Christophe
Marabotto
SCIA
2021

FTML

Exercice 1

a) Matrice de confusion

Estimateur 1

		T	F
y'	T	6	8
	F	1	2

(y en colonnes)

Estimateur 2

		T	F
y'	T	4	1
	F	3	9

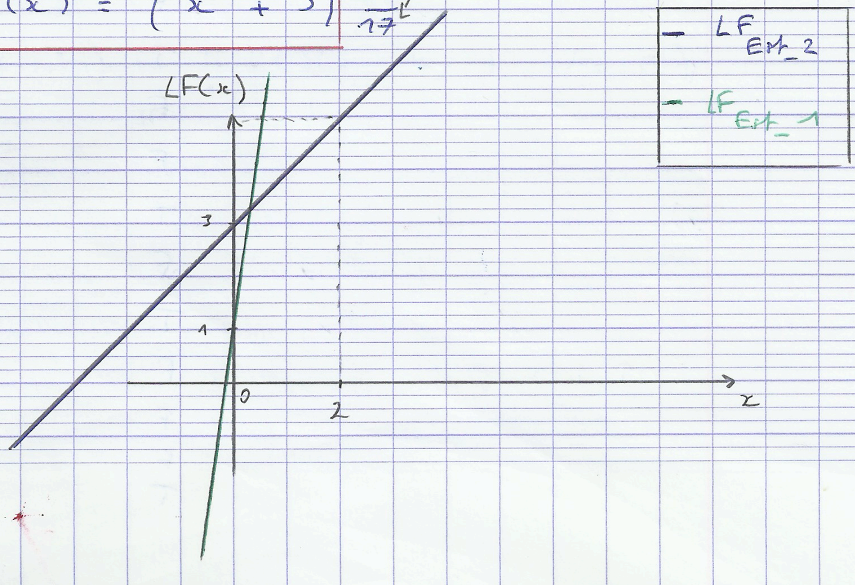
(y en colonnes)

Fonction de perte :

$$\begin{cases} LF(y, y') = 1 \ , \ (y, y') = (F, T) \\ LF(y, y') = x \ , \ (y, y') = (T, F) \end{cases}$$

$$\boxed{\begin{aligned} LF_{Est\_1}(x) &= (8x + 1)\,\frac{1}{17} \\ LF_{Est\_2}(x) &= (x + 3)\,\frac{1}{17} \end{aligned}}$$

taille de l'échantillon (non pris en compte dans le graphe)



b) On suppose que $x = 2$
À une constante multiplicative près, on a :

$$LF_{Est\_1}(2) = 17$$

$$LF_{Est\_2}(2) = 5 < LF_{Est\_1}(2)$$

L'estimateur 2 est plus précis si $x = 2$.

c)
Pour la suite de l'exercice, nous allons construire un estimateur bayésien naïf.

Remarque : je n'ai pas compris la question b.
On pourrait donner une nouvelle colonne Y tel que

Y_EST :
T
F
T
F
T
F
F
T
F
F
T
F
T
F
F
F
T

Néanmoins, cet estimateur ne pourrait pas exister car il devrait à la fois répondre
True pour $(x_1, x_2) = (2, 1)$ et
False pour $(x_1, x_2) = (2, 1)$.
Ce qui n'est pas possible avec nos données (il en manque).

d)

Y = True	1	2	Σ
x1	5	2	7 ✓
x2	2	5	7

Y = False	1	2	
x1	3	7	10 ✓
x2	6	4	10

e)

Cas 1 : $P(Y = T) = ?$

On calcule :

$$P(Y = T) \times \prod_i P(X_i = k_i \mid Y = T)$$

$$= \frac{7}{17} \times \prod_i \frac{k_i}{7}$$

Cas 2 : $P(Y = F) = ?$

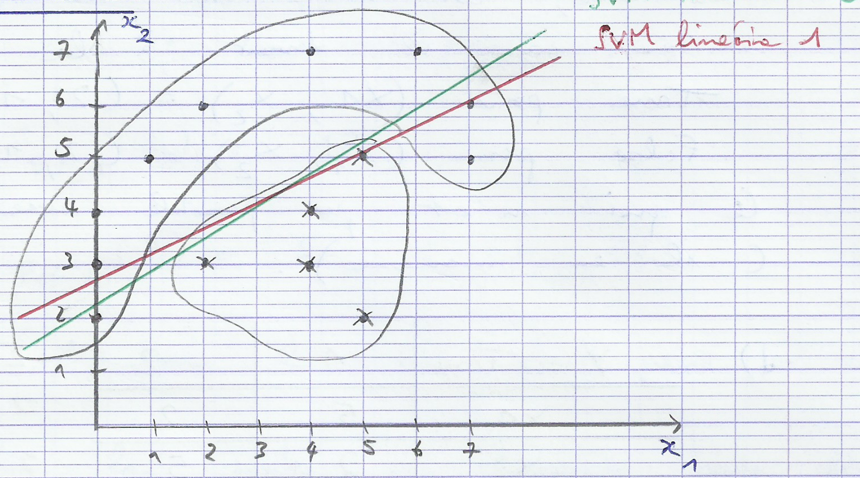
On calcule :

$$P(Y = F) \times \prod_i P(X_i = k_i \mid Y = F)$$

$$= \frac{10}{17} \times \prod_i \frac{k_i}{10}$$

Exercice 2 :

a)



Les méthodes suivantes sont capables d'effectuer ce clustering (non-supervisé) :

- K-means (moins bien que les 2 suivantes)
- DBSCAN
- Regroupement hiérarchique (CAH par exemple)

b)

c) Il n'est pas possible d'obtenir une séparation parfaite de ces 2 clusters avec un SVM linéaire car il n'existe pas de droite passant entre les 2 nuages de points.

d) cf. a)
Plus la pénalisation est grande, moins le modèle laisse passer de classe.
SVM linéaire 2 a donc une plus forte pénalisation.

Exercice 3

a) On a un tableau d'effectif croisé et on veut savoir si les variables sont indépendantes :

On choisit une p_value de 0,05, valeur minimale (utilisée en sociologie) à partir de laquelle l'hypothèse d'indépendance ne sera pas rejetée.
On calcul le χ^2 (chi2) avec l'outil chi2_contingency() et la p_value.
Si p_value << 0,05, on considère que les variables sont bien indépendantes.
⚠ Pour être plus rigoureux, une analyse statistique prenant en compte les degré de libertés des variables est nécessaire.

b) La surface d'une sphère peut pulvériser 3 point alignés mais pas 4.

⇒ Vapnik - Chervonenkis a une dimension de 3.

c) Différence entre risque et ambiguïté :

risque : réfère aux décisions avec des probabilités explicites

ambiguïté : réfère aux décisions sans connaissance des probabilités explicites
⇒ la prise de décision est bloquée