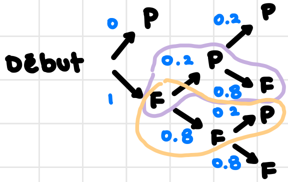
## Q1 - Rappels de probabilités : probabilité conditionnelle et règle de Bayes

(a) $\boxed{P(X|Y) = \dfrac{P(X \cap y)}{P(y)}}$

(b) $P(\text{face}) = 4/5 = 0.8$

$P(\text{pile}) = 1/5 = 0.2$




| évènement | probabilité |
|---|---|
| F - P - F | $1 \times 0.2 \times 0.8 = 0.16$ |
| F - F - P | $1 \times 0.8 \times 0.2 = 0.16$ |
| | + $\boxed{0.32}$ |

(c) Expressions équivalentes de $P(X,y)$

(i) $P(y|X) = \dfrac{P(X,y)}{P(X)}$

$\Leftrightarrow$ $P(X,y) = P(y|X) \cdot P(X)$

(ii) $P(X|y) = \dfrac{P(X,y)}{P(y)}$

$\Leftrightarrow$ $P(X,y) = P(X|y) \cdot P(y)$

(d) Théorème de Bayes

Selon la définition de la probabilité conditionnelle, on a :

$$P(X|Y) = \frac{P(X \cap y)}{P(y)} \quad \text{et} \quad P(y|X) = \frac{P(X \cap y)}{P(X)}$$

En réarrangeant la deuxième équation, on a :

$$P(X \cap y) = P(y|X) \cdot P(X)$$

En substituant ce dernier dans la première équation, on a:

$$P(x|y) = \frac{P(y|x) \cdot P(x)}{P(y)}$$

Ainsi, on obtient la formule du théorème de Bayes:

$$P(x|y) = \frac{P(y|x) \cdot P(x)}{P(y)} \quad \blacksquare$$

(e) $P(\text{UdeM}) = 0.56$

(i) $P(\text{McGill}) = 1 - P(\text{UdeM})$
$= 1 - 0.56$
$= 0.44$

(ii) $P(\text{UdeM} \mid \text{bilingue}) = 0.7$
$P(\text{McGill} \mid \text{bilingue}) = 0.4$

1 $P(\text{bilingue}) = P(\text{bilingue} \mid \text{UdeM}) \cdot P(\text{UdeM}) + P(\text{bilingue} \mid \text{McGill}) \cdot P(\text{McGill})$
$= 0.7 \cdot 0.56 + 0.44 \cdot 0.4$
$= 0.568$

2 $P(\text{McGill} \mid \text{bilingue}) = \frac{P(\text{bilingue} \mid \text{McGill}) \cdot P(\text{McGill})}{P(\text{bilingue})}$
$= \frac{0.44 \cdot 0.4}{0.568}$
$= 0.309859$
$\approx 0.31$

## Q2 - Bag of words & modèle de sujet unique

$P(sport) = 1/3$
$P(politique) = 2/3$

(a) $P(mot = goal \mid sujet = politique) = 12/1000$

(b) $E = P(mot = congress \mid sujet = sport) \cdot \text{nb total de mots}$
$= 4/1000 \cdot 1000$
$= 4$

(c) $P(mot = goal) = P(mot = goal \mid sujet = sport) \cdot P(sport) + P(mot = goal \mid sujet = politique) \cdot P(politique)$
$= 4/100 \cdot 1/3 + 12/1000 \cdot 2/3$
$= 8/375$
$\approx 0.02$

(d) ① $P(mot = kick) = P(mot = kick \mid sujet = sport) \cdot P(sport) + P(mot = kick \mid sujet = politique) \cdot P(politique)$
$= 15/100 \cdot 1/3 + 2/1000 \cdot 2/3$
$= 0.051\bar{3}$

② $$P(sujet = sport \mid mot = kick) = \frac{P(mot = kick \mid sujet = sport) \cdot P(sujet = sport)}{P(mot = kick)}$$
$$= \frac{15/100 \cdot 1/3}{0.0513}$$
$= 0.974689$
$\approx 0.97$

(e) ① $P(mot = congress) = P(mot = congress \mid sujet = sport) \cdot P(sujet = sport) + P(mot = congress \mid sujet = politique) \cdot P(sujet = congress)$
$= 4/1000 \cdot 1/3 + 6/100 \cdot 2/3$
$= 31/750$

② $P(A) = P(\text{sujet} = \text{sport} \mid \text{mot} = \text{congress})$

$$= \frac{P(\text{mot} = \text{congress} \mid \text{sujet} = \text{sport}) \cdot P(\text{sujet} = \text{sport})}{P(\text{mot} = \text{congress})}$$

$$= \frac{4/1000 \cdot 1/3}{31/750}$$

$$= 1/31$$

③ $P(A') = P(\text{sujet} = \text{politique} \mid \text{mot} = \text{congress})$

$= 1 - P(A)$

$= 1 - 1/31$

$= 30/31$

④ $P(\text{goal} \mid \text{congress}) = P(\text{mot} = \text{goal} \mid \text{sujet} = \text{sport}) \cdot P(\text{sujet} = \text{sport} \mid \text{mot} = \text{congress}) + P(\text{mot} = \text{goal} \mid \text{sujet} = \text{politique}) \cdot P(\text{sujet} = \text{politique} \mid \text{mot} = \text{congress})$

$= 4/100 \cdot 1/31 + 12/1000 \cdot 30/31$

$= 2/155$

$\simeq 0.013$

(f) **Estimation des probabilités des sujets**

Il suffit de compter le nombre de documents associés à chaque sujet. Prenons $N$ comme le nombre total de documents :

- $N_{sport}$ : le nombre de documents de sport.
- $N_{politique}$ : le nombre de documents de politique.

Alors la probabilité de chaque sujet :

- $P(\text{sujet} = \text{sport}) = \dfrac{N_{sport}}{N}$
- $P(\text{sujet} = \text{politique}) = \dfrac{N_{politique}}{N}$

## Estimation des probabilités conditionnelles

Premièrement, il faut compter le nombre d'occurrences de chaque mot dans les documents de sport et de politique afin de construire une table de fréquence pour chaque mot et chaque sujet.

Ensuite, on peut estimer les probabilités à partir des fréquences.

- $C_{goal, sport}$ : le nombre de fois que le mot goal apparaît dans les documents de sport.
- $C_{total, sport}$ : le nombre total de mots dans les documents de sport.

Alors, la probabilité conditionnelle pour que le mot soit goal sachant que le sujet était sport :

$$\boxed{P(mot = goal \mid sujet = sport) = \frac{C_{goal, sport}}{C_{total, sport}}}$$

## Q3 - Estimateur du maximum de vraisemblance

(a) $f_\theta(x_1, x_2, \ldots, x_n) = f_\theta(x_1) \cdot f(x_2) \cdot \ldots \cdot f_\theta(x_n)$

$$= \boxed{\prod_{i=1}^{n} f_\theta(x_i)}$$

(b) $f_\theta(x_1, x_2, \ldots, x_n) = f_\theta(x_1) \cdot f(x_2) \cdot \ldots \cdot f_\theta(x_n)$

$$= \prod_{i=1}^{n} f_\theta(x_i)$$

$$= \prod_{i=1}^{n} \frac{1}{\theta} \quad \text{pour} \quad x_i = ]0, \theta]$$

$$= \begin{cases} \left(\frac{1}{\theta}\right)^n & \text{si } 0 < x_i \leq \theta \\ 0 & \text{sinon} \end{cases}$$

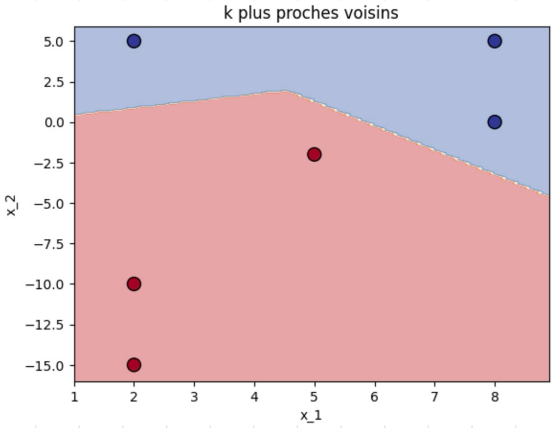
$$= \begin{cases} \frac{1}{\theta^n} & \text{pour } \theta \geq \max(x_1, x_2, \ldots, x_n) \\ 0 & \text{sinon} \end{cases}$$

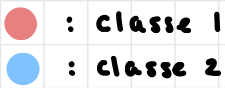
La valeur de $\frac{1}{\theta^n}$ est inversement proportionnelle à $\theta$. Ainsi, puisque l'on veut maximiser notre fonction, on cherche à avoir le plus petit $\theta$ tout en respectant que $\theta \geq \max(x_1, x_2, \ldots, x_n)$. L'estimateur du maximum de vraisemblance $\theta$ est donc égal à $\max(x_1, \ldots, x_n)$.

■

# Q4 - k plus proches voisins

(a)



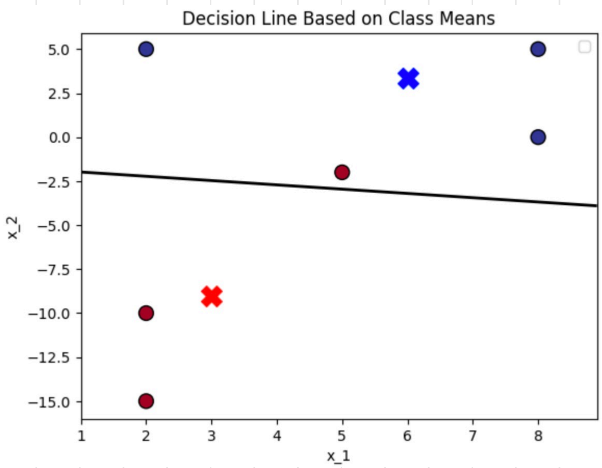

(b) ① Moyenne classe 1 = $(3, -9)$

$$x_1 = \frac{5+2+2}{3} = 3 \qquad x_2 = \frac{-2+(-10)+(-15)}{3} = -9$$

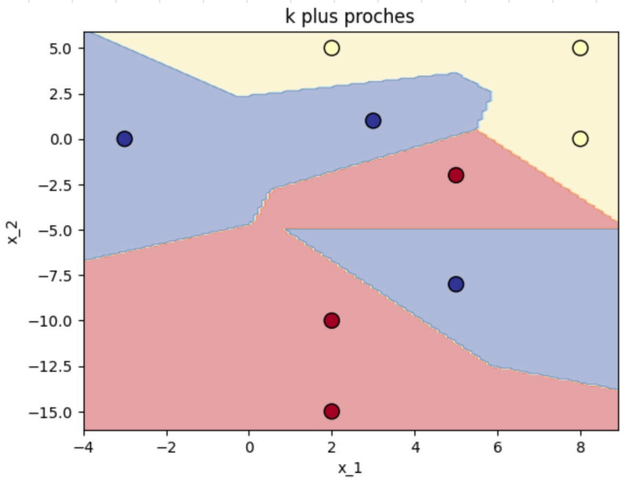
② Moyenne classe 2 = $(6, 3.\overline{3})$

$$x_1 = \frac{8+2+8}{3} = 6 \qquad x_2 = \frac{0+5+5}{3} = 3.\overline{3}$$

③



(c)




: classe 1

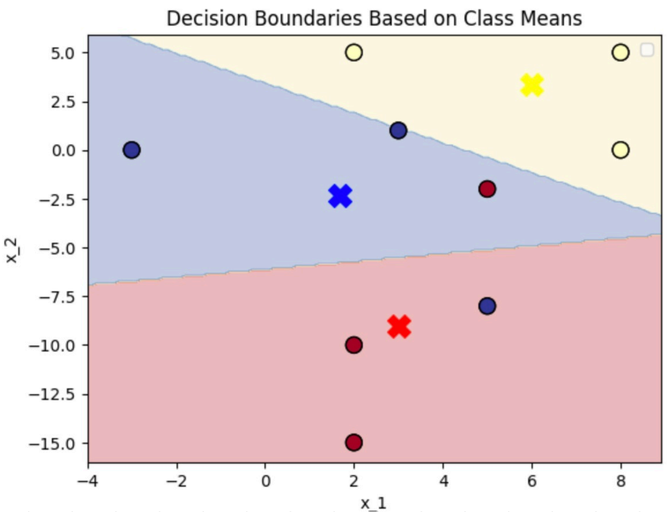
: classe 2

: classe 3

**Moyenne classe 3 = $(1.\bar{6}, -2.\bar{3})$**

$$x_1 = \frac{3 + (-3) + 5}{3} = 1.\bar{6}$$

$$x_2 = \frac{1 + 0 + (-8)}{3} = -2.\bar{3}$$



## Q5 - Méthodes à base d'histogrammes

(a) $1_{\{x \in S\}} = \begin{cases} 1, & \text{si } x \in S \\ 0, & \text{sinon} \end{cases}$

Ainsi, $\mathbb{E}[1_{\{x \in S\}}] = 1 \cdot P(x \in S) + 0 \cdot (1 - P(x \in S))$
$= P(x \in S)$

$\blacksquare$

(b) $p_i = P(x \in V_i)$
$= \int_{V_i} f(x)\, dx$

où $p_i$ est la vraie probabilité d'être dans la région $i$, $V_i$ est le volume de la région, et $f(x)$ est la fonction PDF inconnue

$\hat{p}_i = \frac{k_i}{n}$

où $\hat{p}_i$ est la probabilité empirique d'être dans la région $i$, $k_i$ est le nb de points dans la région $i$ et $n$ est le nb total de points

Ainsi, par la loi des grands nombres, on a :

$$\frac{k_i}{n} \to P(x \in V_i) = \int_{V_i} f(x)\, dx = p_i \quad \text{lorsque } n \to \infty$$

En effet, lorsque $n \to \infty$, la probabilité empirique $\hat{p}$ s'approche de la vraie probabilité $p_i$.

$\blacksquare$

(c) $d = 464$
$m = 2$

nb de régions $= m^d$
$= 2^{464}$

(d) précision = 90%
$\epsilon$ = 10%

① 90% - 10% = 80%

② (80%) ÷ (5%) = 16 (précision augmente en 16 intervalles de 5%)

③ 4 × 16 = 64 nouveaux points / région

(e) nb région = $m^d$

P( point région ) = $\frac{1}{m^d}$

P(point région vide ) = $\boxed{\left(1 - \frac{1}{m^d}\right)^n}$