Partiel FTML

Exercice 1

1) Est_1

TP 6	FP 8
FN 1	TN 2

Est_2

TP 4	FP 1
FN 3	TN 9

$$\Rightarrow R_{emp}(est\_1) = \frac{1 \times x + 8 \times 1}{17} = \frac{x}{17} + \frac{8}{17}$$

$$R_{emp}(est\_2) = \frac{3 \times x + 1 \times 1}{17} = \frac{3}{17}x + \frac{1}{17}$$

$$\Rightarrow R_{emp}(est\_1) > R_{emp}(est\_2)$$

$$\frac{1}{17}x + \frac{8}{17} > \frac{3}{17}x + \frac{1}{17}$$

$$\frac{7}{17} > \frac{2}{17}x$$

$$7 > 2x \Rightarrow x < \frac{7}{2}$$

$\Rightarrow$ le risque empirique de l'estimateur 1 est supérieur à celui de l'estimateur 2 pr $x < \frac{7}{2}$

2) le faux négatif coûtent plus cher que le faux positif, on souhaite minimiser le nombre de faux négatif de notre estimateur.

	T	F	Est_3
1,2	□	–	→ T
2,2	–	⊓	→ F
1,1	–	⊤	→ T /F (equivalent)
2,1	–	□	→ F

Notre estimateur renverra donc :

- vrai pour 1,2
- faux pour 2,2
- faux pour 1,1 (choix faux car je veux
- faux pour 2,1 minimiser les faux negatifs)

Et aura un risque empirique de $\frac{7}{17}$

3) Dans ce cadre, il n'est pas nécessaire de se placer dans le cas du Bayesien naif car :

- nous avons suffisement de données sur la distribution de x_1 et x_2 pour ne pas avoir besoin de l'hypothèse d'indépendance

4)

$Y=$ True	$X=1$	$X=2$
X_1	5	2
X_2	2	5

$Y=$ False	$X=1$	$X=2$
X_1	3	7
X_2	6	4

5) $P(Y=T)=7 \qquad P(Y=F)=10$

- $P(Y=T \mid X_1=1, X_2=1) = \dfrac{P(X_1=1 \mid Y=T) \times P(X_2=1 \mid Y=T) \times P(Y=T)}{P(X_1=1, X_2=1)}$

$$= \frac{\frac{5}{7} \times \frac{2}{7} \times \frac{7}{17}}{\frac{3}{17}} = \frac{17}{3} \times \frac{70}{833}$$

$$P(Y=F \mid X_1=1, X_2=1) = \frac{\frac{3}{10} \times \frac{6}{\cancel{10}} \times \frac{\cancel{10}}{17}}{\frac{3}{17}} = \frac{17}{3} \times \frac{18}{170}$$

$\Rightarrow P(Y=T \mid X_1=1, X_2=1) < P(Y=F \mid X_1=1, X_2=1)$

$\Rightarrow Y=F$ si $X_1=1, X_2=1$

Similairement :

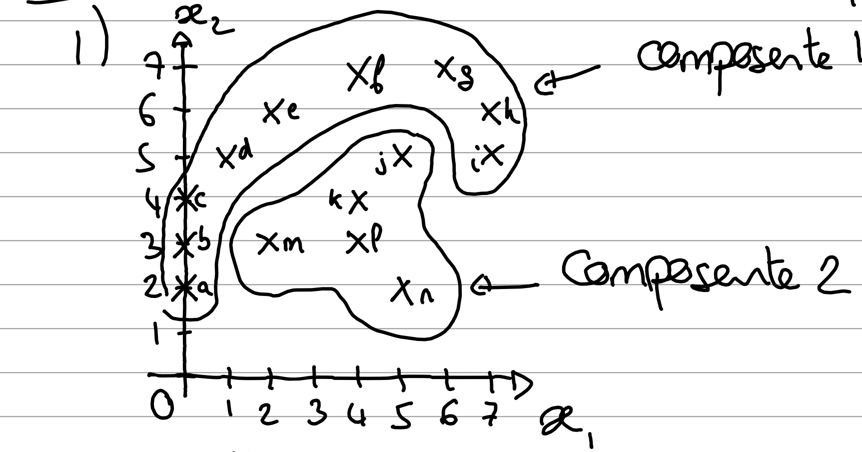
- pour $X_1=1$ et $X_2=2$, $Y=T$
- pour $X_1=2$ et $X_2=1$, $Y=F$
- pour $X_1=2$ et $X_2=2$, $Y=F$

$\Rightarrow$ Notre estimateur bayésien renvoie les mêmes valeurs que l'estimateur optimal

Exercice 2

(les lettres sont utilisés pour le q2)

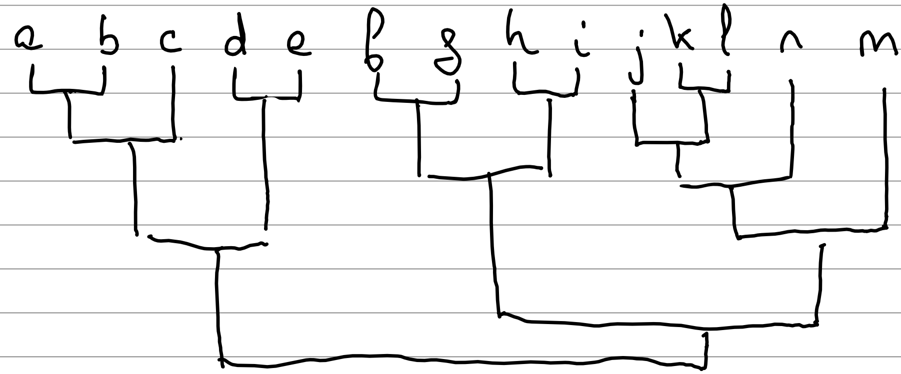
1)



Les méthodes non supervisés qui paraissent à même de les distinguer sont :

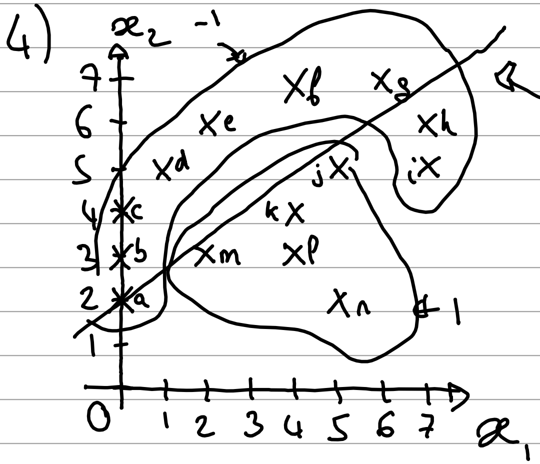
- le spectral clustering
- la classification hiérarchique
- le DBSCAN
- l'optics

2) Le coût de l'intégration sera calculé avec une distance de Manhattan

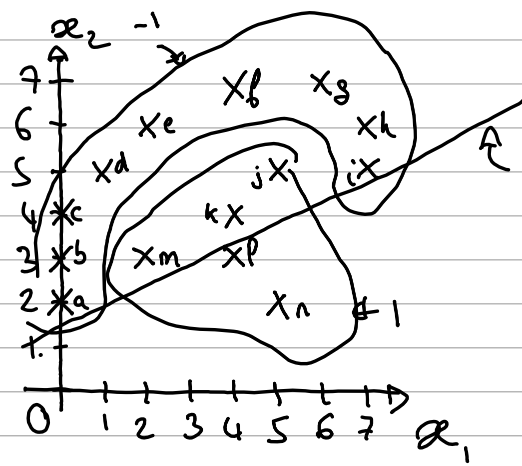


3) On ne pourra pas séparer parfaitement les 2 classes avec un SVM linéaire car l'un encercle l'autre

4)



exemple d'SVM linéaire pénalisant les erreurs de classe 1



SVM linéaire pénalisant les erreurs de classe -1

e) On peut utiliser le kernel RBF (radial basis function) pour séparer les 2 classes.

$$\left(k(x, x') = \exp\left(- \frac{\| x - x' \|^2}{2\sigma^2} \right) \right)$$

Exercice 3

1) À partir d'un tableau croisé d'effectifs, on peut utiliser l'indice chi2 (χ_2) pour tester si les variables sont indépendantes

2) Pour trouver le nombre maximum de points pulvérisable à la surface d'une sphère, je chercherais combien de points aux maximum je peux séparer avec l'outil choisi (triangle, rectangle, ...)

3) La différence entre risque et ambiguité est qu'un risque à une probabilité connu, tandis que l'ambiguité n'est pas connu à l'avance