Devoir 1 rapport théorique :

1.

a) $P(X|Y) = \dfrac{P(X,Y)}{P(Y)}$

b)

Soit X = " obtenir exactement 2 faces "

Y = " obtenir face au premier lancer "

$P(X|Y) = \frac{1}{5} \times \frac{4}{5} + \frac{4}{5} \times \frac{1}{5} = .32$

c)

i) $P(X,Y) = P(Y|X) \times P(X)$

ii) $P(X,Y) = P(X|Y) \times P(Y)$

d)

D'après c) $P(X,Y) = P(Y|X) \cdot P(X) = P(X|Y) \cdot P(Y)$

$$\Rightarrow \quad P(X|Y) = \frac{P(Y|X) \cdot P(X)}{P(Y)}$$

e)

i) $P(\text{McGill}) = 1 - .56 = .44$

ii)

$$P(\text{McGill} \mid \text{Bilingue}) = \frac{P(\text{McGill}, \text{Bilingue})}{P(\text{Bilingue})}$$

$$P(\text{McGill}, \text{Bilingue}) = P(\text{Bilingue} \mid \text{McGill}) \cdot P(\text{McGill})$$
$$= .4 \times .44 = .176$$

$$P(\text{Bilingue}) = P(\text{Bilingue}, \text{McGill}) + P(\text{Bilingue}, \text{UdeM})$$

$$= .176 + .7 \times .56 = .568$$

$$\Rightarrow P(\text{McGill} \mid \text{Bilingue}) = \frac{.176}{.568} = .31$$

2.

a) $P(\text{goal} \mid \text{politique}) = \frac{12}{1000} = .012$

b) Soit N = nombre fois où le mot «congrès» apparaît dans un document de 1000 mots sachant que le sujet est le sport

$N \sim \text{Binomial}(1000, 4/1000)$

on a : $E(N) = 1000 \times 4/1000 = 4$

c)

$$P(\text{goal}) = P(\text{goal}, \text{sport}) + P(\text{goal}, \text{politique})$$

$$P(\text{goal}, \text{sport}) = P(\text{goal} \mid \text{sport}) \times P(\text{sport})$$

$$= \frac{4}{100} \times \frac{1}{3} = .013$$

$$P(\text{goal}, \text{politique}) = P(\text{goal} \mid \text{politique}) \times P(\text{politique})$$

$$= \frac{12}{1000} \times \frac{2}{3} = .008$$

$$P(\text{goal}) = .013 + .008 = .021$$

d)

$$P(\text{sport} \mid \text{kick}) = \frac{P(\text{kick} \mid \text{sport}) \times P(\text{sport})}{P(\text{kick})}$$

$$P(\text{kick}) = P(\text{kick}, \text{sport}) + P(\text{kick}, \text{politique})$$

$$P(\text{kick}, \text{sport}) = \frac{15}{100} \times \frac{1}{3} = .05$$

$$P(\text{kick}, \text{politique}) = \frac{2}{1000} \times \frac{2}{3} = .0013$$

$$P(\text{kick}) = .0013 + .05 = .0513$$

$$P(sport \mid kick) = \frac{.05}{.0513} = .975$$

e)

Etant donné que les mots dans un document sont indépendants on a :

$$P(goal, congress) = P(goal, congress, sport) + P(goal, congress, politique)$$

$$= P(goal, congress \mid sport) \times P(sport)$$

$$+ P(goal, congress \mid politique) \times P(politique)$$

$$= P(goal \mid sport) \times P(congress \mid sport) \times P(sport)$$

$$+ P(goal \mid politique) \times P(congress \mid politique)$$

$$\times P(politique)$$

$$= \frac{4}{100} \times \frac{4}{1000} \times \frac{1}{3} + \frac{12}{1000} \times \frac{6}{100} \times \frac{2}{3}$$

$$= 5.33 \times 10^{-4}$$

$$P(\text{congress}) = P(\text{congress} \mid \text{sport}) \times P(\text{sport})$$

$$+ P(\text{congress} \mid \text{politique}) \times P(\text{politique})$$

$$= \frac{4}{1000} \times \frac{1}{3} + \frac{6}{100} \times \frac{2}{3} = .041$$

$$\Rightarrow P(\text{goal} \mid \text{congress}) = \frac{P(\text{goal}, \text{congress})}{P(\text{congress})}$$

$$P(\text{goal} \mid \text{congress}) = \frac{5.33 \times 10^{-4}}{.041} = .013$$

f)

Pour estimer par exemple $P(\text{goal} \mid \text{politique})$ on compte le nombre total de fois que le mot "goal" apparaît dans tous les documents de type "politique" puis on le divise par le nombre total de mots dans tous les documents de type "politique".

Pour estimer par exemple P(politique) on compte le nombre total de documents de type "politique" ensuite on le divise par N.

3.

a)

Soit $\underline{x} = (x_1, x_2, \ldots, x_n)$

on a : $f_\theta(\underline{x}) = f_\theta(x_1) \times f_\theta(x_2) \times \ldots \times f_\theta(x_n)$

b)

on a : $f_\theta(\underline{x}) = \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} \frac{1}{\theta} = \frac{1}{\theta^n}$

$$\frac{\partial}{\partial \theta} f_\theta(\underline{x}) = -n\theta^{-n-1} < 0 \quad \forall \theta \geq 0$$

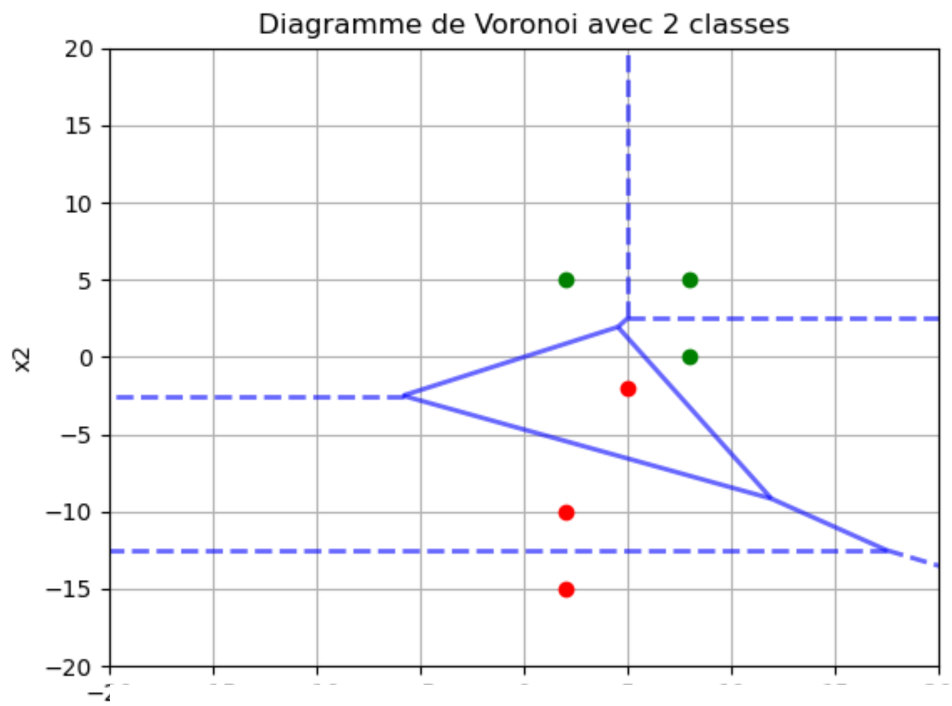
alors la vraisemblance est une fonction décroissante sur θ donc elle est maximale pour des petites valeurs de θ

Etant donné que θ doit toujours être plus grande que les observations on en déduit que la plus petite valeur que $\hat{\theta}$ peut prendre est le maximum des observations

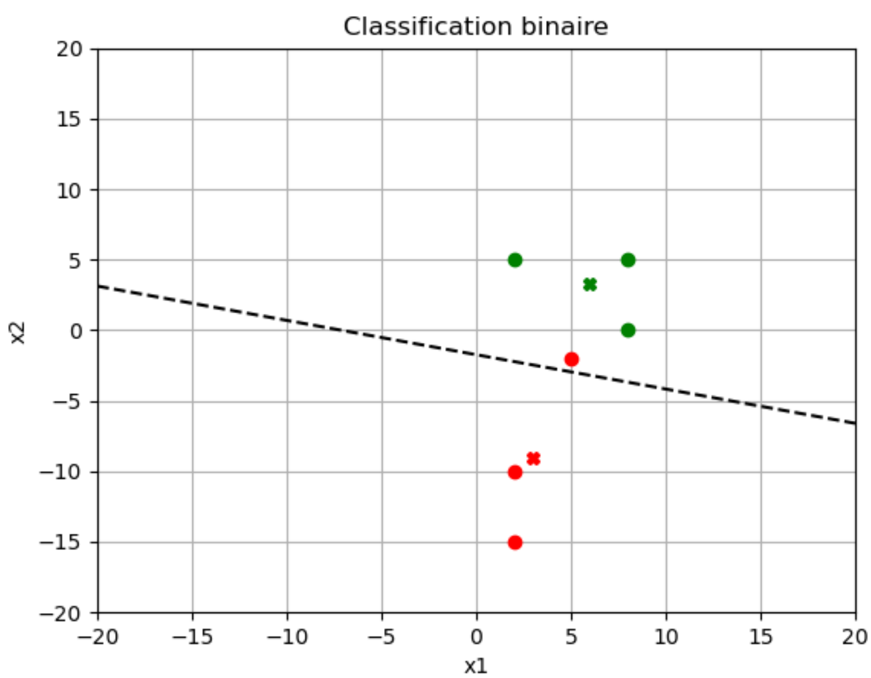
d'où $\hat{\theta} = \max\{x_1, x_2, \dots, x_n\}$

4.

a)



b)



c)

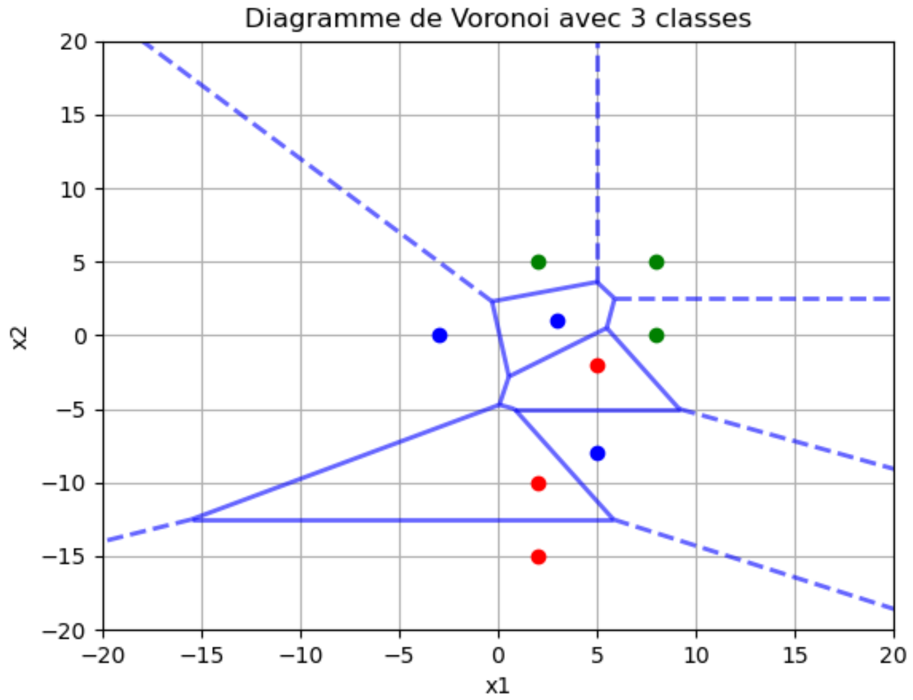
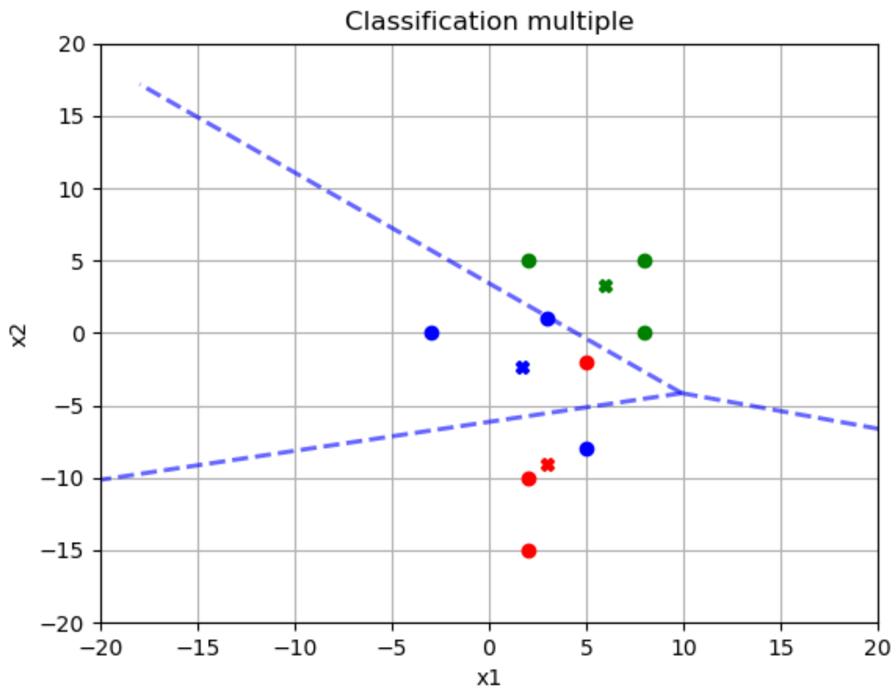


Diagramme de Voronoi avec 3 classes
x1
x2

d)


Classification multiple
x1
x2

5.

a)

on a : $\mathbb{1}_{\{x \in S\}} = \begin{cases} 1 & \text{si} \quad x \in S \\ 0 & \text{sinon} \end{cases}$

alors $E\left[\mathbb{1}_{\{x \in S\}}\right] = 1 \times P(x \in S) + 0 \times P(x \notin S)$

$\Rightarrow E\left[\mathbb{1}_{\{x \in S\}}\right] = P(x \in S)$

b)

Soit $Y_j = \mathbb{1}_{\{X_j \in V_i\}}$

on a : $E(Y_j) = E\left[\mathbb{1}_{\{X_j \in V_i\}}\right] = P(X_j \in V_i)$

$= \int_{V_i} f(x)\,dx$ pour X_j continue

Supposons que $\{x_j\}_{j=1}^{n}$ est iid alors les Y_i sont aussi iid en tant que fonctions de variables iid.

Soit $\bar{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$.

Par la loi des grands nombres $\bar{Y}_n$ converge en probabilité vers $E(Y_j) = \int_{V_i} f(x)\,dx$.

Le résultat devient immédiat en remarquant que $\hat{p}(x_j \in V_i) = \bar{Y}_n$

c)

Nombre total de regions $= 2^{464}$

Nombre de chiffres $= \lfloor \log_{10}(2^{464}) \rfloor + 1 = 140$

d)

plus petit nombre $= 16 \times 4 \times 2^{464}$
$= 3.05 \times 10^{141}$

e)

Soit R une certaine region

n

on a : $P(R \text{ soit vide}) = \left(1 - \frac{1}{m^d}\right)^n$

Sachant que chaque dimension est subdivisé en d regions, nous avons m^d regions au total en d dimensions.
Si les points sont uniformément répartis sur toutes les regions alors tous les points ont la même chance d'appartenir à une region qui est égal à

$$\frac{1}{\# \text{total de regions}} = \frac{1}{m^d}$$

Donc $P(R \text{ soit vide}) = P(\text{tous les points} \notin \bar{a}\ R)$

$$= \prod_{i=1}^{n} P(\text{point } i \notin R) = \prod_{i=1}^{n}\left(1 - \frac{1}{m^d}\right) = \left(1 - \frac{1}{m^d}\right)^n$$