Chapitre 2. Classifieur bayésien naïf

Claude Petit, Insee et université de Rennes - claude.petit@univ-rennes.fr

Juin 2026

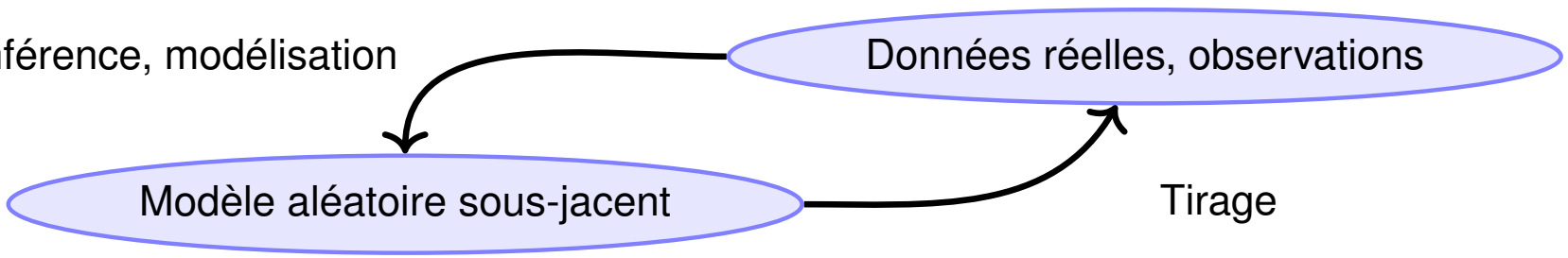
Overview

Modèle statistique

Définitions

- Expérience aléatoire $\iff$ espace probabilisé $\iff$ $(\mathbb{X}, \mathfrak{F}, \mathbb{P})$.
  - $\mathbb{X}$ ensemble des issues possibles de l'expérience aléatoire.
  - $\mathfrak{F}$ tribu sur $\mathbb{X}$ (évènements de l'expérience).
  - $\mathbb{P}$ mesure de probabilité sur $\mathfrak{F}$.
- Modèle statistique : famille d'expériences aléatoires, $\hat{\mathrm{m}}$ $\mathbb{X}$ et $\hat{\mathrm{m}}$ $\mathfrak{F}$: $(\mathbb{X}, \mathfrak{F}, \mathscr{P})$.
- Observation x : réalisation d'une variable aléatoire $X \sim \mathbb{P}$.
- Objectif de l'inférence statistique : déterminer $\mathbb{P} \in \mathscr{P}$ sachant x.

Savoir distinguer $(\mathbb{X}, \mathfrak{F}, \mathbb{P})$ et $(\mathbb{X}, \mathfrak{F}, \mathscr{P})$!



Modèle paramétrique

Definition

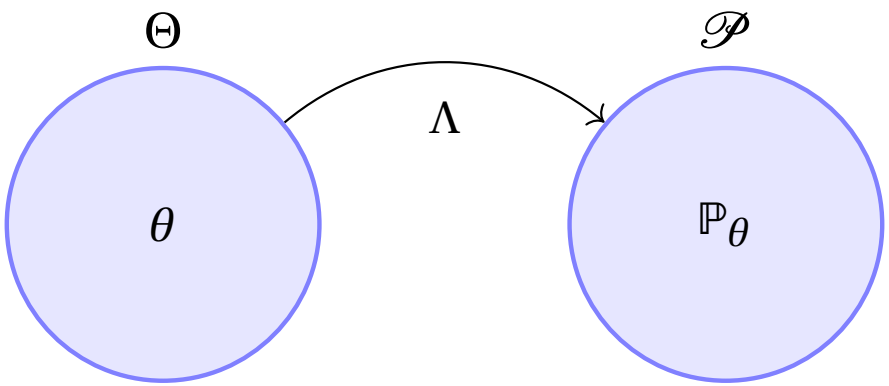
Modèle statistique $(\mathbb{X}, \mathfrak{F}, \mathscr{P})$ paramétré si les éléments de $\mathscr{P}$ peuvent être décrit par un paramètre : c.-à-d. il existe Λ surjection :

$$\Lambda : \Theta \longrightarrow \mathscr{P} \tag{1}$$

$$\theta \mapsto \mathbb{P}_\theta \tag{2}$$

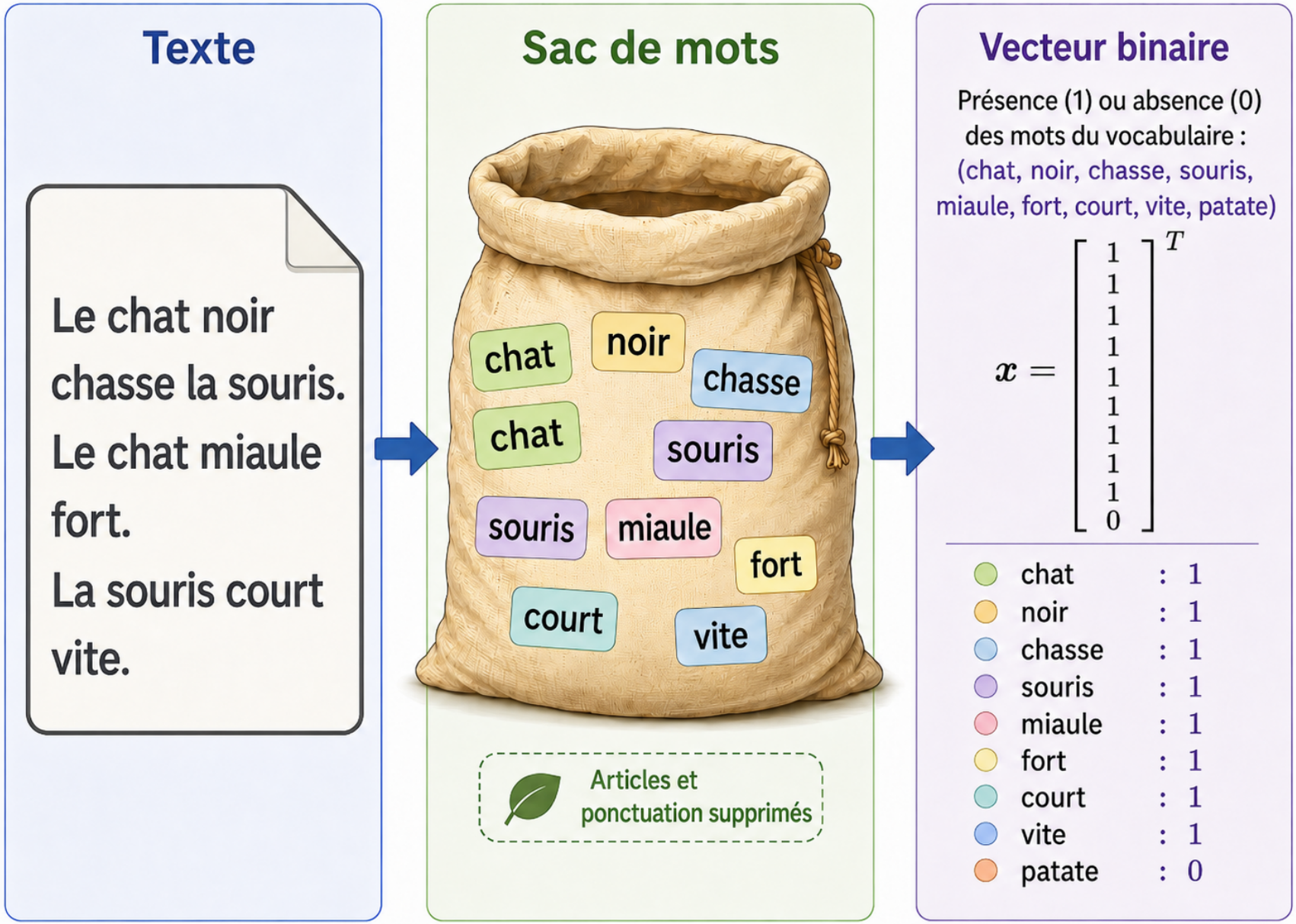
$\mathscr{P} = \{\mathbb{P}_\theta, \theta \in \Theta\}$. Si Λ bijection, on dit que le modèle est identifiable.

Modèle paramétrique si $\Theta \subset \mathbb{R}^p$ pour $p \in \mathbb{N}$. Sinon, non-paramétrique.

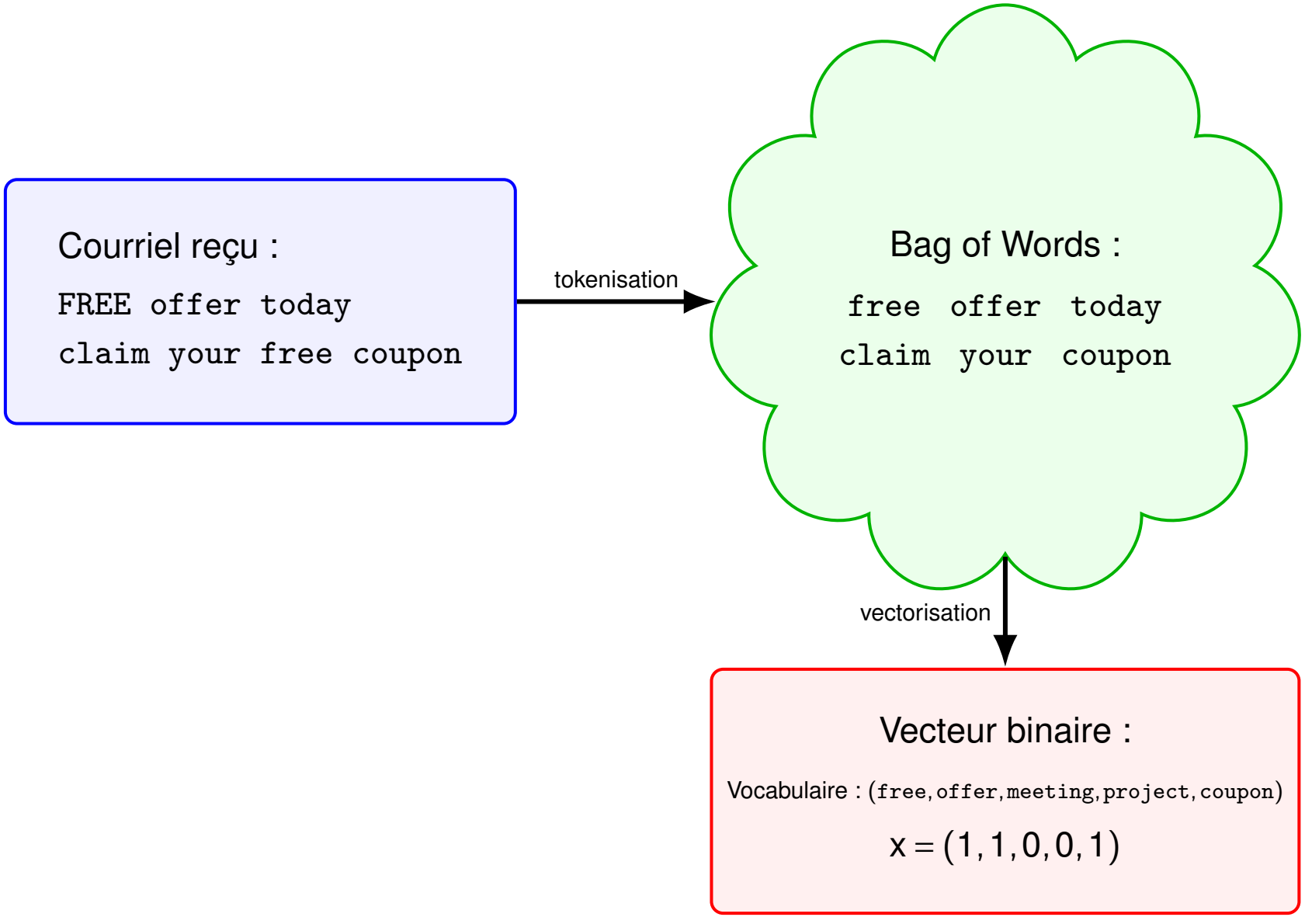


Un premier exemple traditionnel : la détection de spam

Détection de spam -1- : sac de mots



Détection de spam -2- : traitement d'un courriel

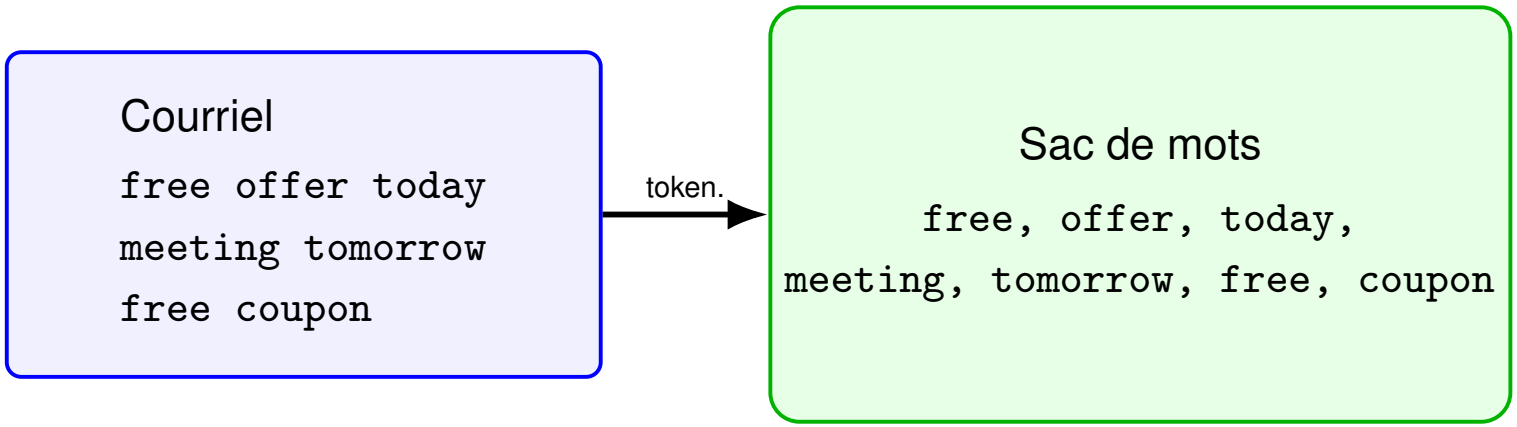


Courriel

```
free offer today
meeting tomorrow
free coupon
```

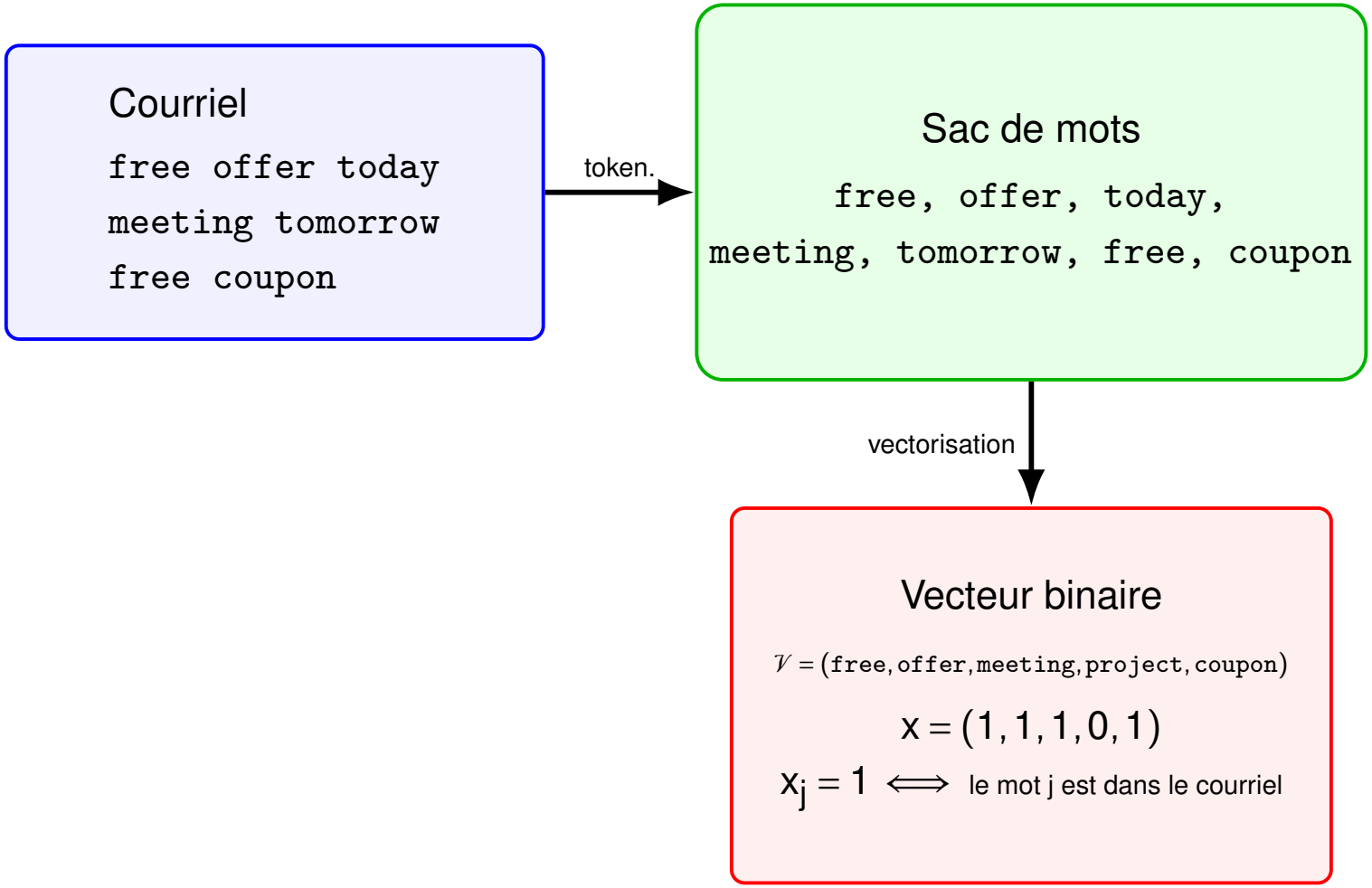
- Étape 1. On part du message brut.

Détection de spam -3- : traitement d'un courriel



- Étape 2. L'ordre local des mots est oublié : on conserve les occurrences.

Détection de spam -3- : traitement d'un courriel



• Étape 3. Après projection sur le vocabulaire, on code la présence/absence de chaque terme.

Détection de spam - 2 : traitement d'un courriel

- On définit le vocabulaire commun à l'ensemble des courriels :

$$\mathcal{V} = \{\texttt{free}, \texttt{offer}, \texttt{meeting}, \texttt{project}, \texttt{coupon}\}.$$

- Pour chaque vecteur, on veut décider s'il s'agit d'un spam ou non :

Document	Vecteur binaire	Classe
free offer coupon	(1,1,0,0,1)	Spam
free offer	(1,1,0,0,0)	Spam
meeting project	(0,0,1,1,0)	Non spam
project meeting	(0,0,1,1,0)	Non spam
free meeting	(1,0,1,0,0)	Non spam
free	(1,0,0,0,0)	Spam ?

- Comment décider ? À partir d'un échantillon de courriels déjà reçus,
  - Classer les spams et les autres.
  - Pour chaque vecteur, compter le nombre d'apparitions dans les spams.
  - Peut-on décider qu'un vecteur caractérise un spam si on le trouve plus souvent dans les spams que dans les non spams ?

Détection de spam -4- : traitement d'un courriel

- Peut-on décider qu'un vecteur caractérise un spam si on le trouve plus souvent dans les spams que dans les non spams ?

Document	Vecteur binaire	Classe
free offer coupon	(1,1,0,0,1)	Spam
free offer	(1,1,0,0,0)	Spam
meeting project	(0,0,1,1,0)	Non spam
project meeting	(0,0,1,1,0)	Non spam
free	(1,0,0,0,0)	?

- Non, car cela dépend du contexte.
- Non, car un courriel peut ne pas contenir pas tous les mots du vecteur.
- Non, car le nombre de vecteurs augmente exponentiellement avec la taille.
- Il faut travailler directement sur les statistiques des mots.

⇒ nécessité d'un cadre probabiliste.

Cadre probabiliste

Formalisation statistique -1-

- $\mathscr{D}_n = \{Z_1, \ldots, Z_n\}$ n-échantillon de v.a.i.i.d. $Z_i = (X_i, Y_i)$.
- X_i observations issues d'une v.a. $X \in \{0,1\}^d$: données, variables explicatives, « features ». Ici : vecteurs binaires du contenu d'un courriel.
- $X_{ij} = 1 \iff$ le mot j est présent dans le courriel i.
- Y_i issues d'une v.a. Y, catégories des X_i : étiquettes ou labels.
- $Y_i = 1 \iff$ le courriel est un spam.
- $X \in \mathbb{X} = \{0,1\}^d$, $Y \in \mathbb{Y} = \{0,1\}$.
- $\mathbb{P}_\theta$ proba sur $\mathscr{E} = \mathbb{X} \times \mathbb{Y}$: loi inconnue de (X, Y) et (X_i, Y_i).
- θ paramètre inconnu.
- $h \in \mathscr{F} = \mathscr{F}(\mathbb{X}, \mathbb{Y})$ fonction de prédiction (classifieur, régresseur) : $h(x) = y$.

Objectif de l'apprentissage supervisé : déterminer Y sachant X à partir des seules observations de $Z_1, \ldots, Z_n$

Formalisation statistique -2-

• Pour un courriel de vecteur x, on cherche l'étiquette $y = 0, 1$ la plus vraisemblable.

• On calcule $\mathbb{P}[Y = 1|X = x]$ et si $\mathbb{P}[Y = 1|X = x] > \mathbb{P}[Y = 0|X = x]$, on décide que c'est un spam.

• Ce sont des probabilités conditionnelles... Formule de Bayes :

$$\begin{cases} \mathbb{P}[Y = 1|X = x] & = \dfrac{\mathbb{P}[X = x|Y = 1]\ \mathbb{P}[Y = 1]}{\mathbb{P}[X = x]} \\ \mathbb{P}[Y = 0|X = x] & = \dfrac{\mathbb{P}[X = x|Y = 0]\ \mathbb{P}[Y = 0]}{\mathbb{P}[X = x]} \end{cases} \tag{3}$$

• Pour comparer les deux probabilités, le dénominateur ne sert à rien. On cherche donc :

$$\mathbb{P}[Y = 1|X = x] \sim \mathbb{P}[X = x|Y = 1]\ \mathbb{P}[Y = 1]. \tag{4}$$

Rappel : la formule de Bayes

$$\overbrace{\mathbb{P}(A|B)}^{\textit{a posteriori}} = \frac{\overbrace{\mathbb{P}(B|A)}^{\text{vraisemblance}} \overbrace{\mathbb{P}(A)}^{\textit{a priori}}}{\underbrace{\mathbb{P}(B)}_{\text{marginale}}}$$

Formalisation statistique -3-

$$\mathbb{P}[X = x|Y = 1] \times \mathbb{P}[Y = 1] \; ? \tag{5}$$

• Calcul de $\mathbb{P}[Y = 1]$? La probabilité *a priori* de recevoir un spam peut être approchée par la fréquence de spams dans l'échantillon d'entrainement.

• Calcul de $\mathbb{P}[X = x|Y = 1]$? C'est la probabilité *a posteriori* que le vecteur de mots X soit égal à $x = (x_1, \ldots, x_d)$ sachant que le courriel est un spam. On peut l'estimer aussi à partir de l'échantillon d'entraînement.

• Le vecteur caractéristique x représente le courriel et x_i indique si le mot i est présent ou non dans le courriel (il y a d mots dans le vocabulaire $\mathcal{V}$) :

$$[X = x] = [X_1 = x_1, X_2 = x_2, \ldots, X_d = x_d] = \bigcap_{j=1}^{d} [X_j = x_j] \tag{6}$$

Formalisation statistique -4-

$$\mathbb{P}[X = x|Y = 1] = \mathbb{P}[X_1 = x_1, X_2 = x_2, \ldots, X_d = x_d|Y = 1] \quad (7)$$

- 😱 Il existe 2^d vecteurs (courriels) possibles. Pour en estimer la fréquence de chacun d'eux, il faudrait donc un échantillon avec beaucoup plus de 2^d courriels.
- Mais si les X_i sont indépendants conditionnellement à Y,

$$\mathbb{P}[X = x|Y = 1] = \mathbb{P}[X_1 = x_1, X_2 = x_2, \ldots, X_d = x_d|Y = 1] \quad (8)$$

$$= \prod_{j=1}^{d} \mathbb{P}[X_j = x_j|Y = 1] \quad (9)$$

- Il suffit alors de la connaissance de n probabilités conditionnelles pour l'estimation.

Rappel : indépendance conditionnelle -1-

- Deux évènements A et B sont indépendants si

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B) \tag{10}$$

- Deux évènements A et B sont indépendants conditionnellement à un évènement C si

$$\mathbb{P}(A \cap B|C) = \mathbb{P}(A|C) \times \mathbb{P}(B|C) \tag{11}$$

- Ces deux notions sont.... indépendantes ! Indépendance n'implique pas indépendance conditionnelle et vice versa (hélas).

 Indépendance conditionnelle : notion terriblement dangereuse.

Rappel : indépendance conditionnelle -2-

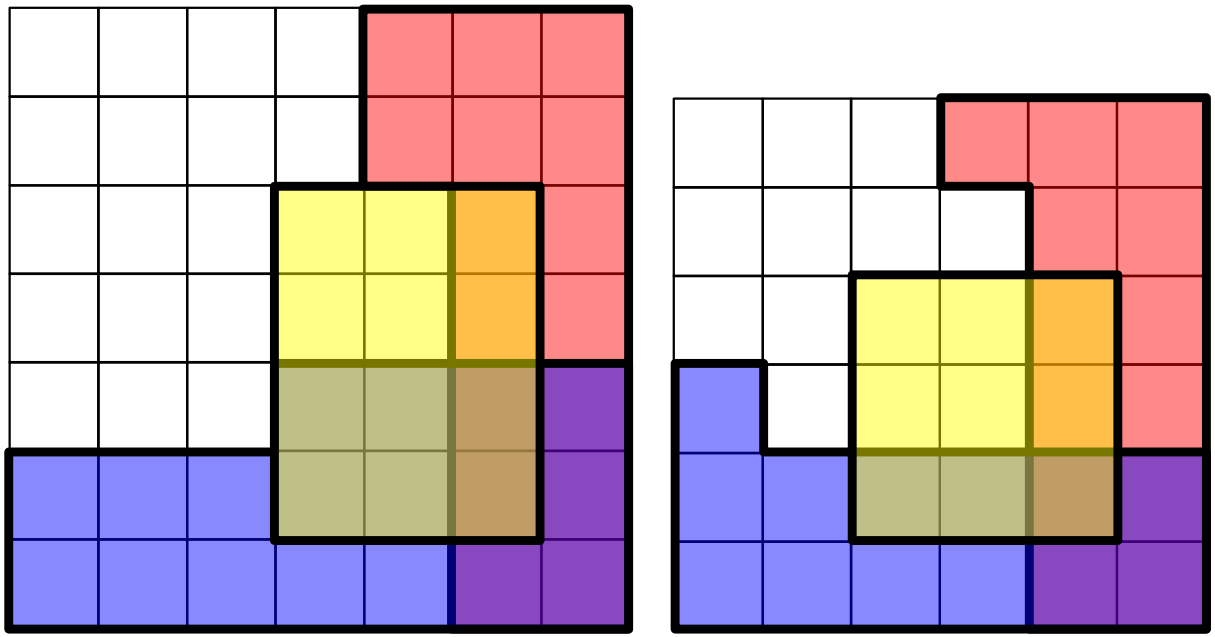


Figure 1 – $\mathbb{P}(A \cap B | C) = \mathbb{P}(A|C)\mathbb{P}(B|C)$ mais $\mathbb{P}(A \cap B | \overline{C}) \neq \mathbb{P}(A|\overline{C})\mathbb{P}(B|\overline{C})$. La figure provient de wikipedia.

Rappel : indépendance conditionnelle -3- : $A \perp\!\!\!\perp B \mid C \not\Rightarrow A \perp\!\!\!\perp B$

On dispose de 2 pièces, l'une équilibrée, l'autre avec deux faces « face ». On en choisit une au hasard de façon équitable. Soit C l'évènement « je choisis la pièce équilibrée ». $\mathbb{P}(C) = 1/2$.

On lance deux fois de suite cette pièce. Soit A (resp. B) l'évènement le premier lancer est « pile » (resp. le second lancer est « face »).

$$\mathbb{P}(A \cap B|C) = \frac{1}{2} \times \frac{1}{2} = \mathbb{P}(A|C)\mathbb{P}(B|C) \tag{12}$$

Mais (formule des probabilités totales)

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \cap B|C)\mathbb{P}(C) + \mathbb{P}(A \cap B|\overline{C})\mathbb{P}(\overline{C}) = \frac{1}{4} \times \frac{1}{2} + 0 \times \frac{1}{2} = \frac{1}{8}$$

$$\mathbb{P}(A) = \mathbb{P}(A|C)\mathbb{P}(C) + \mathbb{P}(A|\overline{C})\mathbb{P}(\overline{C}) = \frac{1}{4} + 0$$

$$\mathbb{P}(B) = \mathbb{P}(B|C)\mathbb{P}(C) + \mathbb{P}(B|\overline{C})\mathbb{P}(\overline{C}) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4}$$

$$\mathbb{P}(A)\mathbb{P}(B) = \frac{3}{16} \neq \frac{2}{16}$$

Donc A et B ne sont pas indépendants.

Rappel : indépendance conditionnelle -4- : $A \perp\!\!\!\perp B \not\Rightarrow A \perp\!\!\!\perp B \mid C$

On lance deux dés équilibrés indépendants (⚅ , ⚀). Soit X et Y les résultats des lancers (ce sont des v.a. et plus des évènements). Soit

$$Z = \mathbb{1}_{[X=Y]}$$

Sans conditionnement, les deux dés sont indépendants : $X \perp\!\!\!\perp Y$

Mais sachant $Z = 1$ (les deux dés sont égaux), si $X =$ ⚂, alors nécessairement $Y =$ ⚂ donc X et Y ne sont pas indépendants conditionnellement à Z.

- Cause commune Z : X et Y peuvent être $\perp\!\!\!\perp$ conditionnellement à Z.
- Effet commun Z : conditionner par Z peut créer une dépendance.

Revenons à nos courriels et à nos spams : « Idiot Bayes »

$$\mathbb{P}[X = x|Y = 1] = \mathbb{P}[X_1 = x_1, X_2 = x_2, \ldots, X_d = x_d|Y = 1] \tag{13}$$

$$= \prod_{j=1}^{d} \mathbb{P}[X_j = x_j|Y = 1] \tag{14}$$

- Dans la vraie vie, les mots clefs sont-ils indépendants conditionnellement au fait qu'un courriel est un spam ?
- Pas du tout.....
- ... mais ça marche bien quand même... et sinon on ne peut pas le calculer.
- Hypothèse « naïve » : indépendance conditionnelle de X sachant Y.

Formalisation statistique -5- : estimations *a priori*

- C'est la partie « entraînement » de la méthode sur l'échantillon $\mathscr{D}_n = \{(x_i, y_i), i = 1, \ldots, n\}$.

- x_i est un courriel composé des mots $j = 1, \ldots, d$ pour lesquels $x_{ij} = 1$ et y_i vaut 1 ou 0 selon que x_i est un spam ou pas. 1 : spam / 0 : pas spam.

- Soit n_1 le nombre de spams parmi les n courriels y_i de l'échantillon. On estime $\theta_1 = \mathbb{P}[Y = 1]$ par $\widehat{\theta}_1 = \widehat{\mathbb{P}}[Y = 1] = n_1/n$.

- Pour chaque mot j du vocabulaire ($j = 1, \ldots, d$), on calcule le nombre n_{j1} de spams dans lequel il se trouve (dans l'échantillon). On estime $\theta_{j1} = \mathbb{P}[X_j = 1 | Y = 1]$ par $\widehat{\theta}_{j1} = \widehat{\mathbb{P}}[X_j = 1 | Y = 1] = n_{j1}/n_1$.

$$n_1 = \sum_{i=1}^{n} \mathbb{1}_{[y_i = 1]} \quad : \text{ Nombre de spams.}$$

$$n_{j1} = \sum_{i=1}^{n} \mathbb{1}_{[y_i = 1]} \mathbb{1}_{[x_{ij} = 1]} \quad : \text{ Nombre de spams contenant le mot } j.$$

- Finalement,

$$\begin{aligned}
\mathbb{P}[Y=1|X=x] &\approx \mathbb{P}[X=x|Y=1]\,\mathbb{P}[Y=1] \ \star \\
&= \mathbb{P}[X_1=x_1, X_2=x_2, \ldots, X_n=x_n|Y=1]\,\mathbb{P}[Y=1] \\
&\approx \mathbb{P}[Y=1]\prod_{j=1}^{d}\mathbb{P}[X_j=x_j|Y=1] \ \star \\
&\approx \frac{n_1}{n}\prod_{j=1}^{d}\frac{n_{j1}}{n_1} = \widehat{\theta}_1\prod_{j=1}^{d}\widehat{\theta}_{j1} \ \star \\
\mathbb{P}[Y=0|X=x] &\approx \mathbb{P}[X=x|Y=0]\,\mathbb{P}[Y=0] \\
&\approx \frac{n-n_1}{n}\prod_{j=1}^{d}\frac{n_1-n_{j1}}{n-n_1} = (1-\widehat{\theta}_1)\prod_{j=1}^{d}(1-\widehat{\theta}_{j1})
\end{aligned}$$

- Si $\widehat{\mathbb{P}}[Y=1|X=x] > 1/2$ le nouveau courriel est (classé comme) un spam.
-  Attention ! Il s'agit d'une (série d') approximation(s) : $\star\star\star$

- Si un mot x_j n'apparaît pas dans les spams de l'échantillon, alors $n_{j1} = 0$ et tout le produit est nul :

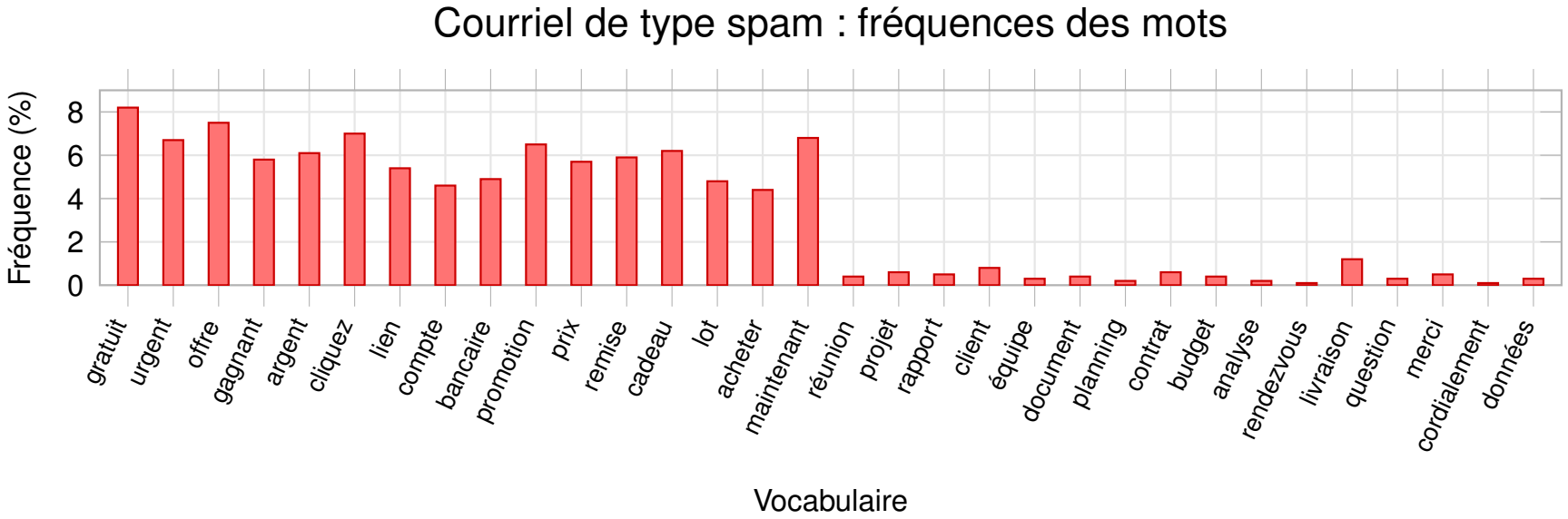
$$\widehat{\mathbb{P}}[Y = 1 | X = x] = \frac{n_1}{n} \prod_{j=1}^{d} \frac{n_{j1}}{n_1} = 0.$$

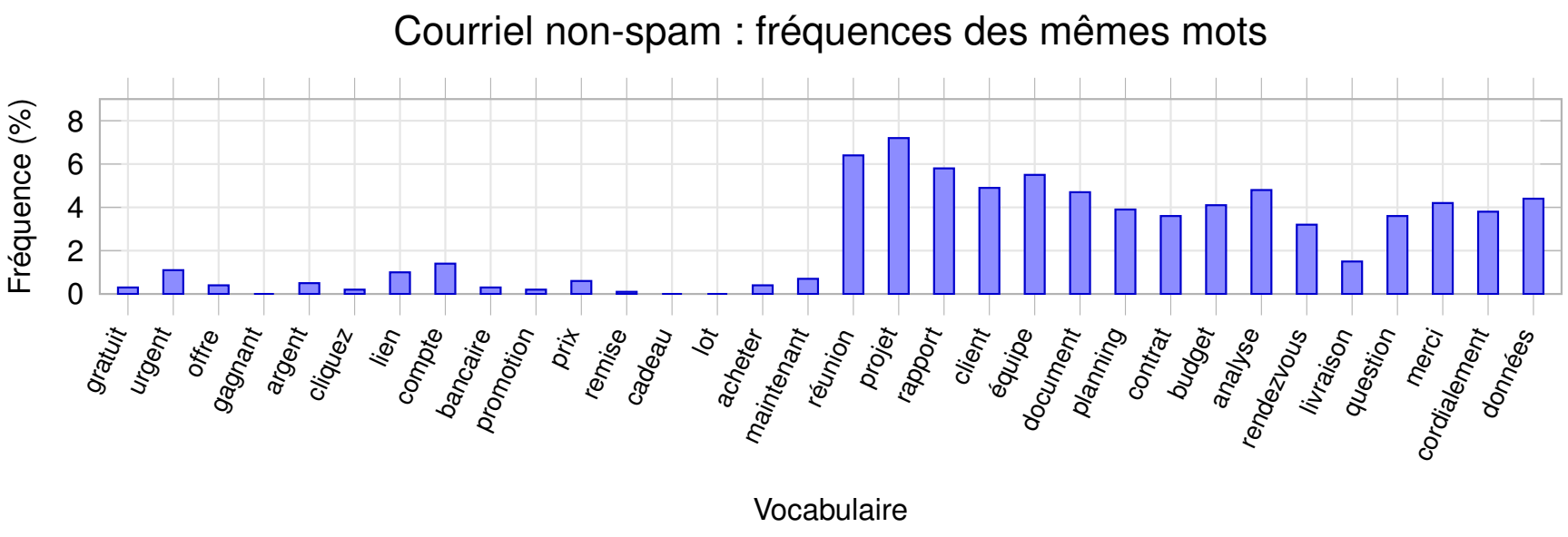
- Tous les nouveaux messages contenant ce mot auront une probabilité nulle d'être des spams, ce qui n'est pas réaliste.
- Pour éviter cela, on utilise le lissage de Laplace :

$$\widehat{\mathbb{P}}[Y = 1 | X = x] = \frac{n_1}{n} \prod_{j=1}^{d} \frac{n_{j1} + 1}{n_1 + 2}.$$

- Pourquoi $+2$ au dénominateur ?

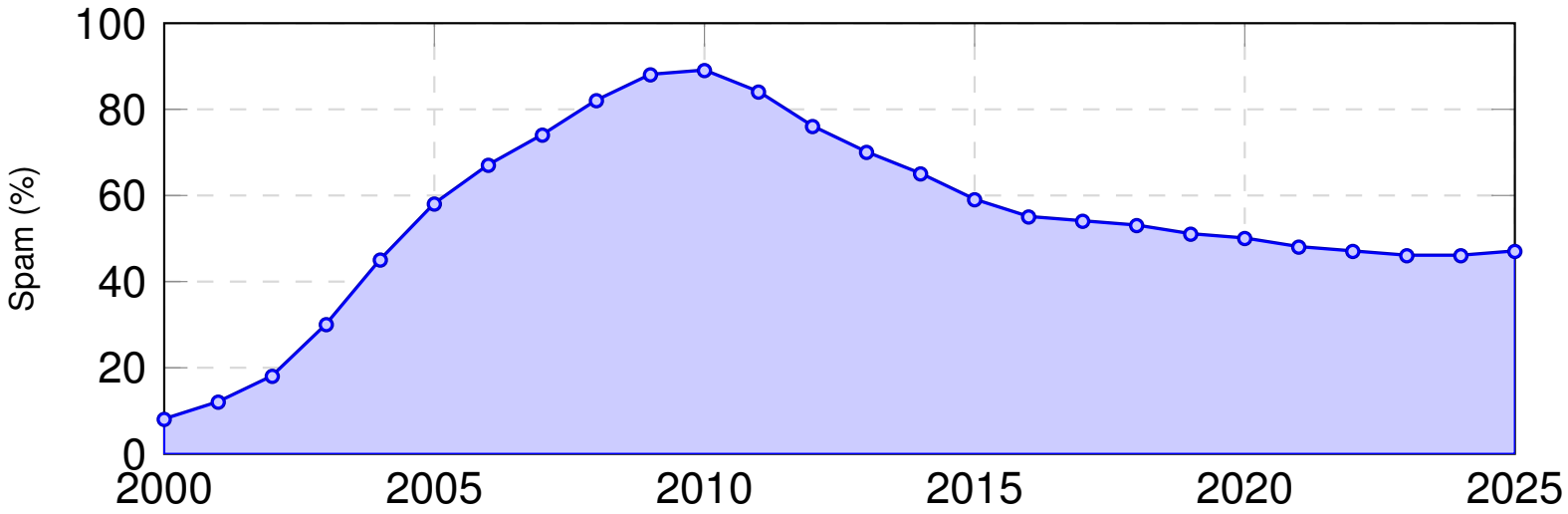
Les spams « in real life » -1-



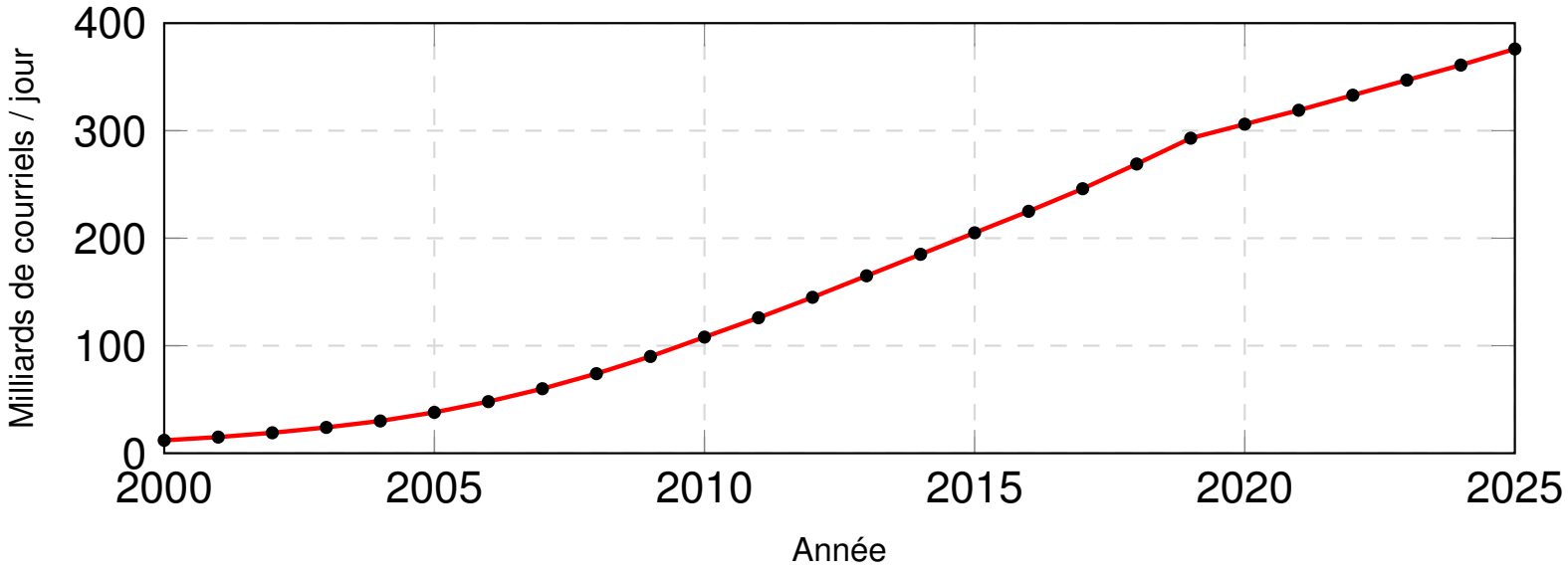


Les spams « in real life » -2-

Pourcentage de spams parmi les courriels envoyés (2000–2025)



Nombre de courriels envoyés dans le monde (2000–2025)



Naive Bayes : cas général discret

Classifieur bayésien naïf : cas général

- $\mathscr{D}_n = \{Z_1, \ldots, Z_n\}$ n-échantillon de v.a.i.i.d. $Z_i = (X_i, Y_i)$ pour $i = 1, \ldots, n$.
- $X_i = (X_{ij})_j \in \mathbb{N}^d$ observations issues d'un v.a. $X \in \mathbb{N}^d$: données, variables explicatives, « features ».
- Y_i issues d'une v.a. Y : étiquettes ou labels. $Y = y_1, \ldots, y_K$.
- $\mathbb{P}_\theta$, loi de (X, Y) inconnue, caractérisée par θ.
- $\theta_k = \mathbb{P}[Y = y_k]$, $k = 1, \ldots, K$: loi *a priori* de Y (inconnue).
- $\theta_{jk} = \mathbb{P}[X_j = x_j | Y = y_k]$, $j = 1, \ldots, d$: loi conditionnelle de X sachant Y.
- Le classifieur bayésien naïf est alors :

$$h_{NB}(x) = \underset{y=y_1,\ldots,y_K}{\arg\max} \left(\mathbb{P}[Y = y_k] \prod_{j=1}^{d} \mathbb{P}[X_j = x_j | Y = y_k] \right) \tag{15}$$

$$= \underset{y=y_1,\ldots,y_K}{\arg\max} \left(\theta_k \prod_{j=1}^{d} \theta_{jk} \right) \tag{16}$$

Exemple : classifieur de Bayes multinomial

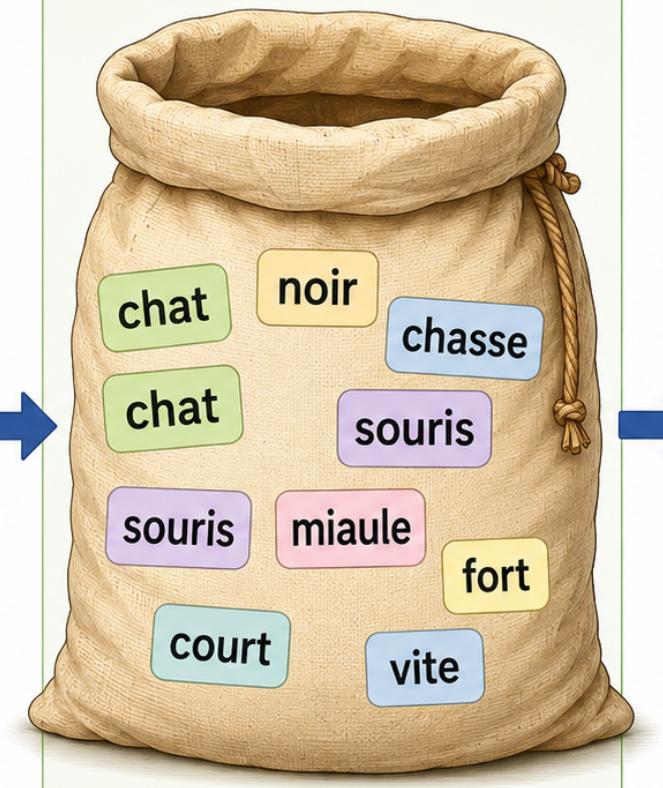
- Documents (textes de longueur m mots) appartenant à K catégories.
- Si $K = 2$ on retrouve les spams / non spams.
- Le vocabulaire de mots est $\mathcal{V} = \{w_1, \ldots, w_d\}$.
- X_j : nombre d'occurences du mot w_j dans le texte X.
- Ici, $x_j \in \mathbb{N}$ et $y = y_1, \ldots, y_K$.
- $X = (X_1, \ldots, X_d)$ document contenant $\sum_{j=1}^{d} X_j = m$ mots.
- Les mots sont indépendants conditionnellement à $[Y = y]$.
- La loi de X sachant $[Y = y_k]$ est multinomiale de paramètres $m, \theta_{1k}, \ldots \theta_{dk}$.

$$\mathbb{P}[X = x | Y = y_k] = \frac{m!}{x_1! \ldots x_d!} \prod_{j=1}^{d} \theta_{jk}^{x_j} \tag{17}$$

Texte

Le chat noir chasse la souris. Le chat miaule fort. La souris court vite.

Sac de mots



Articles et ponctuation supprimés

Vecteur de comptages

Ordre du vocabulaire : (chat, noir, chasse, souris, miaule, fort, court, vite)

$$x = \begin{bmatrix} 2 \\ 1 \\ 1 \\ 2 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}^T$$

- chat : 2
- noir : 1
- chasse : 1
- souris : 2
- miaule : 1
- fort : 1
- court : 1
- vite : 1

Rappels sur la loi multinomiale

- Loi d'un vecteur $X = (X_1, .., X_d)$ d'entiers.
- On dispose d'une urne avec des boules de d couleurs.
- On tire de façon indépendante et avec remise m boules.
- X_j : nombre de boules de couleur j parmi les m tirées.
- θ_{jk} : probabilité de tirer une boule de couleur j (on oublie k pour l'instant).
- La loi est donnée par :

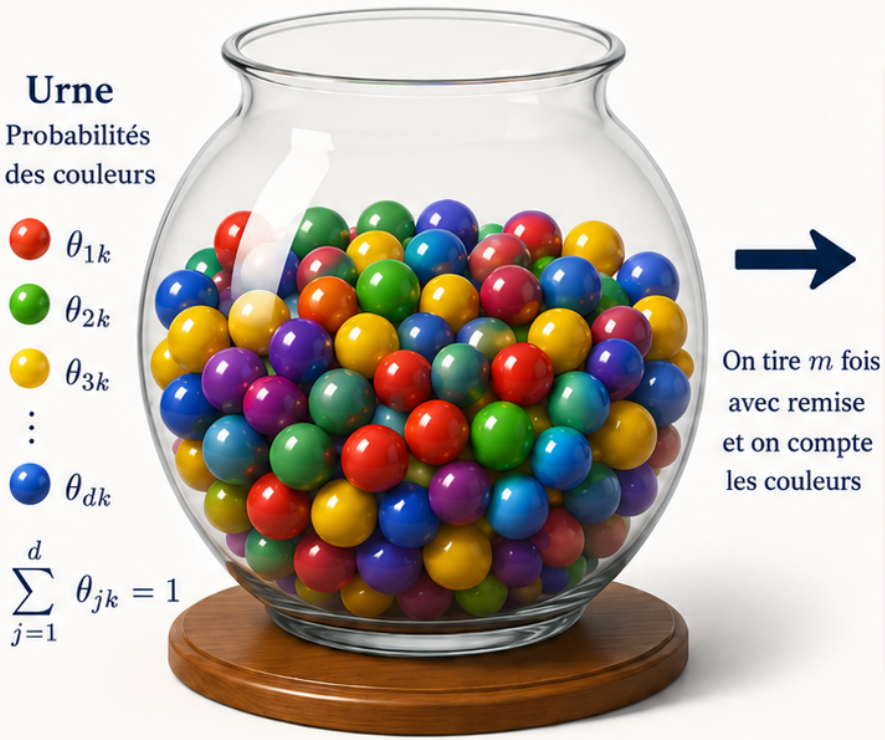
$$\mathbb{P}[X = x | Y = y_k] = \mathbb{P}[X_1 = x_1, \ldots, X_d = x_d | Y = y_k] = \frac{m!}{x_1! \ldots x_d!} \prod_{j=1}^{d} \theta_{jk}^{x_j} \qquad (18)$$

- La moyenne du vecteur X est

$$\mathbb{E}[X] = \left(m\theta_{j1}, \ldots, m\theta_{jd}\right) \qquad (19)$$

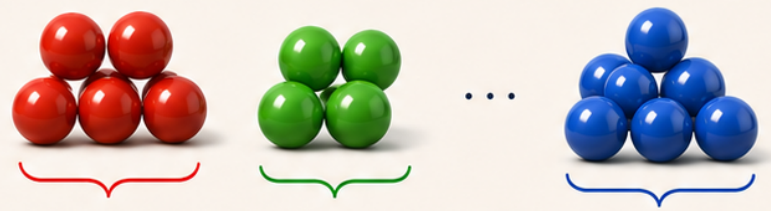
Loi multinomiale

On effectue m tirages indépendants avec remise dans une urne contenant des boules de d couleurs.



On tire m fois avec remise et on compte les couleurs

Résultat d'un tirage de m boules



x_1 boules rouges — x_2 boules vertes — x_d boules bleues

Nombre total de boules : m

Le vecteur des comptages $X = (X_1, X_2, \ldots, X_d)$ suit une loi multinomiale de paramètres m et $(\theta_{1k}, \theta_{2k}, \ldots, \theta_{dk})$.

$$X \sim \mathcal{M}(m, \theta_{1k}, \theta_{2k}, \ldots, \theta_{dk})$$

avec $\sum_{j=1}^{d} X_j = m$ et $X_j \geq 0$.

$$\mathbb{P}(X_1 = x_1, \ldots, X_d = x_d) = \frac{m!}{x_1!\, x_2! \cdots x_d!}\, \theta_{1k}^{x_1}\, \theta_{2k}^{x_2} \cdots \theta_{dk}^{x_d}.$$

si $\sum_{j=1}^{d} x_j = m$ et $x_j \geq 0$.

Estimation des paramètres multinomiaux

$$\mathbb{P}[X = x|Y = y_k] = \mathbb{P}[X_1 = x_1, \ldots, X_d = x_d|Y = y_k] = \frac{m!}{x_1!\ldots x_d!} \prod_{j=1}^{d} \theta_{jk}^{x_j} \tag{20}$$

- La moyenne du vecteur X est

$$\mathbb{E}[X] = (m\theta_{1k}, \ldots, m\theta_{dk}) \tag{21}$$

(22)

- On va estimer les probabilités par

$$\widehat{\theta}_{jk} = \frac{n_{jk}}{n_k}$$

$n_k = \sum_{i=1}^{n} \mathbb{1}_{[Y_i = y_k]}$, nombre de documents de catégorie k.

$n_{jk} = \sum_{i=1}^{n} x_{jk} \mathbb{1}_{[Y_i = y_k]}$, nombre d'occurrences du mot j dans la catégorie k.

Estimateur bayésien naïf pour le modèle multinomial

- Le coefficient multinomial ne dépend pas de la classe y_k, on peut donc l'oublier.

- Le classifieur bayésien naïf est alors :

$$h_{NB}(x) = \underset{y=y_1,\dots,y_K}{\arg\max} \left(\mathbb{P}[Y = y_k] \prod_{j=1}^{d} \mathbb{P}[X_j = x_j | Y = y_k] \right) \tag{23}$$

$$\widehat{h}_{NB}(x) = \underset{k=1,\dots,K}{\arg\max} \left(\widehat{\theta}_k \prod_{j=1}^{d} \widehat{\theta}_{jk}^{x_j} \right) \tag{24}$$

$$= \underset{k=1,\dots,K}{\arg\max} \left(\frac{n_k}{n} \prod_{j=1}^{d} \left(\frac{n_{jk}}{n_k} \right)^{x_j} \right) \tag{25}$$

Deux remarques sur le modèle multinomial

- On peut utiliser un lissage pour éviter des probabilités nulles :

$$\widehat{\theta}_{jk} = \frac{n_{jk} + 1}{n_k + d}$$

- Dans un vecteur X de loi multinomiale, les coordonnées X_j ne sont pas indépendantes, alors que l'hypothèse naïve parle d'indépendance ? ? ? ?

⇒ L'hypothèse d'indépendance ne porte pas sur les coordonnées X_j mais sur le tirage des mots individuels.

- La « naïveté » vient du fait qu'on ignore l'ordre des mots, la syntaxe, et les dépendances entre les mots. Une fois le comptage effectué par agrégation des mots, les X_j deviennent dépendant car $\sum_j X_j = m$.

Naive Bayes : cas général continu

Classifieur bayésien naïf : si X est réel -1-

- $\mathscr{D}_n = \{Z_1, \ldots, Z_n\}$ n-échantillon de v.a.i.i.d. $Z_i = (X_i, Y_i)$ pour $i = 1, \ldots, n$.
- $X_i = (X_{ij})_j \in \mathbb{R}^d$ observations issues d'un v.a. $X \in \mathbb{R}^d$: données, variables explicatives, « features ».
- Y_i issues d'une v.a. Y : étiquettes ou labels. $Y = y_1, \ldots, y_K$.
- $\mathbb{P}_\theta$, loi de (X, Y) inconnue, caractérisée par θ.
- $\theta_k = \mathbb{P}[Y = y_k]$, $k = 1, \ldots, K$: loi *a priori* de Y (inconnue).
- $f(x|Y = y_k)$ est la densité conditionnelle de X sachant $[Y = y_k]$.

Classifieur bayésien naïf : si X est réel -2-

- Hypothèse naïve : sachant $[Y = y_k]$, les lois conditionnelles des coordonnées $X_1, \ldots, X_d$ sont indépendantes :

$$f(x_1, \ldots, x_d | Y = y_k) = \prod_{j=1}^{d} f(x_j | Y = y_k) \tag{26}$$

- Le prédicteur est alors

$$h_{NB}(x) = \underset{k=1,\ldots,K}{\arg\max} \left(\mathbb{P}[Y = y_k] \prod_{j=1}^{d} f(x_j | Y = y_k) \right) \tag{27}$$

Classifieur bayésien naïf gaussien (GNB) -1-

- Un cas classique : la loi conditionnelle de X_j sachant $[Y = y_k]$ est gaussienne de moyenne μ_k et variance σ_k^2 :

$$f(x_j|Y = y_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{1}{2}\left(\frac{x_j-\mu_k}{\sigma_k}\right)^2} \tag{28}$$

- Estimation des paramètres :

$$n_k = \sum_{i=1}^{n} \mathbb{1}_{[Y_i=y_k]} \tag{29}$$

$$\widehat{\theta}_k = n_k/n \tag{30}$$

$$\widehat{\mu}_{jk} = \frac{1}{n_k} \sum_{i=1}^{n} x_{ij} \mathbb{1}_{[Y_i=y_k]} \tag{31}$$

$$\widehat{\sigma^2_{jk}} = \frac{1}{n_k} \sum_{i=1}^{n} (x_{ij} - \widehat{\mu}_{jk})^2 \mathbb{1}_{[Y_i=y_k]} \tag{32}$$

Classifieur bayésien naïf gaussien (GNB) -2-

- Sans hypothèse d'indépendance naïve, les estimateurs seraient :

$$\widehat{\mu}_k = \frac{1}{n_k} \sum_{i=1}^{n} x_i \mathbb{1}_{[Y_i = y_k]} \tag{33}$$

$$\widehat{\sigma_k^2} = \frac{1}{n_k} \sum_{i=1}^{n} (x_i - \widehat{\mu}_k)(x_i - \widehat{\mu}_k)^\top \mathbb{1}_{[Y_i = y_k]} \tag{34}$$

- Par exemple, avec $d = 2$,

Sans hypothèse naïve :

$$\widehat{\mu}_k = \begin{pmatrix} \widehat{\mu}_{1k} \\ \widehat{\mu}_{2k} \end{pmatrix}$$

$$\widehat{\Sigma}_k = \begin{pmatrix} \widehat{\sigma_{1k}^2} & \widehat{\sigma}_{12k} \\ \widehat{\sigma}_{21k} & \widehat{\sigma_{2k}^2} \end{pmatrix}$$

Avec hypothèse naïve de Bayes :

$$\widehat{\mu}_k = \begin{pmatrix} \widehat{\mu}_{1k} \\ \widehat{\mu}_{2k} \end{pmatrix}$$

$$\widehat{\Sigma}_k = \begin{pmatrix} \widehat{\sigma_{1k}^2} & 0 \\ 0 & \widehat{\sigma_{2k}^2} \end{pmatrix}$$

Classifieur bayésien naïf gaussien (GNB) -3-

- Classifieur bayésien naïf « théorique » :

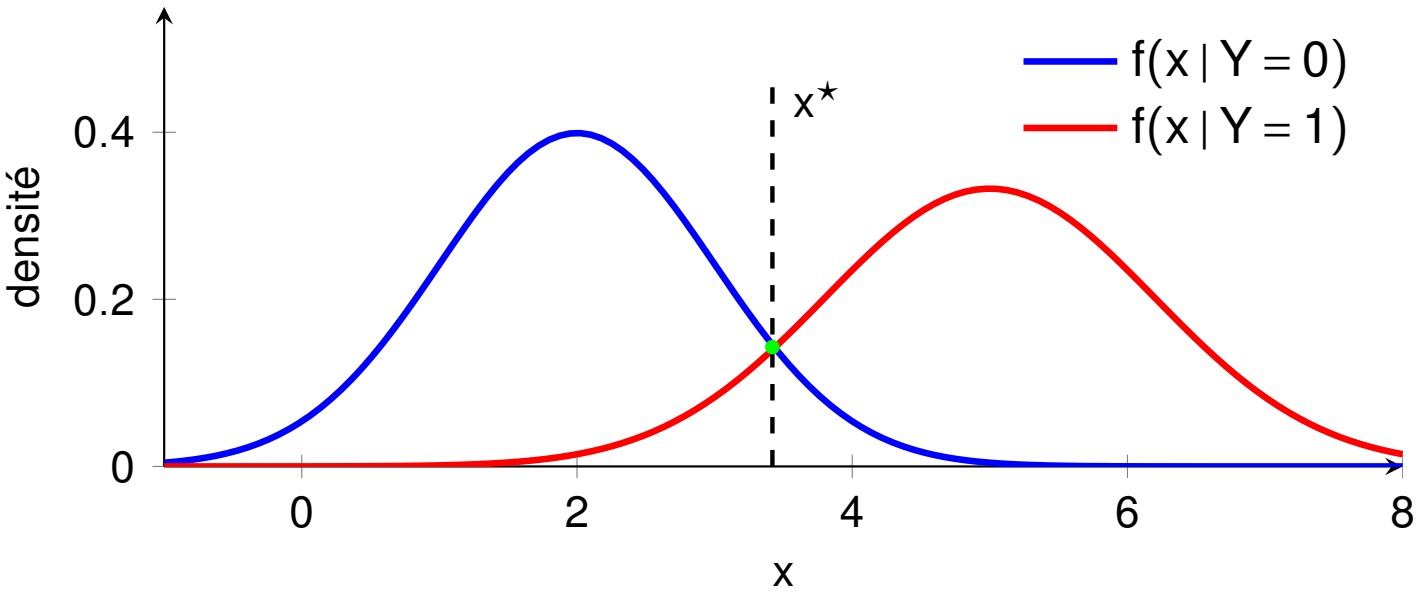
$$h_{NB}(x) = \underset{k=1,\dots,K}{\arg\max} \left(\mathbb{P}[Y = y_k] \prod_{j=1}^{d} f(x_j | Y = y_k) \right) \tag{35}$$

- Classifieur bayésien naïf « empirique » (le seul calculable) :

$$\widehat{h}_{NB}(x) = \underset{k=1,\dots,K}{\arg\max} \left[\widehat{\theta}_k \prod_{j=1}^{d} \left(\frac{1}{\sqrt{2\pi\widehat{\sigma}_k^2}} e^{-\frac{1}{2}\left(\frac{x_j - \widehat{\mu}_k}{\widehat{\sigma}_k}\right)^2} \right) \right] \text{😱} \tag{36}$$

Classifieur bayésien naïf gaussien (GNB) -4-

• Si K = 2 (spam / non spam) on peut visualiser la façon dont le choix s'effectue :



• Le point de coupure $x^\star$ satisfait :

$$\theta_0 f(x^\star|Y=0) = \theta_1 f(x^\star|Y=1) \tag{37}$$

à gauche de $x^\star$ la classe 0 est la plus probable, à droite, la classe 1 est la plus probable.

Exemple : prédire le genre -1-

- 8 individus à classer selon le genre :
- $Y \in \{\text{Homme}, \text{Femme}\}$
- $X = (\text{taille}, \text{poids}, \text{pointure})$...
- ... en unités US.
- $X_j \mid Y = k \sim \mathcal{N}(\mu_{jk}, \sigma^2_{jk})$.
- Hypothèse :
  $f(x \mid Y = y_k) = \prod_{j=1}^{3} f(x_j \mid Y = y_k)$

Id	Genre	Taille	Poids	Pointure
1	H	6.00	180	12
2	H	5.92	190	11
3	H	5.58	170	12
4	H	5.92	165	10
5	F	5.00	100	6
6	F	5.50	150	8
7	F	5.42	130	7
8	F	5.75	150	9

- Après calculs :

$$\widehat{\theta}_H = \frac{1}{2}, \qquad \widehat{\mu}_H = \begin{pmatrix} 5.9 \\ 176.2 \\ 11.2 \end{pmatrix}, \qquad \widehat{\sigma}^2_H = \begin{pmatrix} 0.03 \\ 92.2 \\ 0.69 \end{pmatrix}$$

$$\widehat{\theta}_F = \frac{1}{2}, \qquad \widehat{\mu}_F = \begin{pmatrix} 5.4 \\ 132.5 \\ 7.5 \end{pmatrix}, \qquad \widehat{\sigma}^2_F = \begin{pmatrix} 0.07 \\ 418.8 \\ 1.2 \end{pmatrix}$$

Exemple : prédire le genre -2-

- On veut prédire la classe de $x = (6, 130, 8)$.
- On compare les deux scores :

$$\widehat{\mathbb{P}}[Y = y_k | X = (6, 130, 8)] = S_k = P(Y = y_k) \prod_{j=1}^{3} p(x_j \mid Y = y_k).$$

- Numériquement : $S_H = 7.10^{-11}$, $S_F = 4,510^{-4}$, donc $S_F \gg S_H$.

$$\boxed{\widehat{y} = F}$$

- Pas de lissage de Laplace ici car les densités gaussiennes sont positives.

Naive Bayes et régression logistique

Naive Bayes et régression logistique -1-

- Plaçons-nous dans le cas où $Y = 0, 1$ et $X_j = 0, 1$. La décision $Y = 1$ est prise si le LLR (rapport de vraisemblance logarithmique) suivant est ≥ 0 :

$$
\begin{aligned}
&\log\left(\frac{\mathbb{P}[Y=1|X=x]}{\mathbb{P}[Y=0|X=x]}\right) \\
&= \log\left(\frac{\mathbb{P}[Y=1]\prod_{j=1}^{d}\mathbb{P}[X_j=x_j|Y=1]}{\mathbb{P}[Y=0]\prod_{j=1}^{d}\mathbb{P}[X_j=x_j|Y=0]}\right) \\
&= \log\left(\frac{\mathbb{P}[Y=1]\prod_{j=1}^{d}\theta_{j1}^{x_j}(1-\theta_{j1})^{1-x_j}}{\mathbb{P}[Y=0]\prod_{j=1}^{d}\theta_{j0}^{x_j}(1-\theta_{j0})^{1-x_j}}\right) \\
&= \underbrace{\log\frac{\mathbb{P}[Y=1]}{\mathbb{P}[Y=0]} + \sum_{j=1}^{d}\log\left(\frac{1-\theta_{j1}}{1-\theta_{j0}}\right)}_{b} + \sum_{j=1}^{d}\underbrace{\left(\log\frac{\theta_{j1}}{1-\theta_{j1}} - \log\frac{\theta_{j0}}{1-\theta_{j0}}\right)}_{w_j}x_j \\
&= b + \sum_{j=1}^{d} w_j x_j = b + w^\top x
\end{aligned}
$$

Naive Bayes et régression logistique -2-

$$\frac{\mathbb{P}[Y=1|X=x]}{\mathbb{P}[Y=0|X=x]} = e^{b+w^\top x}$$

$$\iff \frac{\mathbb{P}[Y=1|X=x]}{1-\mathbb{P}[Y=1|X=x]} = e^{b+w^\top x}$$

$$\iff \mathbb{P}[Y=1|X=x] = \frac{1}{1+e^{-b-w^\top x}}$$

$$\iff \mathbb{P}[Y=1|X=x] = \sigma(b+w^\top x)$$

- $\sigma(x) = (1+e^{-x})^{-1}$ fonction logistique.
- C'est exactement une régression logistique !
- Le classifieur NB est :

$$h_{NB}(x) = \mathbb{1}\left[\frac{\mathbb{P}[Y=1|X=x]}{\mathbb{P}[Y=0|X=x]} > 1\right]$$

$$= \mathbb{1}_{[b+w^\top x>0]}$$

$$= \text{sgn}(b+w^\top x > 0)$$

Naive Bayes et régression logistique -3-

- On suppose que $Y = 0, 1$ et que la loi de X sachant $[Y = y_k]$ est gaussienne $\sim \mathcal{N}(\mu_{jk}, \sigma_j^2)$ (les variances ne dépendent pas de Y). On a vu que $\mathbb{P}[Y = 1|X = x] = \sigma(b + w^\top x)$, avec

$$b + w^\top x = \log \frac{\theta_1}{1-\theta_1} + \sum_{j=1}^{d} \log \frac{f(x_j|Y=1)}{f(x_j|Y=0)}$$

$$= \log \frac{\theta_1}{1-\theta_1} + \sum_{j=1}^{d} \log \frac{(2\pi\sigma_j^2)^{-\frac{1}{2}} \exp\left(-\frac{(x_j-\mu_{j1})^2}{2\sigma_j^2}\right)}{(2\pi\sigma_j^2)^{-\frac{1}{2}} \exp\left(-\frac{(x_j-\mu_{j0})^2}{2\sigma_j^2}\right)}$$ 😱

$$= \log \frac{\theta_1}{1-\theta_1} + \sum_{j=1}^{d} \frac{(X_j-\mu_{j1})^2 - (X_j-\mu_{j0})^2}{2\sigma_j^2}$$

$$= \log \frac{\theta_1}{1-\theta_1} + \sum_{j=1}^{d} \left(\frac{\mu_{j1}-\mu_{j0})^2}{\sigma_j^2} x_j + \frac{\mu_{j0}^2 - \mu_{j1}^2}{2\sigma_j^2} \right)$$

Naive Bayes et régression logistique -4-

- On peut généraliser cela au cas multinomial.
- La méthode du classifieur bayésien naïf et celle de la régression logistique diffèrent pourtant :
- Naïve Bayes : méthode générative,
- les paramètres sont estimés à partir des probabilités empiriques.
- Régression logistique : méthode discriminante.
- les paramètres sont estimés à partir de la méthode du gradient.

Propriétés des classifieurs bayésiens naïfs

Classifieur bayésien naïf : un exemple de modèle génératif -1-

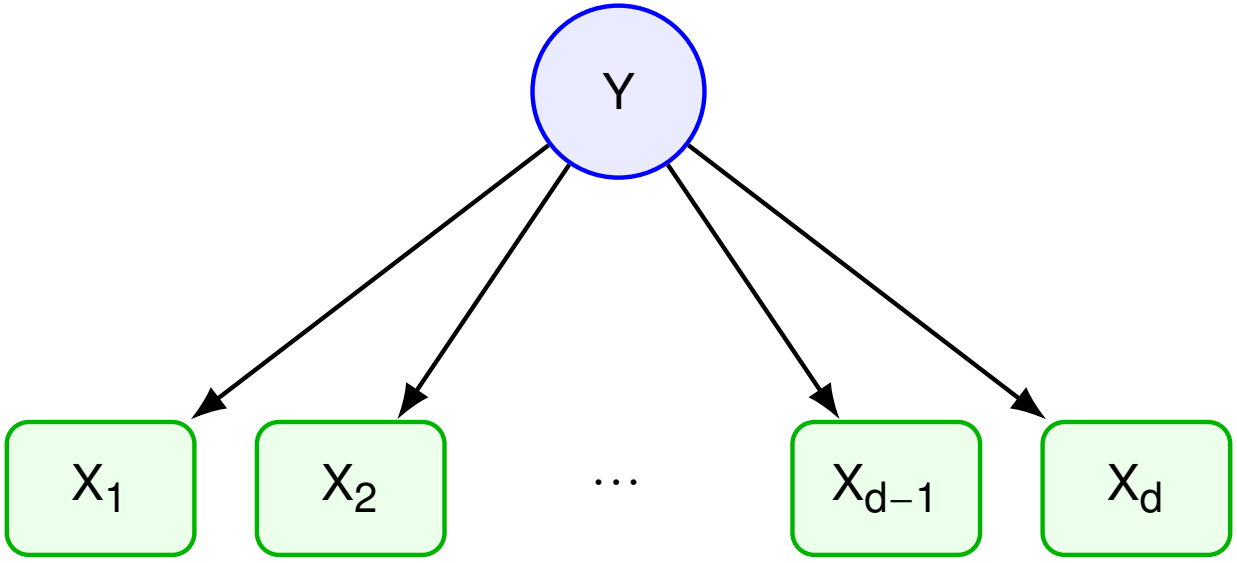


Figure 2 – Y est la cause commune des X_j

$$f_{(X,Y)}(x, y_k) = \mathbb{P}[Y = k] \prod_{j=1}^{d} f(x_j | Y = y_k). \tag{38}$$

- Modèle génératif : on modélise la loi jointe (X, Y), puis on en déduit les liens de causalité entre X et Y : génération de X à partir de Y.
- Les dépendances entre variables observées sont ignorées une fois la classe connue.

Classifieur bayésien naïf : un exemple de modèle génératif -2-

• Modèle génératif : modélise la loi jointe $\mathbb{P}_\theta$ de (X, Y). Estime comment les données sont structurées dans l'espace. Comprend les données, puis utilise un formalisme bayésien.

• Exemples : Naïve Bayes, LDA, HMM, GAN, LLM.

• Modèle discriminant : modélise la loi conditionnelle de Y sachant $[X = x]$. Sépare l'espace en classes et évalue les frontières des classes.

• Exemples : régression logistique, k-NN, arbres de décision, SVM, MLP.

• Génératif : « comment les données sont-elles produites ? » vs. discriminant : « où est la frontière entre les classes ? »

• Pour chacun des deux types de modèles, on peut être en classification ou en régresssion. Apprentissage supervisé : génératif et discriminant possibles. Apprentissage non supervisé : seulement du génératif.

Naive Bayes : pas une méthode bayésienne... mais parfois linéaire

- Malgré son nom, il ne s'agit pas d'une méthode bayésienne !

- Méthode bayésienne : estimateurs *a posteriori* étant donné une loi *a priori* et des observations de la vraisemblance.

- Exemple : MAP (maximum *a posteriori).*

- Naive Bayes : pas de loi *a priori*, hypothèse naïve non licite et les paramètres sont estimés de façon fréquentiste : on n'estime pas leur loi *a posteriori*.

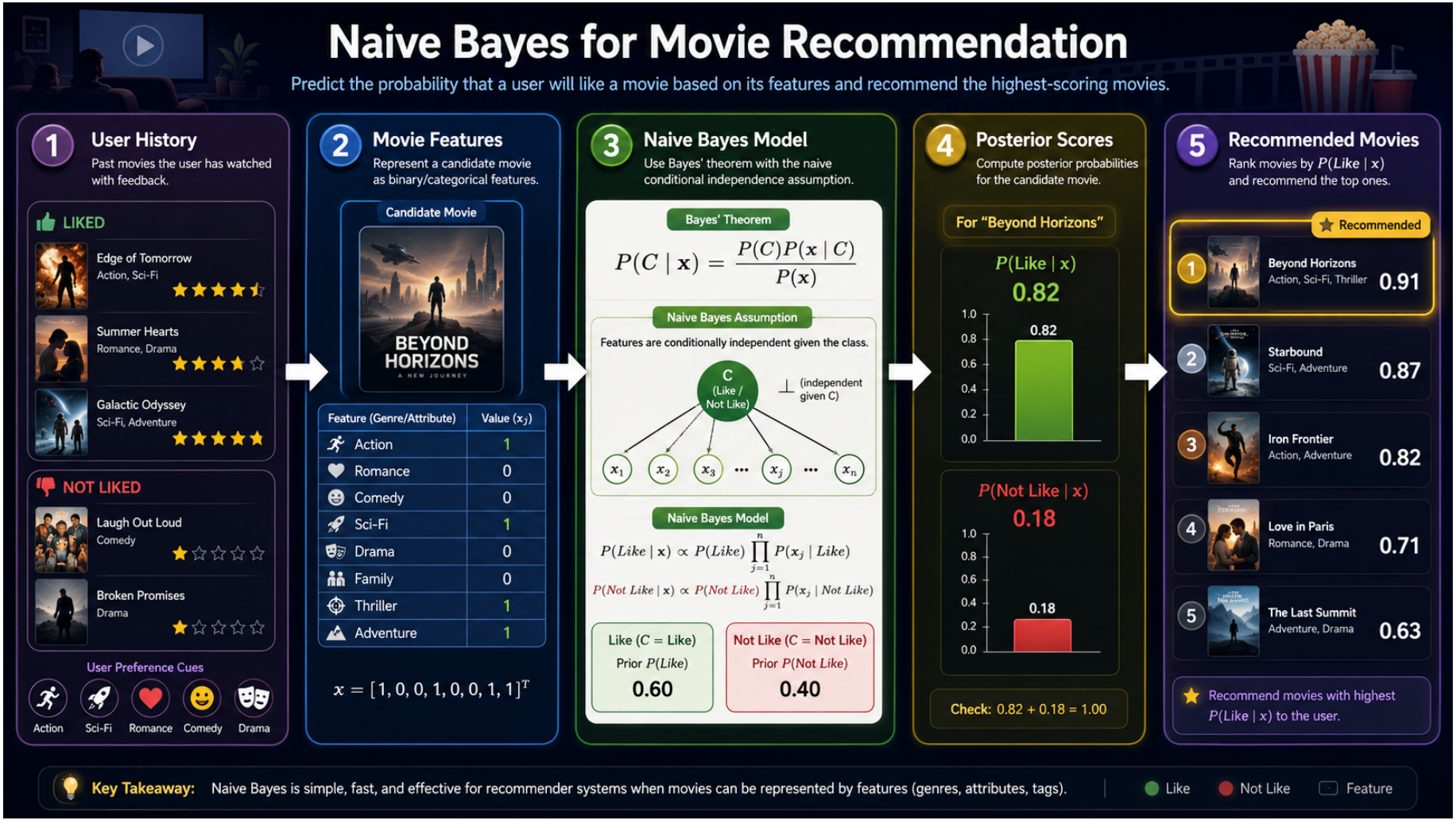
- ... mais classifieur linéaire si la loi de X sachant Y appartient au modèle exponentiel (bernoulli, binomial, multinomial, gaussien, etc.) et que les variances ne dépendent pas de Y :

$$\log \frac{\mathbb{P}[Y = 1|X = x]}{\mathbb{P}[Y = 0|X = x]} = b + w^\top x.$$

Applications -1-

- Filtrage de spam.
- Détection de phishing.
- Détection d'intrusion dans un réseau informatique.
- Classification de textes (articles de presse, détection de thème, tickets clients, documents juridiques ou médicaux).
- Avis clients.
- Systèmes de recommandations.
- Classification d'images simples (MNIST chiffres manuscrits).
- Détection de langue.

Applications -2-



Conclusion

Tableau de synthèse

Méthode	Inter-préta-tion	Perf.	Calibra-tion	Overfit.	Coût calc.
Régression linéaire	▲	▼	✗	▲	▲
Régression logistique	▲	▼	✗	▲	▲
Naive Bayes	▲	▼	▲	▲	▲
SVM (kernel RBF)	✗	▲	▼	▼	▼
Réseau de neurones	✗	▲	✗	✗	✗
Arbres de décision	▲	✗	▲	✗	▲
Boosting	✗	▲	▲	▼	▼
Bagging (forêt)	✗	▲	▼	▲	▼

▲ : bon ▼ : moyen ✗ : faible ou problématique

• Ce tableau est une sorte de moyenne de ce que l'on trouve sur internet, il vaut ce qu'il vaut et ne donne qu'une idée générique. Le contenu des cases est finalement assez aléatoire. En fait, ne vous fiez pas du tout à ce type de tableau et quand vous traitez un problème de Machine Learning, essayez toutes les méthodes !

Blibliographie

- Machine Learning, a Probabilistic Perspective, Kevin Murphy, MIT Press, 2012.
- An Empirical Study of the Naïve Bayes Classifier, Irina Rish, 2001, IJCAI 2001 Work Empir Methods Artif Intell.
- The Optimality of Naive Bayes, Harry Zhang, Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, 2004.
- Idiot's Bayes : Not So Stupid after All ? David J. Hand and Keming Yu, International Statistical Review, Vol. 69, No. 3, 2001.
- On the Optimality of the Simple Bayesian Classifier under Zero-One Loss, Domingos, P., Pazzani, M. Machine Learning, 29, 1997.
- On Discriminative vs. Generative Classifiers : A comparison of logistic regression and naive Bayes, Andrew Y. Ng, Michael I. Jordan, Advances in Neural Information Processing Systems 14 (NIPS 2001).
- Scikit Learn documentation.

Matériel pédagogique

- Polycopié de Machine Learning, diapos, exercices corrigés, sujets de TP, notebooks et matériel pédagogique sur cpmath.fr/ML. En particulier :
  - Une fiche de petits exercices corrigés sur Naive Bayes.
  - Un petit TP sur Naive Bayes sous forme de Notebook Python.
- Des questions sur le cours : par mail (claude.petit@univ-rennes.fr).
- SAV gratuit sans supplément et éternel.

Crédits photos et figures

- Diapo 4 : bag of words. ChatGPT.
- Diapo 15 : indépendance conditionnelle. Wikipedia, Azatoth, CC.
- Diapo 22 : SpamAssassin Public Mail Corpus.
- Diapo 23 : DeBounce email spam statistics. Ici.
- Diapo 26 : bag of words. ChatGPT.
- Diapo 28 : loi multinomiale. ChatGPT.
- Diapo 49 : système de recommandation. ChatGPT.