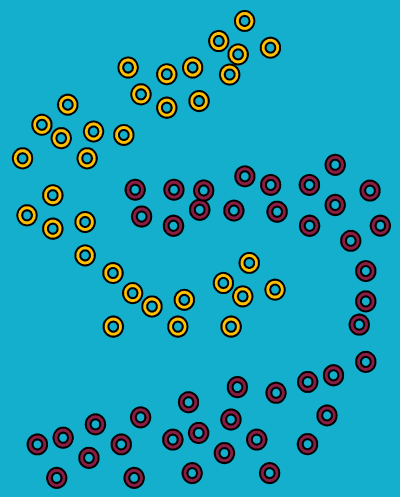
1



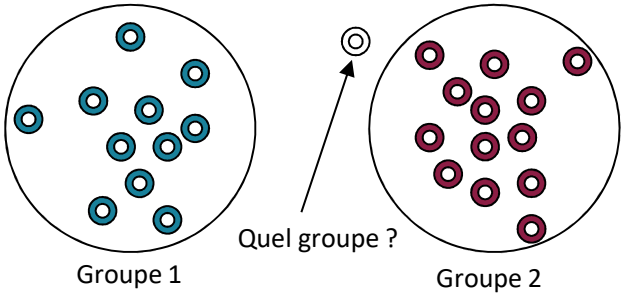
Calcul de distance

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

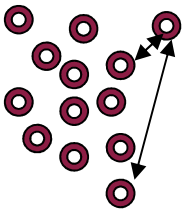
FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Plan

- Contexte : prédiction avec K-NN



- Notion de distance
  - Applications
  - Définition et calcul



FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Qu'est-ce que la classification supervisée ?

Elle permet de **prédire** si un élément est membre d'un groupe ou d 'une catégorie donné.

- Classes
  - Identification de groupes avec des profils particuliers
    - ex: sain/ malade,
    - normal / fraude
    - …
  - Possibilité de décider de l'appartenance d'une entité à une classe

Caractéristiques

- Apprentissage supervisé : classes connues à l'avance
  - Pb : qualité de la classification (taux d'erreur)

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Illustration

Attributs

- Infos sur les observations
  - Montant, heure, nb articles,…
  - Couleur voiture, nb chevaux,…
  - Age, antécédents, diagnostics,…

Classe

Ce que l'on veut prédire

Observations

- Tickets caisse
- Assurés
- Patients
- Séjours
- …

ID	attribut1	attribut2	attribut3	attribut4	Classe
Obs1	1	Oui	65	1.1	Non
Obs2	1	Non	15	1.5	Non
Obs3	3	Oui	30	1.0	Oui
Obs4	4	Oui	36	0.5	Non
Obs5	1	Oui	25	2.0	Non

ObsX	5	Oui	30	2.0	?

Objectif : Prédire la classe de ObsX

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Exemple : comestible ou pas ?

odor	spore-print-color	habitat	cap-color	poisonous
n	n	g	w	e
y	w	l	n	p
f	h	p	g	p
f	h	g	y	p
n	n	g	g	e
f	h	u	w	p
l	n	d	y	e
s	w	l	e	p
s	w	l	n	p
f	h	d	g	?

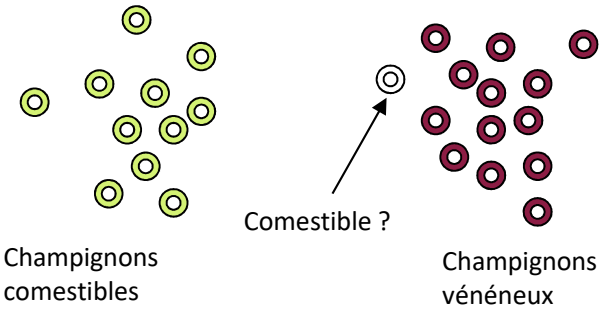
Source: mushroom dataset (https://archive.ics.uci.edu/ml/datasets/mushroom)
From Audobon Society Field Guide; mushrooms described in terms of physical characteristics; classification: poisonous or edible

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Principe K-NN

Idée simple

1) calculer distance de similarité



2) regarder les K voisins plus proches

(K-Nearest Neighbors)

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Distance

Notion **indispensable** au clustering

Comment définir le degré de ressemblance entre deux observations ?

- Quelle observation (**obs3** ou **obs6**) est la plus proche de **Obs4** ?

Obs3	3	Oui	30	1.0	Oui
Obs6	5	Non	49	1.5	Oui
Obs4	4	Oui	36	0.5	Non

- Quel mot (**Dupont** ou **Durant**) est le plus proche de **Dumont** ?

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Distances : applications

Segmentation de données

Classe 1	Obs1	1	Oui	65	1.1	Non
	Obs2	1	Non	15	1.5	Non
	Obs5	1	Oui	25	2.0	Non
Classe 2	Obs3	3	Oui	30	1.0	Oui
	Obs4	4	Oui	36	0.5	Non
Classe 3	Obs6	5	Non	49	1.5	Oui

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Quelques applications

Marketing

segmentation du marché en découvrant des groupes de clients distincts à partir de bases de données d'achats.

Environnement

identification des zones terrestres similaires (en termes d'utilisation) dans une base de données d'observation de la terre.

Assurance

identification de groupes d'assurés distincts associés à un nombre important de déclarations.

Planification de villes

identification de groupes d'habitations suivant le type d'habitation, valeur, localisation géographique, …

Médecine

Localisation de tumeurs dans le cerveau

- Nuage de points du cerveau fournis par le neurologue
- Identification des points définissant une tumeur

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Distance : définition

Mesure la dissimilarité entre 2 objets

- Distance élevée -> objets différents

Propriétés

1. $d(x, y) \geq 0$
2. $d(x, y) = 0 \text{ iff } x = y$
3. $d(x, y) = d(y, x)$
4. $d(x, z) \leq d(x, y) + d(y, z)$

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Distance entre 2 attributs (1)

Avant de calculer la distance entre 2 objets

ID	attribut1	attribut2	attribut3	attribut4	attribut5
Obs1	1	Oui	65	1.1	Non
Obs2	1	Non	15	1.5	Non
Obs3	3	Oui	30	1.0	Oui
Obs4	4	Oui	36	0.5	Non
Obs5	1	Oui	25	2.0	Non
Obs6	5	Non	49	1.5	Oui

Distance entre 2 valeurs ?

d('oui','non') ?

d(1.5, 1.0) ?

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Distance entre 2 attributs (2)

Pour les données numériques

- Ex: « **age** » [18,118]

$$d(a_1, a_2) = \left| \frac{a_1 - a_2}{d_{\max}} \right|$$

→ **Pour distance normalisée**

- d(80,40) = 0,4 (d_{max} = 118 – 18 = 100)
- d(18,19) = 0,01

Pour les données binaires :

- 0 si identique, 1 sinon
- d(0,0)=d(1,1)=0
- d(0,1)=d(1,0)=1

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Distance entre 2 attributs (3)

Pour les données qualitatives :

- Ex: **‘rouge’, ‘vert’, ‘bleu’, ‘orange’, ‘noir’**
- 0 si identique, 1 sinon
- d(‘rouge’,’bleu’) = d(‘vert’,’bleu’) = 1
- d(‘bleu’,’bleu’) = 0

Pour les données qualitatives ordonnées :

- Ex: **‘Très satisfait’, ‘satisfait’, ‘passable’, ‘mécontent’, ‘très mécontent’**
- d(‘très satisfait’,’très mécontent’) = 1
- d(‘très satisfait’, ’mécontent’) = 0,75
- d(‘très satisfait’,‘passable’) = 0,5
- d(‘très satisfait’,’satisfait’) = 0,25

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

D’autres distances entre 2 attributs

Entre mots

- Ex: d(**‘dupont’**,**’dumont’**)
- Jaro-winkler
- Levenshtein

Données binaires asymétriques

- Ex: test VIH : 1 infecté, 0 sinon
- 2 personnes ayant la valeur 1 sont plus proches que 2 personnes ayant la valeur 0
- Jaccard

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

C'est à vous

Complétez le tableau du bas (distance normalisée)

ID	Name	subtype	mass	width	height	color_score
F1	lemon	spanish_belsan	194	7.2	10.3	0.70
F2	lemon	unknown	120	6.0	8.4	0.74
F3	orange	turkey_navel	158	7.2	7.8	0.77

	mass	width	height	color_score
d(F1,F2)				
d(F2,F3)				
d(F1,F3)				

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Distance entre 2 objets (1)

ID	attribut1	attribut2	attribut3	attribut4	attribut5
Obs1	1	Oui	65	1.1	Non
Obs2	1	Non	15	1.5	Non
Obs3	3	Oui	30	1.0	Oui
Obs4	4	Oui	36	0.5	Non
Obs5	1	Oui	25	2.0	Non
Obs6	5	Non	49	1.5	Oui

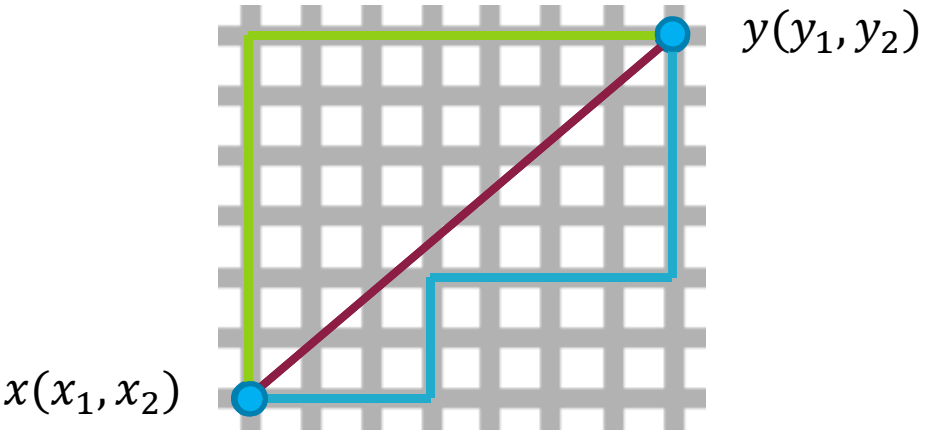
Comment combiner distances entre 2 valeurs ?

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Distance entre 2 objets (2)

Distance $d(x, y)$ entre deux objets $x(x_1, x_2, \ldots, x_n)$ et $y(y_1, y_2, \ldots, y_n)$

- n: nombre de dimensions
- $dessin : n = 2$



Distance Euclidienne

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Distance de Manhattan

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

Distance de Minkowski

$$d(x, y) = \sqrt[q]{\sum_{i=1}^{n} [x_i - y_i|^q}$$

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Exemple : distance de Manhattan

d(obs2,obs3) ?

ID	attribut1	attribut2	attribut3	attribut4	attribut5
Obs1	1	Oui	65	1.1	Non
Obs2	**1**	**Non**	**15**	**1.5**	**Non**
Obs3	**3**	**Oui**	**30**	**1.0**	**Oui**
Obs4	4	Oui	36	0.5	Non
Obs5	1	Oui	25	2.0	Non
Obs6	5	Non	49	1.5	Oui

$$d(obs2,obs3) = d(1,3) + d('Non','Oui') + d(15,30) + d(1.5,1.0) + d('Non','Oui')$$

$$d(obs2,obs3) = \frac{|1-3|}{5-1} + 1 + \frac{|15-30|}{65-15} + \frac{|1.5-1.0|}{2.0-0.5} + 1$$

$$d(obs2,obs3) = 3{,}13$$

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

C’est à vous !

A partir du tableau obtenu précédemment, calculez la distance de Manhattan normalisée

	mass	width	height	color_score	$d_{manhattan}$
d(F1,F2)	1	1	0,76	0,57	
d(F2,F3)	0,51	1	0,24	0,42	
d(F1,F3)	0,48	0	1	1	

Quels sont les deux fruits les plus proches ?