Yossra Antari
SCIA 2021

EXAM FTML

N.B : Par manque de temps je n'ai pas pu faire tout et même pour ce que j'ai fais je suis un peu déçue de ce que je rend. Le partiel était intéressant mais trop complexe (de mon point de vue) avec le temps imposé. Bonne correction tout de même !

Exercice 1

a) *Quel est le risque empirique, sur l'échantillon considéré, associé respectivement aux estimateurs EST_1 et EST_2 ? Comparez-les en fonction des valeurs de x.*

Matrice de confusion pour l'estimateur 1 :

	TRUE	FALSE
TRUE	6	1
FALSE	8	2

Matrice de confusion pour l'estimateur 2:

	TRUE	FALSE
TRUE	4	3
FALSE	1	9

Le risque empirique associé respectivement aux estimateurs 1 et 2 sont :

- $8 + x / 17$
- $1 + 3x / 17$

En fonction de la valeur de x, un estimateur peut être plus avantagé que l'autre, on remarque ainsi que :

- si $x < 3,5$: estimateur 1 est avantagé
- si $x >= 3,5$: estimateur 2 est avantagé

b) La question est *"construisez un estimateur qui minimise le risque empirique"*. Je ne suis pas sûre de bien comprendre. Prendre la colonne Y serait un estimateur qui minimise le risque car il n'y aurait plus d'erreur.

c) *Expliquez pourquoi il n'était pas nécessaire ici de se placer dans le cadre bayésien naïf. (Au moins deux raisons)*

d) *Produisez maintenant les tableaux agrégés où la corrélation n'apparaît plus entre X1 et X2*

Pour Y = FALSE

	1	2
1	2	1
2	4	3

Pour Y = TRUE

	1	2
1	1	4
2	1	1

e) *Utilisez ces tableaux pour produire l'estimateur bayesien naïf optimal. Comparez avec l'estimateur optimal trouvé plus haut.*

Pour X1 = 1 et X2 = 1 :

- P(Y = T, X1 = 1 et X2 = 1) = 7 / 17 * 1 /7 = **7/119**
- P(Y = F, X1 = 1 et X2 = 1) = 10 / 17 * 2 / 10 = **20/170**

Pour X1 = 1 et X2 = 2:

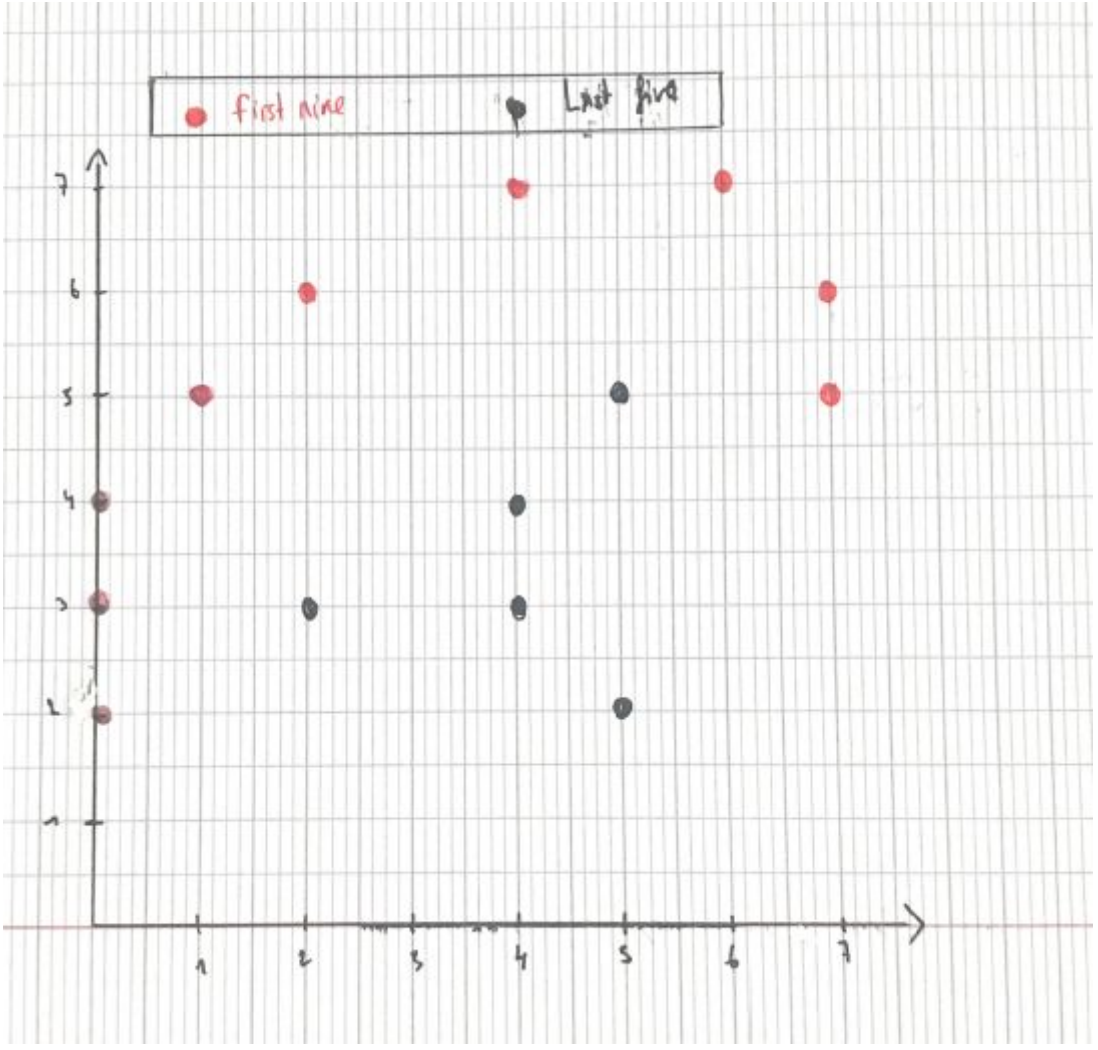
- P(Y = T, X1 = 1 et X2 = 2) = 7 / 17 * 4 / 7 = **28/119**
- P(Y = F, X1 = 1 et X2 = 2) = 10 / 17 * 1 / 10 = **10/170**

Nous faisons la même chose pour le reste des combinaisons restantes (je n'ai pas le temps de toutes les écrire) et nous nous retrouvons avec le tableau d'estimateur bayésien naïf suivant :

	1	2
1	**FALSE**	**TRUE**
2	**FALSE**	**FALSE**

Exercice 2

a) *Dessinez les points et identifiez deux composantes probables (les 9 premiers points contre les 5 derniers). Quelles méthodes non supervisées vous paraissent à même de les distinguer ?*



La méthode des **k-means** semble être approprié pour notre cas.

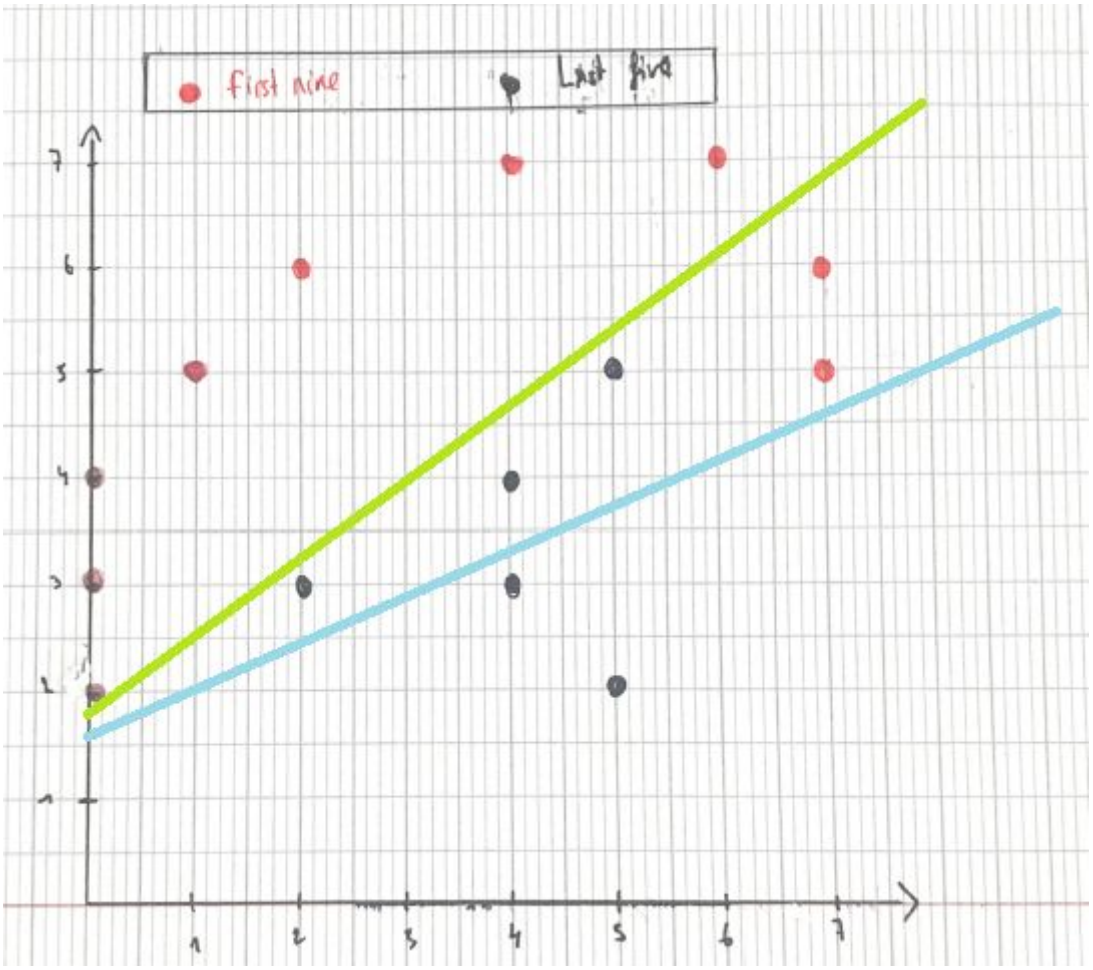
b) *Représentez le dendogramme d'une classication hiérarchique ascendante. N'oubliez pas de préciser quelle fonction vous avez choisie pour évaluer le coût de l'intégration (l'axe des ordonnées) mais pas besoin de faire les calculs exacts ; par contre l'ordre des associations doit être exact*

??

c) *Il s'avère que les deux composantes correspondent effectivement à deux classes qu'on appellera 1 (à l'intérieur) et -1 (à l'extérieur). Est-il possible de les séparer avec un SVM linéaire ?*

Il semble possible de les séparer avec un SVM linéaire, néanmoins il ne sera précis et on aura de la "perte" au niveau de la séparation

d) *Fixez deux niveaux de pénalisation arbitraires et dessinez approximativement les SVM linéaires optimaux associés*



Ici on peut voir que la droite vert pénalise points rouges et que la droite bleu pénalise les points noires.

Exercice 3

a) *Si, étant donné un tableau croisé d'effectifs, je voulais savoir si les deux variables concernées sont indépendantes ?*

Pour savoir si les deux variables concernées sont indépendantes, on utilise le **test du x^2**. Pour plus de détail sur le test, voici les trois étapes qui le compose :

1. On calcul les effectifs croisés espérés

$$E_{i,j} = \frac{1}{n} \sum \sharp\{X = i\} \sharp\{Y = j\}$$

 # = nombre d'événements

2. On calcul les effectifs croisés observés :

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

3. On peut aussi calculer l'écart relatif entre les deux variables :

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

Un T faible signifie que les variable sont indépendantes

b) Si je voulais trouver le nombre maximum de points pulvérisables à la surface d'une sphère ?

Pour trouver le nombre maximum de points pulvérisables à la surface d'une surface, il serait possible de calculer la dimension de Vapnik-Chervonenkis car cette dimension corresponds au nombre maximum d'éléments pulvérisable.

c) Si je voulais expliquer à une personne non-spécialiste la différence entre risque et ambiguité ?

L'ambiguïté se base sur le manque d'information. L'erreur, elle se calcule et se base sur l'information (sa véracité)