne

$$D(a, b) = \sqrt{\sum (a_i - b_i)^2}$$

# Exemple

0 or 1    0 & 1    discret (paral, pas pareil)

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	?

- tous normaliser







$$\frac{V - \min}{\max - \min}$$

ex: 
$$\frac{35 - 22}{63 - 22}$$

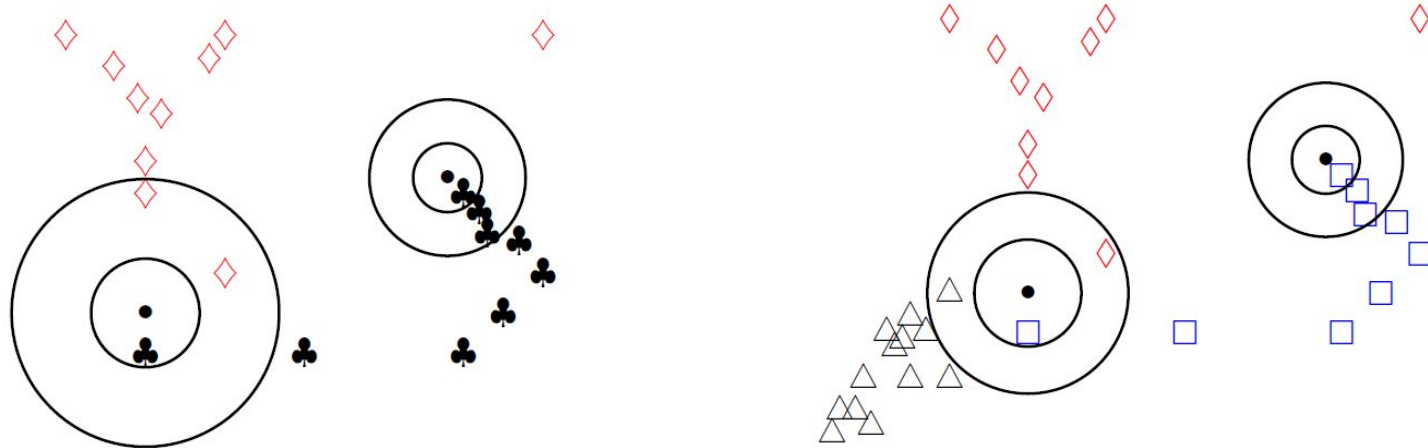
# Exemple

*K= doit être impair de préférence*

*! attention normaliser avant !  
sans normaliser*

Customer	Age	Income	No. credit cards	Loyal	Distance from David
John 	35	35K	3	No	$\text{sqrt} [(35-37)^2 + (35-50)^2 + (3-2)^2] = 15.16$
Rachel 	22	50K	2	Yes	$\text{sqrt} [(22-37)^2 + (50-50)^2 + (2-2)^2] = 15$
Hannah 	63	200K	1	No	$\text{sqrt} [(63-37)^2 + (200-50)^2 + (1-2)^2] = 152.23$
Tom 	59	170K	1	No	$\text{sqrt} [(59-37)^2 + (170-50)^2 + (1-2)^2] = 122$
Nellie 	25	40K	4	Yes	$\text{sqrt} [(25-37)^2 + (40-50)^2 + (4-2)^2] = 15.74$
David 	37	50K	2	Yes	

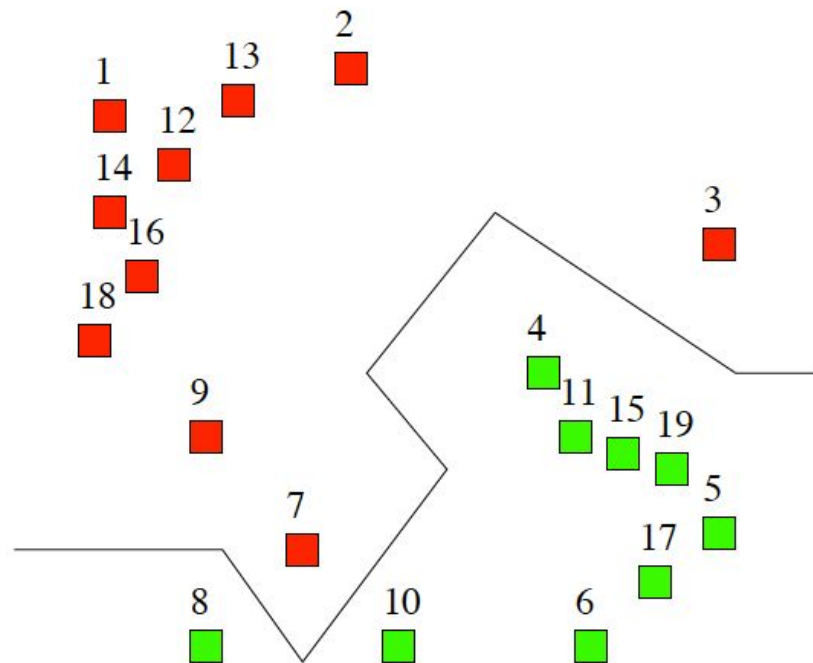
# Choix de K



Il n'y a pas de méthode absolue permettant de déterminer k à priori. De manière empirique, la formule ci-à-côté est un bon estimateur (n nombre d'exemples et C le nombre de classes)

$$k \approx \sqrt{\overset{\text{nbr exemple}}{n} / \overset{\text{nbr class}}{C}}$$

# Discussion



# Discussion

2 colonne faire un ratio pour trouver 'y'

Calculs Precision - relative  
- absolue

nettoyage complet pour le p

- nettoyage
  - retirer l'inutile
  - normaliser
  - réduire le nbr d'attribut
  - vérifier les spécifications (expliquer les données veulent dire quoi)
  - classificateur
- justifier dans le power point.