

FTML

Exercice 1

1)

Matrice de confusion pour Est_1:

		Obtenu	
		T	F
Prédire	T	6	8
	F	1	2

Risque empirique de Est_1:

$$LF(\tilde{f}, f) = 8 \cdot 1 + x \cdot 1$$

Matrice de confusion pour Est_2

		Obtenu	
		T	F
Prédire	T	4	1
	F	3	9

Risque empirique de Est_2:

$$LF(\tilde{f}, f) = 1 \cdot 1 + x \cdot 3$$

On peut changer la valeur x pour donner l'avantage à un estimateur ou à l'autre:

Si on veut avantager l'estimateur 1 on a :

- $x + 8 < 3x + 1 \iff x > \frac{7}{2}$

Et donc si on veut avantager l'estimateur 2 on a :

- $x + 8 > 3x + 1 \iff x < \frac{7}{2}$

b)

c)

On peut aisément discerner un modèle permettant de séparer

d)

$Y = \text{True}$	1	2
X_1	5	2
X_2	2	5

$Y = \text{False}$	1	2
X_1	3	7
X_2	6	4

e)

c)

$$P(Y=T / X_1=1, X_2=1) = P(Y=T) \times P(X_1=1 / Y=T) \times P(X_2=1 / Y=T)$$

$$= \frac{7}{17} \times \frac{5}{7} \times \frac{2}{7} \approx 0,084$$

$$P(Y=T / X_1=1, X_2=2) = \frac{7}{17} \times \frac{5}{7} \times \frac{5}{7} \approx 0,210$$

$$P(Y=T / X_1=2, X_2=1) = \frac{7}{17} \times \frac{2}{7} \times \frac{2}{7} \approx 0,034$$

$$P(Y=T / X_1=2, X_2=2) = \frac{7}{17} \times \frac{2}{7} \times \frac{5}{7} \approx 0,084$$

$$P(Y=F / X_1=1, X_2=1) = \frac{10}{17} \times \frac{3}{10} \times \frac{6}{10} \approx 0,106$$

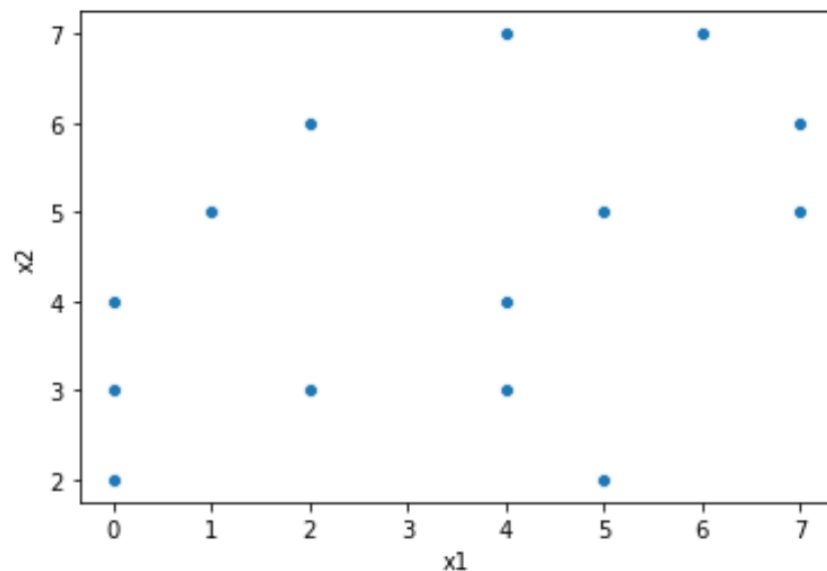
$$P(Y=F / X_1=1, X_2=2) = \frac{10}{17} \times \frac{3}{10} \times \frac{4}{10} \approx 0,071$$

$$P(Y=F / X_1=2, X_2=1) = \frac{10}{17} \times \frac{7}{10} \times \frac{6}{10} \approx 0,247$$

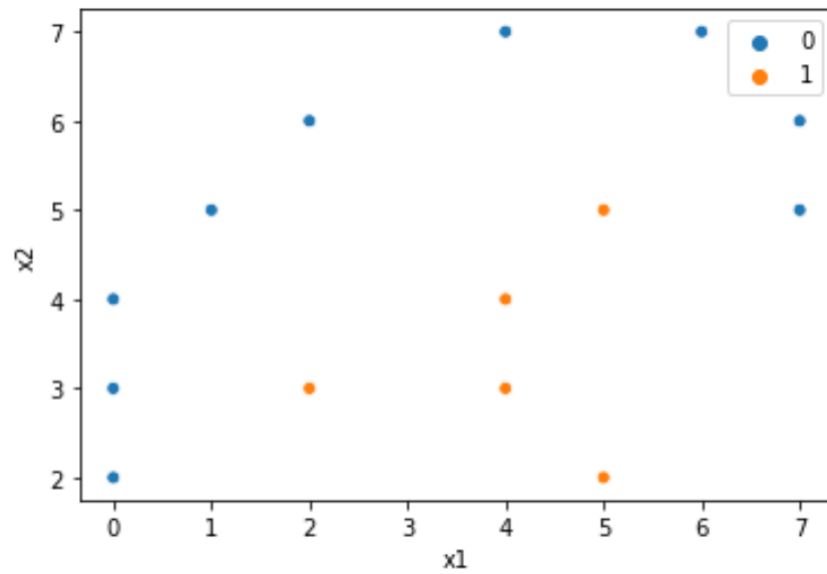
$$P(Y=F / X_1=2, X_2=2) = \frac{10}{17} \times \frac{7}{10} \times \frac{4}{10} \approx 0,165$$

Exercice 2

a)



Si on connaît le nombre de classe que l'on veut obtenir on peut faire un k-means ou de l'agglomérative clustering. En l'occurrence un spectral clustering serait assez efficace.



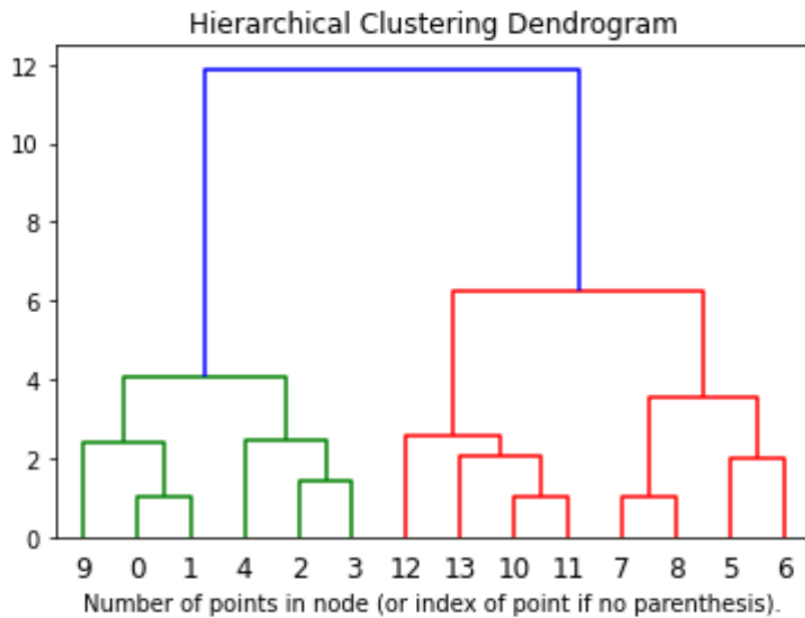
b)

On se munie d'une fonction de distance euclidienne pour faire la distance de Ward.

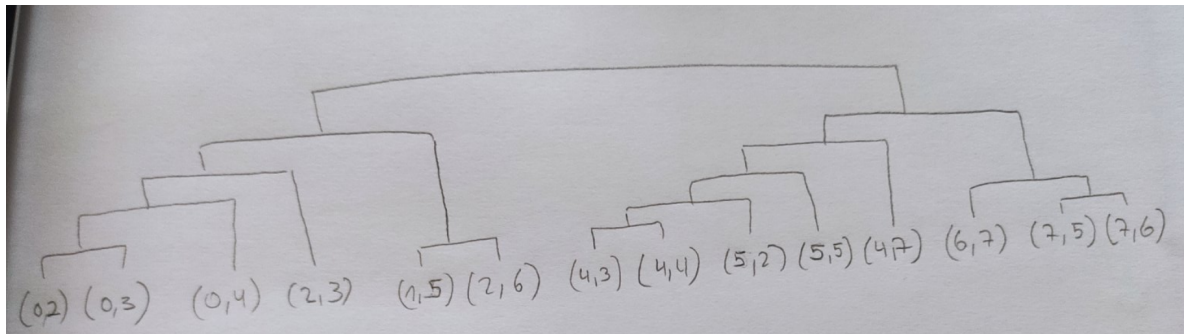
Les points sont les suivants:

```
array([[0, 2],
       [0, 3],
       [0, 4],
       [1, 5],
       [2, 6],
       [4, 7],
       [6, 7],
       [7, 6],
       [7, 5],
       [2, 3],
       [4, 3],
       [4, 4],
       [5, 2],
       [5, 5]])
```

Et le Dendograme associé est donc:



(Ou bien le suivant réaliser à la main)

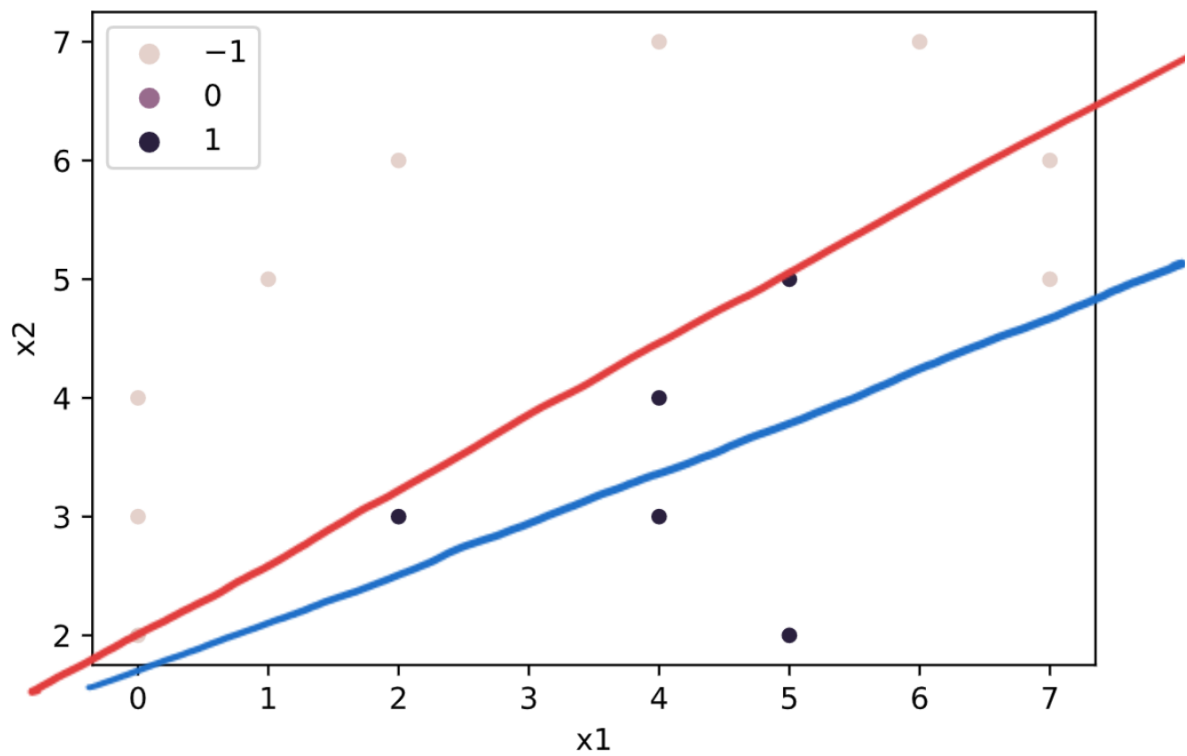


c)

On peut utiliser un SVM linéaire pour séparer le jeu de données en deux classes mais il est clair que la séparation ne serait pas parfaite et que l'on aurait des erreurs de prédiction. En effet un SVM linéaire n'est pas pratique pour séparer deux classes dont l'une est à l'intérieur de l'autre.

d)

On prend une pénalisation C assez grande pour la courbe rouge est moins grande pour la courbe bleue.



e)

Un cercle pourrait parfaitement séparer les deux classes. Ainsi on peut proposer la fonction $f(x_1, x_2) = x_1^2 + x_2^2$

Exercice 3

a)

A partir d'un tableau croisé d'effectifs je peut réaliser un test de χ^2 . Ce test consiste à tester l'existence d'une relation entre deux caractères discrets.

Il tout d'abord à réaliser un tableau d'effectifs théoriques, puis, à partir de ce tableau et du tableau d'effectifs croisés on obtient la matrice de χ^2 . Ensuite on détermine le degré de liberté (nombre de ligne * nombre de colonne).

La valeur de χ^2 s'obtient en sommant les élément de la matrice de χ^2 .

On regarde alors dans la table de χ^2 avec un risque alpha que l'on choisira et le degré de liberté obtenue plus tôt. On pourra alors rejeté ou non l'hypothèse nulle (hypothèse d'indépendance).

b)

Je testerais en tâtonnant en commençant par exemple en essayant de pulvérisé 2, 3 ou 4 points sur la surface de la sphère puis en augmentant petit à petit. J'utiliserais aussi mes connaissance empirique pour mieux juger des nombre de points à tester.

c)

Risque : C'est le fait de connaître des probabilités à travers les expériences et observations connues et déjà obtenues afin de traiter des probabilités inconnues.

Ambiguïté : On ne connaît pas les probabilités des expériences car elles sont inconnues ou sujettes à interprétation