

Exercice 1

a)

X1	X2	Y	Y_EST_1	Y_EST_1
1	2	T	T	T
2	2	F	T	F
1	1	T	F	x
1	2	F	T	x
1	2	T	T	T
2	2	F	T	F
2	1	F	T	x
2	2	T	T	F
2	1	F	T	F
2	1	F	T	x
1	2	T	T	T
1	1	F	F	F
2	1	T	T	x
1	1	F	F	F
2	2	F	T	F
2	1	F	T	F
1	2	T	T	T

* Y_Est_1

predicted

	T	F
actual	6	1
	8	2

* Y_Est_2

predicted

	T	F
actual	4	3
	1	9

Risque empirique :

$$\frac{1}{17} (8 \times 1 + 1 \times x) \\ = \boxed{\frac{8+x}{17}}$$

Risque empirique :

$$\frac{1}{17} (1 \times 1 + 3 \times x) \\ = \boxed{\frac{1+3x}{17}}$$

$$\frac{8+x}{17} - \frac{1-3x}{17} > 0 \\ \Leftrightarrow 7-2x > 0 \\ \Leftrightarrow 7 > 2x \\ \Leftrightarrow 3,5 > x$$

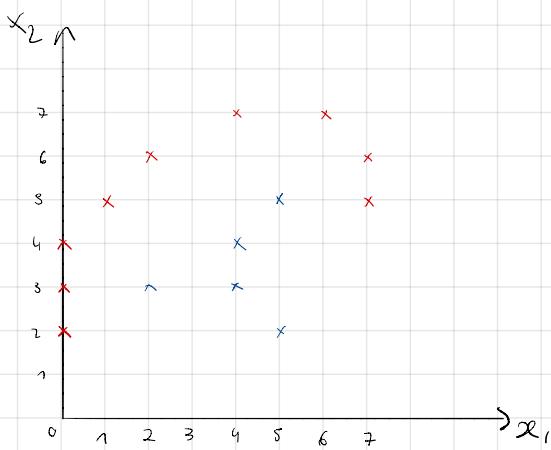
le risque empirique de l'estimateur Y_Est_1 est plus grand que celui de l'estimateur Y_Est_2 si $x < 3,5$

b) $x=2$

c) Le modèle du bayésien naïf n'est pas adapté dans ce cadre car les ne sont pas indépendant, et il n'y a pas assez de variation dans les valeurs de x_i .

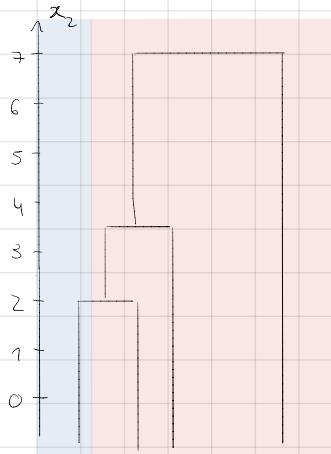
Y	Y_Est_3
T	T
F	F
T	T
F	F
T	T
F	F
F	F
T	T
F	F
F	F
T	T
F	F
F	F
F	F
T	T

Exercice 2



a) La méthode des k-means et la classification ascendante hiérarchique semblent pouvoir les distinguer

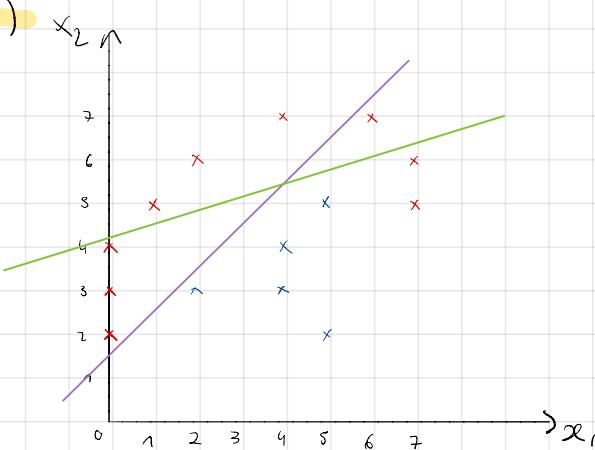
b)



c)

Il est possible de les séparer avec un SVM linéaire, mais il y aura sûrement toujours des points mal classés, même avec un très grand coût de pénalisation. Car comme on peut le voir dans le graphe de la question a) les données ne peuvent pas être séparées avec une droite. On aurait pu utiliser une fonction kernel non linéaire pour projeter les données dans un espace où leur séparation par une droite plus facile.

d)



Dans cette exemple la droite verte à un taux de pénalisation plus petit que la droite violette, on peut facilement remarquer que la droite violette essaie d'avoir moins d'erreur de classification.

$C=5$ C étant le coût de pénalisation.
 $C=100$

Exercice 3

a) Pour savoir si 2 variables sont indépendantes, on utilise le Test du χ^2

On a les effectifs croisés $E_{i,j}$, et on calcule l'effectif croisé observé :

$$O_{i,j} = \#\{X=i \text{ & } Y=j\}$$

Puis on calcule l'écart entre les deux :

$$\tau = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

Si τ suit une loi du χ^2 à $(I-1, J-1)$ DL, alors les variables sont indépendantes.

En pratique on utilise chi2-contingency de la librairie scipy.

b) On peut essayer de le trouver en montant (par dessin ou analyse) que le nombre de point pulvérisable est d'au moins n , puis que $n+1$ point n'est pas pulvérisable. Donc le nombre maximum de points pulvérisable est n .