BURKART
Loïc

Juillet 2020
SCIA

FTML

Ex 1 a) Pour EST-1, on a :

$FP = 8 \qquad FN = 1$

$$ERM1 = \frac{1}{9}(1 \times FP + x \times FN)$$
$$= \frac{8}{9} + \frac{1}{9}x$$

Pour EST-2, on a :

$FP = 1 \qquad FN = 3$

$$ERM2 = \frac{1}{4}(1 \times FP + x \times FN)$$
$$= \frac{1}{4} + \frac{3}{4}x$$

On résout alors l'inéquation :

$$ERM1 < ERM2 \iff \frac{8}{9} + \frac{1}{9}x < \frac{1}{4} + \frac{3}{4}x$$
$$\iff \frac{8}{9} - \frac{1}{4} < \frac{3}{4}x - \frac{1}{9}x$$
$$\iff \frac{32-9}{36} < \frac{27-4}{36}x$$
$$\iff \frac{23}{36} < \frac{23}{36}x$$
$$\iff \underline{1 < x}$$

Donc, le risque empirique de EST_1 est inférieur à celui de EST_2 si et seulement si x est supérieur à 1.

b) On a :

Y = T	X = 1	X = 2
X1	5	2
X2	2	5

Y = F	X = 1	X = 2
X1	3	7
X2	6	4

On peut construire, sans hypothèse, l'EST_3 qui minimise le risque :

X1	X2	Y	Y_EST_3
1	2	T	T
2	2	F	F
1	1	T	T
1	2	F	F
1	2	T	T
2	2	F	F
2	1	F	F
2	2	T	T
2	1	F	F
2	1	F	F
1	2	T	T
1	1	F	F
2	1	T	T
1	1	F	F
2	2	F	F
2	1	F	F
1	2	T	T

ERM_3 = 0.

1/ c) Il n'était pas nécessaire de se placer dans le cadre du bayésien naïf car aucune hypothèse n'était faite. Ainsi, on ne fait pas l'hypothèse $\mathbb{P}(X1 = i \wedge X2 = j) = \mathbb{P}(X1 = i)\,\mathbb{P}(X2 = j)$.

d) Fait en début de question b) page 2.

e) Estimateur bayésien naïf :

$$\mathbb{P}(Y=T \mid X1=1 \wedge X2=1) = \mathbb{P}(Y=T)\,\mathbb{P}(X1=1 \mid Y=T)\,\mathbb{P}(X2=1 \mid Y=T)$$
$$= \frac{7}{17} \times \frac{5}{7} \times \frac{2}{7}$$
$$\simeq 0{,}41 \times 0{,}71 \times 0{,}29 \simeq 0{,}08$$

$$\mathbb{P}(Y=F \mid X1=1 \wedge X2=1) = \mathbb{P}(Y=F)\,\mathbb{P}(X1=1 \mid Y=F)\,\mathbb{P}(X2=1 \mid Y=F)$$
$$= \frac{10}{17} \times \frac{3}{10} \times \frac{6}{10} \simeq 0{,}59 \times 0{,}3 \times 0{,}6$$
$$\simeq 0{,}101$$

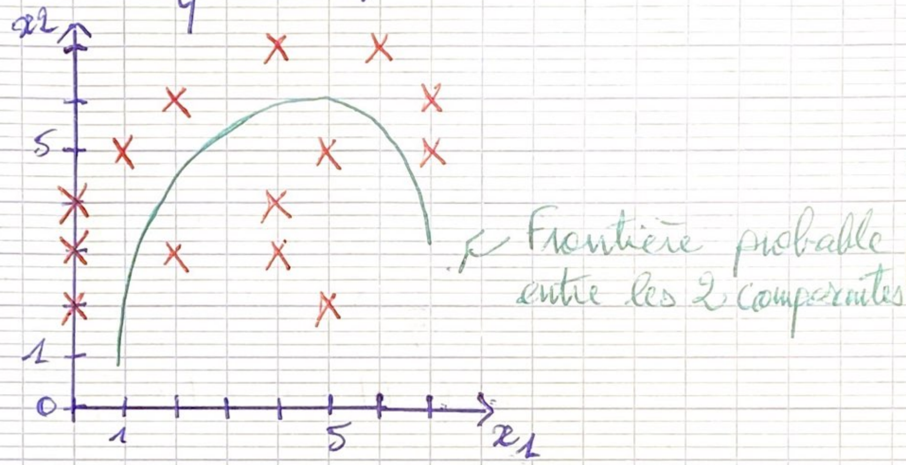
$\mathbb{P}(Y=T \mid X1=1 \wedge X2=2) \simeq 0{,}41 \times 0{,}71 \times 0{,}71 \simeq 0{,}21$
$\mathbb{P}(Y=F \mid X1=1 \wedge X2=2) \simeq 0{,}59 \times 0{,}3 \times 0{,}4 \simeq 0{,}07$.
$\mathbb{P}(Y=T \mid X1=2 \wedge X2=1) \simeq 0{,}41 \times 0{,}29 \times 0{,}29 \simeq 0{,}03$
$\mathbb{P}(Y=F \mid X1=2 \wedge X2=1) \simeq 0{,}59 \times 0{,}7 \times 0{,}6 \simeq 0{,}25$
$\mathbb{P}(Y=T \mid X1=2 \wedge X2=2) \simeq 0{,}41 \times 0{,}29 \times 0{,}7 \simeq 0{,}08$
$\mathbb{P}(Y=F \mid X1=2 \wedge X2=2) \simeq 0{,}59 \times 0{,}7 \times 0{,}4 \simeq 0{,}17$.

X1	X2	Y_EST_BN
1	1	F
1	2	T
2	1	F
2	2	F

Cet estimateur a pour risque :

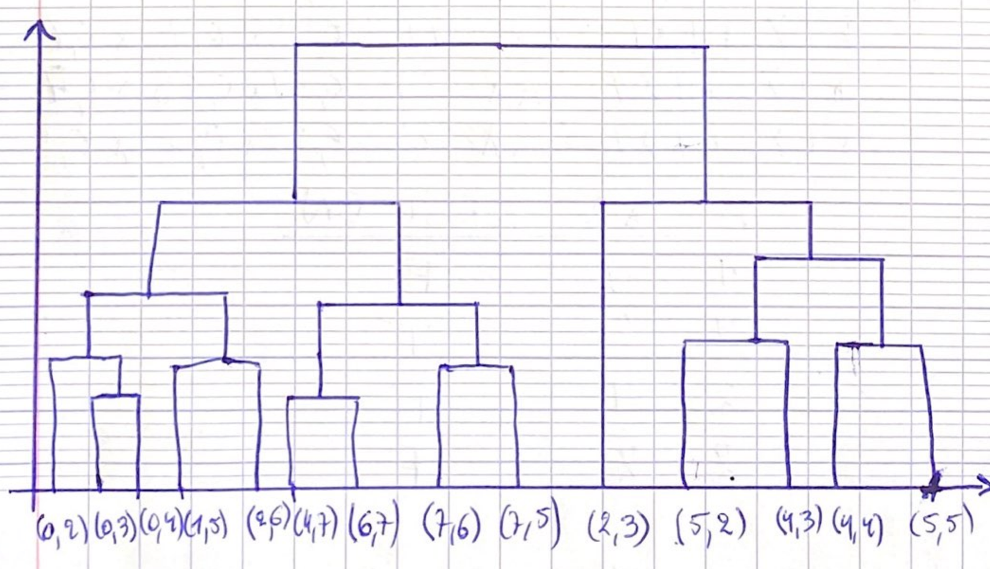
$$ERM = \frac{1}{4}(1+3x) = \frac{7}{4} \simeq 1{,}75.$$

Ex 2 a)



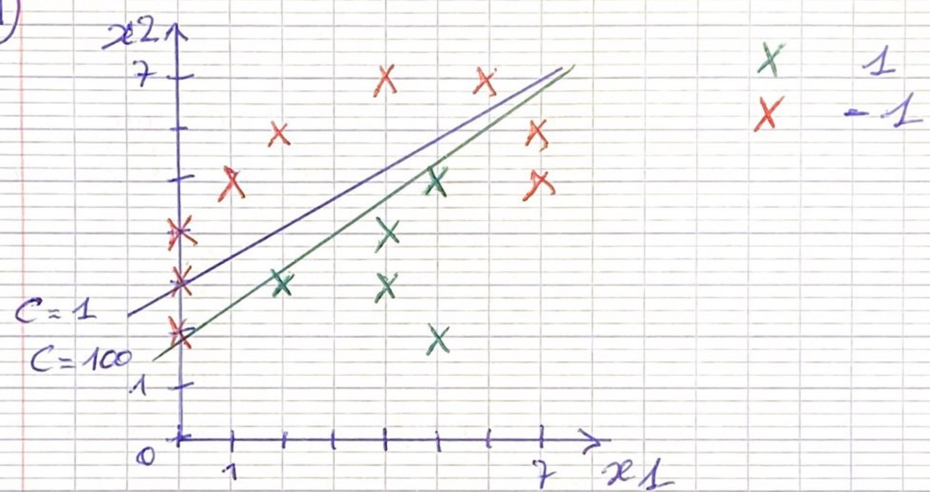
Une classification non supervisée par k-means ne paraît pas pouvoir séparer ces 2 composantes. Un clustering hiérarchique serait plus adapté.

b) Pour évaluer le coût, on choisit la distance euclidienne.



2/ c) Il n'est pas possible de séparer parfaitement les 2 classes par une droite, donc on ne peut pas les séparer parfaitement avec un SVM linéaire.

d)



Droite bleue : SVM linéaire avec $C = 1$.
Droite verte : SVM linéaire avec $C = 100$.

e) Pour séparer ces classes, on peut utiliser un kernel polynômial de degré 2.

On peut proposer :

$$f : (x_1, x_2) \longmapsto (-0{,}37\, x_1^2 + 3{,}17\, x_1 - 0{,}8 \,;\, x_2)$$

6/6

Ex 3 a) Pour savoir si les deux variables sont indépendantes, on effectue un test du chi 2. L'hypothèse nulle est l'indépendance des deux variables. On calcule alors l'écart relatif entre les deux variables. Cet écart suit une loi du χ^2. Si l'écart entre les valeurs obtenues et les valeurs attendues pour une telle loi est trop important (p-value faible), on rejette l'hypothèse nulle.

b) Pour connaître le nombre maximum de points pulvérisables à la surface d'une sphère il faut placer k points à la surface d'une sphère, puis faire deux classes et essayer de les séparer. Il faut faire ça pour toutes les classes possibles, puis avec k de plus en plus grand. Pour une certaine valeur de k on ne pourra plus séparer les classes : on a alors la borne inférieure de la dimension de Vapnik-Chervonenkis.

c) Le risque correspond, pour un estimateur donné, la moyenne des erreurs commises (avec la fonction de coût associée).
L'ambiguïté correspond à une situation dans laquelle aucun estimateur ne peut avoir un risque nul.