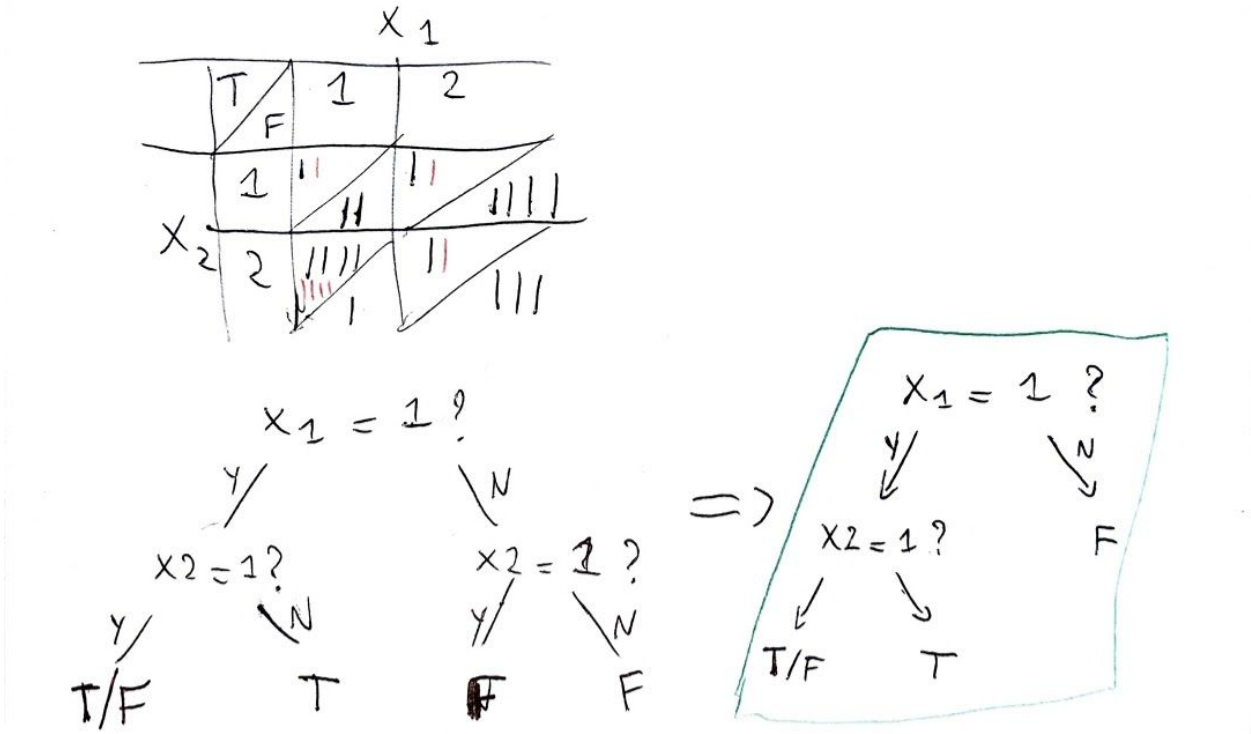
Exercice 1

a) Le risque empirique est égal à la somme des erreurs selon la fonction de coût choisie, divisée par le nombre n=17 d'éléments dans l'échantillon.
R_EST_1 = (8 * 1 + 1 * x) / 17
R_EST_2 = (1 * 1 + 3 * x) / 17
R_EST_1 > R_EST_2 => x > 7 / 2

b) x = 2
Pour prendre en compte le fait que les faux négatifs coûtent deux fois plus chers en terme d'erreur, j'ai doublé en rouge les barres de comptage de mon tableau d'effectif croisé.
X1 = 1 ? then (X2 = 1? then T/F else F) else F
Le risque empirique de ce modèle est de 6 (avec x=2).



c)

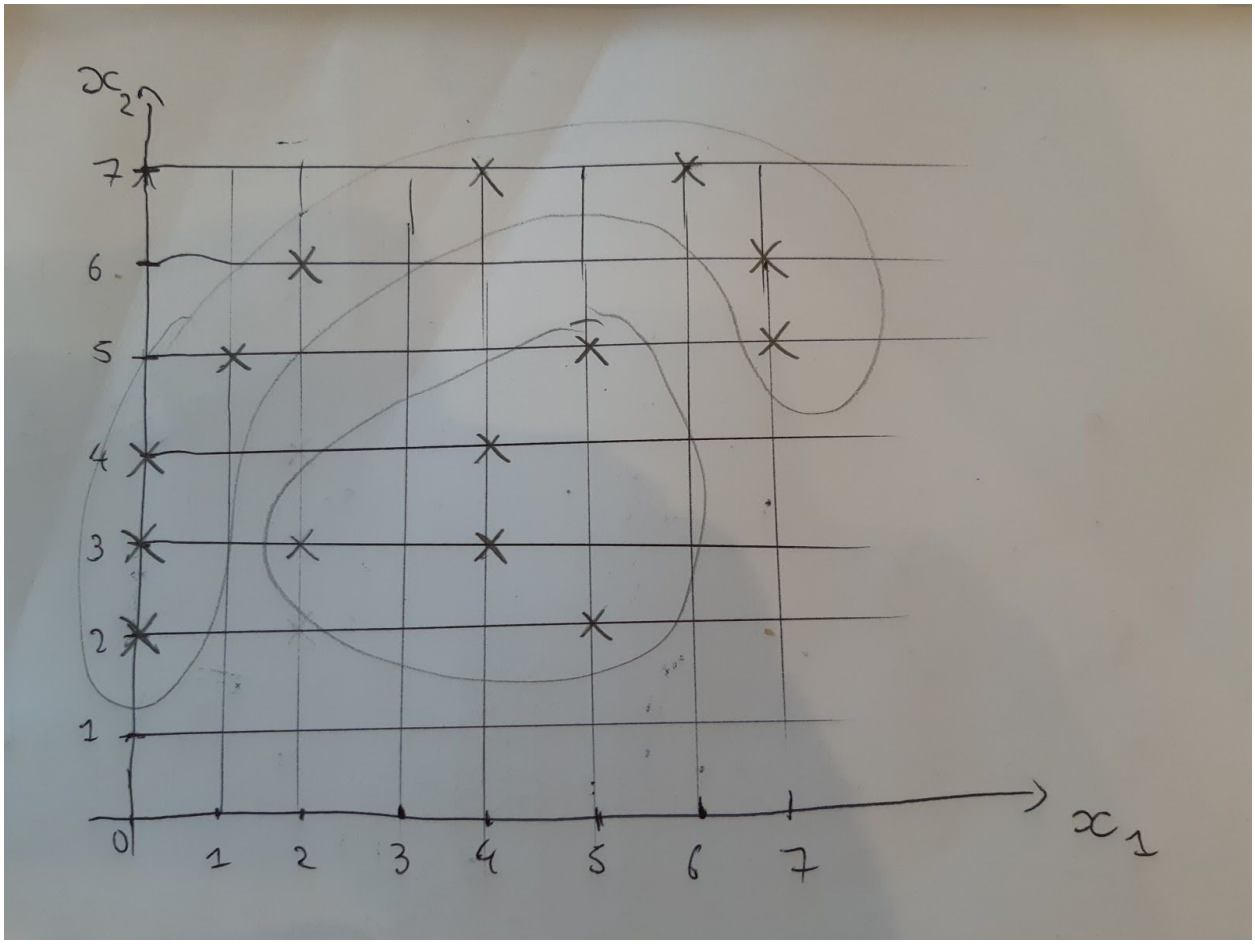
Y=True	1	2
X1	5	2
X2	4	5

Y=False	1	2
X1	3	7
X2	6	4

d) On compare P(Y=T ou F / X1 = a && X2 = b) = P(Y=...) * (pas eu le temps d'appliquer)

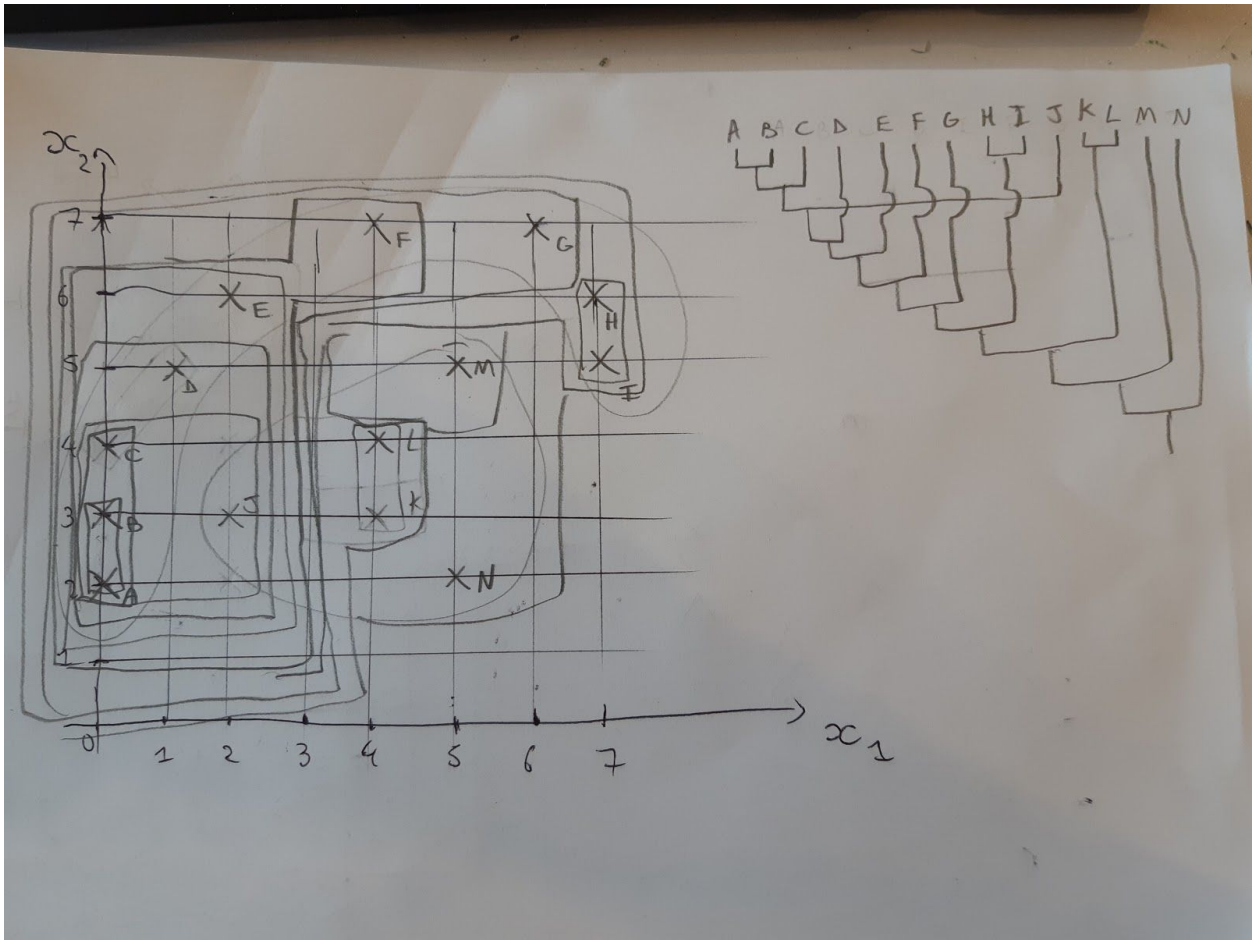
Exercice 2

a) Je n'utiliserais pas un kmeans (en tous cas pas sans avoir transformé l'espace auparavant). J'utiliserais plutôt un kNN avec k=3 (avec 2, plusieurs erreurs de classification m'apparaissent) et comme distance une distance euclidienne (pas de Manhattan, qui elle réduirait relativement l'écart entre les points du blob central et ceux de l'arc externe par rapport aux écarts entre les points de l'arc externe eux-mêmes, qui sont souvent placés en diagonal les uns des autres). Après avoir consulté le cours, je vois que les méthodes "SpectralClustering" et "Agglomerative Clustering" pourraient donner de très bons résultats. Enfin, les méthodes "Ward" et "DBSCAN" pourraient donner de bons résultats.



b) Pour me simplifier la vie, je vais utiliser une distance de Manhattan, et le critère de liaison minimum ("single-linkage clustering"). Après exécution, je me rend compte que

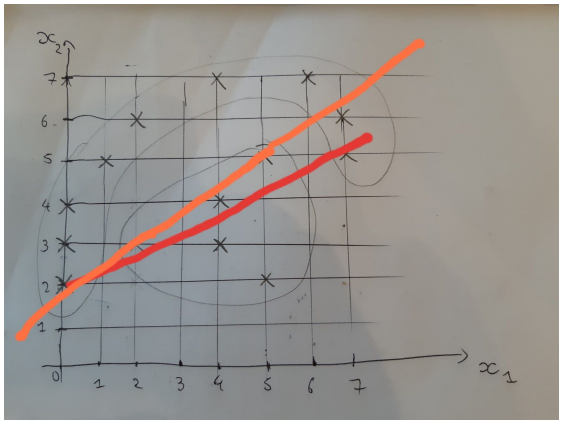
cette méthode donne de très mauvais résultats (ce à quoi je m'attendais, cf propos sur la distance de Manhattan dans la réponse précédente ; avec ici en prime le critère de liaison minimum qui rend l'algorithme encore plus "greedy").



c) Il est possible de les séparer avec un SVM linéaire, puisque la question suivante nous demande de le faire, mais si l'on n'opère pas de transformation sur l'espace (e.g kernel trick), il y aura des pertes (des mauvaises classifications) et la marge maximale sera nulle.

d) Droite rouge : si se tromper sur la classe externe (-1) est une erreur plus grave.

Droite orange : si se tromper sur la classe interne (1) est une erreur plus grave.



Exercice 3

a) Etant donné un tableau croisé d'effectifs, si je voulais savoir si les deux variables concernées sont indépendantes, alors à partir des effectifs croisés espérés E(i,j) et de ceux observés O(i,j), je calculerais leur écart relatif T ; si T suit une loi du khi carré, alors ces variables sont indépendantes. Sinon, elles dépendent l'une de l'autre.

b) Je ne comprends pas la question. Si c'était "pulvérisable par la surface d'une sphère", je lui apporterais la réponse qui suit. Trois points non alignés sont pulvérisables par une sphère (plus ils sont alignés, plus la sphère peut avoir besoin d'avoir un rayon gigantesque, si "l'adversaire sélectionne" les deux points opposés. A partir de 4 points, ils forment dans le meilleur des cas un polygone à 4 côtés, l'adversaire peut prendre les deux points les plus éloignés dans une classe, les autres dans l'autre classe … (pas le temps de terminer).

c) -