FTML - Partiel

Exercice 1

1-a)

On a:

- EST_1:
  - FP: 8
  - FN: 1
- EST_2:
  - FP: 1
  - FN: 3

Le risque empirique est calculé via la formule: 1*FP + x*FN

Ce qui veut dire que:

- x = 7/2 => R(EST_1) = R(EST_2)
- x < 7/2 => R(EST_1) > R(EST_2)
- x > 7/2 => R(EST_1) < R(EST_2)

Comme indiqué, on pose maintenant x = 2

1-b)

On a la répartition suivante `(X1, X2) -> T / F`:

- (1, 1) -> 1 / 2
- (1, 2) -> 4 / 1
- (2, 1) -> 1 / 4
- (2, 2) -> 1 / 3

Sachant que les FN comptent double, l'estimateur minimisant le risque empirique est donc le suivant:

- (1, 1) -> F
- (1, 2) -> T
- (2, 1) -> F
- (2, 2) -> F

1-c)

- Les données sont à peu près équilibrées (quantité des possibles X1 / X2 / Y)
  - Les coefficients du bayésien naïfs seront à peu près égaux

1-d)

Pour X1:

- | X1 | T | F |
- | 1 | 5 | 3 |
- | 2 | 2 | 7 |

Pour X2:

- | X2 | T | F |
- | 1 | 2 | 6 |
- | 2 | 5 | 4 |

1-e)

Estimateur bayésien naïf:

- `P(T | (X1=x1 & X2=x2))` = `P(T)` * `P(X1=x1 | T)` * `P(X2=x2 | T)`
- `P(F | (X1=x1 & X2=x2))` = `P(F)` * `P(X1=x1 | F)` * `P(X2=x2 | F)`
- Si `P(T | (X1=x1 & X2=x2))` > `P(F | (X1=x1 & X2=x2))` => T sinon => F

Avec:

- `P(T)` = 7 / 17
- `P(F)` = 10 / 17
- `P(X1=1 | T)` = 5 / 7
- `P(X1=2 | T)` = 2 / 7
- `P(X1=1 | F)` = 3 / 10
- `P(X1=2 | F)` = 7 / 10
- `P(X2=1 | T)` = 2 / 7
- `P(X2=2 | T)` = 5 / 7
- `P(X2=1 | F)` = 6 / 10
- `P(X2=2 | F)` = 4 / 10

En comparant avec l'estimateur trouvé plus haut:

- `P(T | (X1=1 & X2=1))` = `(7 * 5 * 2) / (17 * 7 * 7)` ~= 0.084

- `P(F | (X1=1 & X2=1))` = `(10 * 3 * 6) / (17 * 10 * 10)` ~= 0.106

    - Donc (1, 1) -> F, ce qui est le même résultat que l'estimateur précédent

- `P(T | (X1=1 & X2=2))` = `(7 * 5 * 5) / (17 * 7 * 7)` ~= 0.210

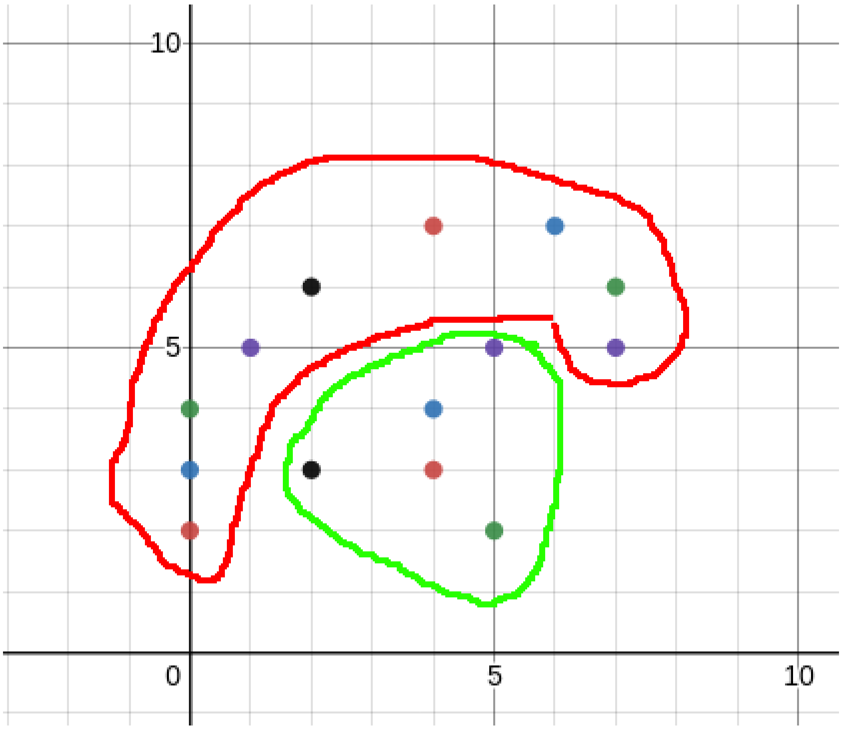
- `P(F | (X1=1 & X2=2))` = `(10 * 3 * 4) / (17 * 10 * 10)` ~= 0.071

    - Donc (1, 2) -> T, ce qui est le même résultat que l'estimateur précédent

---

Exercice 2

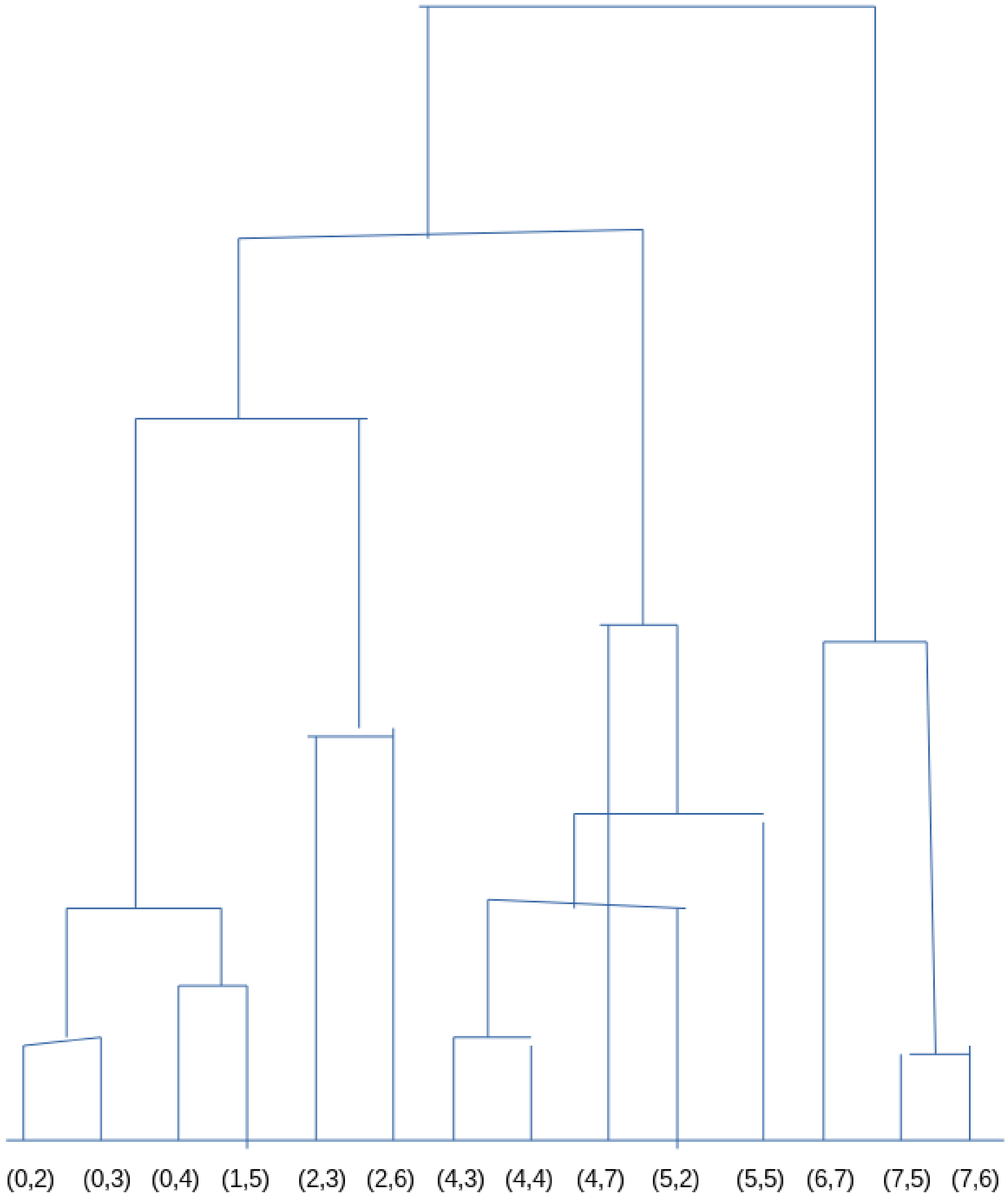
2-a)



Des méthodes non supervisées devraient pouvoir les distinguer, par exemple K-means, DBSCAN ou une classification hiérarchique

2-b)

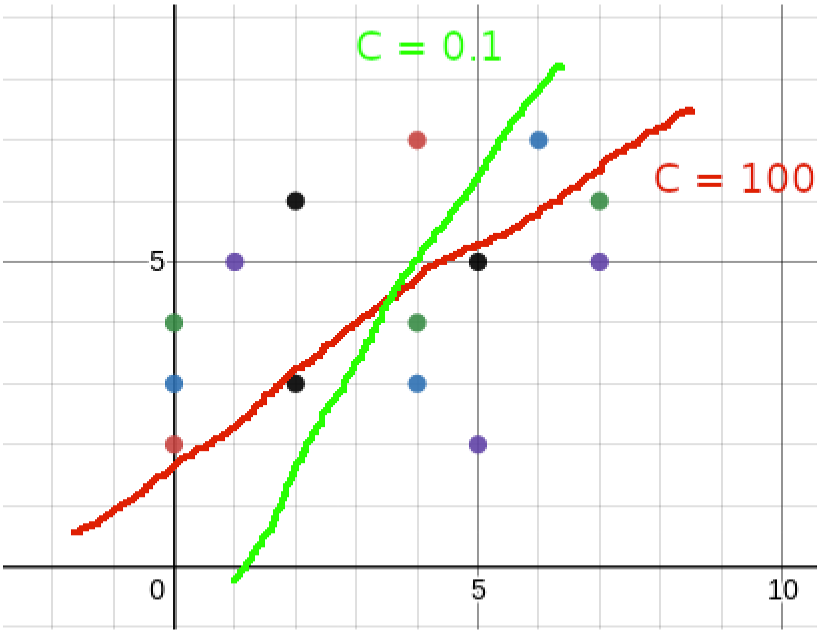
Dendogramme obtenu en prenant comme fonction de coût la distance entre les milieux des 2 groupes de points.



2-c)

Il n'est pas possible de les séparer avec un SVM linéaire car on ne peut tracer une droite séparant les 2 classes (le mieux possible mettrait 2 points de `-1` dans `1`)

2-d)



2-e)

En prenant le kernel suivant `f(x1,x2) = x2-5 + (x-5)^2 / 5`, les points des deux classes sont séparable par le prédicat: `f(x1,x2) > 0`

Exercice 3

3-a)

- Calcul du coefficient de corrélation `r` ou la `p-value`.

Avec le coefficient de corrélation:

- Si r ~= 0 => variables indépendantes
- Si r ~= -1 ou r ~= 1 => variables corrélées

Avec la p-value:

- Si p <= 0.05 => variables indépendantes
- Si p > 0.05 => variables corrélés

3-b)

TODO

3-c)

Le risque est un "taux" représentant à quel point un estimateur se trompe dans ses estimations.

Alors que l'ambiguité est le fait qu'un estimateur situe une donnée à la limite entre plusieurs classes, ce qui fait que de petits changements dans la donnée peut changer le résultat de la prédiction de l'estimateur.