

# APPRENTISSAGE SUPERVISÉ

## DEUXIÈME PARTIE

Cours précédent :

Concept de l'**apprentissage supervisé** : montrer à l'ordinateur des exemples pour qu'il apprenne une règle.

**Phase d'apprentissage** : on connaît l'entrée  $X$  et la sortie/réponse  $Y$ , on cherche la fonction qui les lie  $Y = f(X)$

**Optimisation des paramètres** : pendant la phase d'apprentissage, grâce à la validation croisée

**Phase de test** : on fait prédire des valeurs à l'algorithme et on les compare aux vraies valeurs pour estimer sa performance.

**Deux types de modèles** : algorithmiques (semaine dernière) et probabilistes (cette semaine).

# Naive Bayes

# PROBABILITÉS CONDITIONNELLES

---

Si A et B sont deux événements, la probabilité que A se produise conditionnellement au fait que B se produise se dit **probabilité de A sachant B** s'écrit :  $P(A|B)$ .

*Exemple 1:*

A : "il pleut aujourd'hui"

B : "il pleuvait hier"

$P(A|B)$  : probabilité qu'il pleuve aujourd'hui sachant qu'il pleuvait hier.

$P(B|A)$  : probabilité qu'il ait plu hier sachant qu'il pleut aujourd'hui.

$P(A, B) = P(A \cap B)$  : **probabilité jointe** qu'il ait plu hier **et** qu'il pleuve aujourd'hui.

# PROBABILITÉS CONDITIONNELLES

---

Si A et B sont deux événements, la probabilité que A se produise conditionnellement au fait que B se produise se dit **probabilité de A sachant B** s'écrit :  $P(A|B)$ .

Exemple 2:

A : "j'achète"

B : "j'ai vu une pub"

$P(A|B)$  : probabilité que j'achète sachant que j'ai vu une pub .

$P(B|A)$  : probabilité que j'aie vu une pub sachant que j'ai acheté.

$P(A, B) = P(A \cap B)$  : **probabilité jointe** que j'aie vu une pub **et** que j'aie acheté.

# CALCUL DE CES PROBABILITÉS

Si A et B sont deux événements, alors:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Jour	Lundi	Mardi	Mercredi	Jeudi	Vendredi
Hier	Pluie	Pluie	Soleil	Soleil	Pluie
Aujourd'hui	Pluie	Soleil	Soleil	Pluie	Soleil

**Probabilités simples** :  $P(\text{pluie aujourd'hui}) = 2/5$ ,  $P(\text{pluie hier}) = 3/5$

**Probabilité jointe** :  $P(\text{pluie aujourd'hui} \cap \text{pluie hier}) = 1/5$

**Probabilités conditionnelles** :

$$P(\text{pluie aujourd'hui} | \text{pluie hier}) = \frac{P(\text{pluie aujourd'hui} \cap \text{pluie hier})}{P(\text{pluie hier})} = \frac{1/5}{3/5} = \frac{1}{3}$$

$$P(\text{pluie hier} | \text{pluie aujourd'hui}) = \frac{P(\text{pluie hier} \cap \text{pluie aujourd'hui})}{P(\text{pluie aujourd'hui})} = \frac{1/5}{2/5} = \frac{1}{2}$$

# CALCUL DE CES PROBABILITÉS

Si A et B sont deux événements, alors:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Obs	1	2	3	4	5	6	7	8
Pub	Oui	Non	Oui	Oui	Oui	Non	Oui	Oui
Achat	Non	Oui	Non	Oui	Non	Non	Non	Non

**Probabilités simples** :  $P(\text{pub}) = 6/8$ ,  $P(\text{achat}) = 2/8$

**Probabilité jointe** :  $P(\text{pub} \cap \text{achat}) = 1/8$

**Probabilités conditionnelles** :

$$P(\text{pub}|\text{achat}) = \frac{P(\text{pub} \cap \text{achat})}{P(\text{achat})} = \frac{1/8}{2/8} = \frac{1}{2}$$

$$P(\text{achat}|\text{pub}) = \frac{P(\text{pub} \cap \text{achat})}{P(\text{pub})} = \frac{1/8}{6/8} = \frac{1}{6}$$

# CALCUL DE CES PROBABILITÉS

Si A et B sont deux événements, alors:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Obs	1	2	3	4	5	6	7	8
Pub	Oui	Non	Oui	Oui	Oui	Non	Oui	Oui
Achat	Non	Oui	Non	Oui	Non	Non	Non	Non

**Probabilités simples** :  $P(\text{pub}) = 6/8$ ,  $P(\text{achat}) = 2/8$

**Probabilité jointe** :  $P(\text{pub} \cap \text{achat}) = 1/8$

**Probabilités conditionnelles** :

$$P(\text{pub}|\text{achat}) = \frac{P(\text{pub} \cap \text{achat})}{P(\text{achat})} = \frac{1/8}{2/8} = \frac{1}{2}$$

$$P(\text{achat}|\text{pub}) = \frac{P(\text{pub} \cap \text{achat})}{P(\text{pub})} = \frac{1/8}{6/8} = \frac{1}{6}$$



# THEOREME DE BAYES

Si A et B sont deux événements:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) \times P(A)}{P(B)}$$



Dans une université, il y a 34% de femmes.

Parmi les étudiants en informatique, 22% sont des femmes.

Les étudiants en informatique représentent 20% de l'université.

Quelle est la proportion d'étudiantes en informatique parmi les femmes de l'université ?

# THEOREME DE BAYES

Si A et B sont deux événements:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) \times P(A)}{P(B)}$$



Dans une université, il y a 34% de femmes.

Parmi les étudiants en informatique, 22% sont des femmes.

Les étudiants en informatique représentent 20% de l'université.

Quelle est la proportion d'étudiantes en informatique parmi les femmes de l'université ?

On cherche :  $P(\text{étudier l'informatique} | \text{être une femme})$

# THEOREME DE BAYES

Si A et B sont deux événements:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) \times P(A)}{P(B)}$$



Dans une université, il y a 34% de femmes.

Parmi les étudiants en informatique, 22% sont des femmes.

Les étudiants en informatique représentent 20% de l'université.

Quelle est la proportion d'étudiantes en informatique parmi les femmes de l'université ?

On cherche :  $P(\text{étudier l'informatique} | \text{être une femme})$

Solution:

$$P(\text{info} | \text{femme}) = \frac{P(\text{femme} | \text{info}) \times P(\text{info})}{P(\text{femme})} = \frac{0,22 \times 0,2}{0,34} = 12,9\%$$

# NAIVE BAYES

---

Pour chaque mot  $W$  rencontré dans l'ensemble d'apprentissage, on calcule la probabilité qu'un message soit un spam **sachant** qu'il contient le mot  $W$  (**spamicité** du mot  $W$ ):

$$P(S|W) = \frac{\overbrace{P(W|S)}^{\text{\% de spams contenant } W} \times \overbrace{P(S)}^{\text{\% de spams}}}{\underbrace{P(W)}_{\text{\% de messages contenant } W}}$$

# NAIVE BAYES - TEST

---

Pour un message de l'ensemble de test, on calcule la **probabilité que le message soit un spam** sachant qu'il contient les mots

$W_1, W_2, W_3 \dots W_n$  :

$$P(S|W_1, W_2, \dots, W_n) = \frac{\overbrace{p_1 p_2 \dots p_n}^{\text{spamité des mots du message}}}{\underbrace{p_1 p_2 \dots p_n}_{\text{spamité}} + \underbrace{(1 - p_1)(1 - p_2) \dots (1 - p_n)}_{\text{non-spamité}}}$$

avec  $p_1 = P(S|W_1)$ ,  $p_2 = P(S|W_2)$ ,  $\dots$ ,  $p_n = P(S|W_n)$

# REMARQUES

---

Le classifieur Naive Bayes considère que **tous les mots sont indépendants**. Autrement dit, il regarde l'effet des mots un par un, sans se soucier des autres. C'est la raison pour laquelle il est **naïf**. Et pourtant ! Existant depuis longtemps (1996) et malgré son caractère naïf, il reste très utilisé, en particulier pour la classification du spam et donne de **bons résultats**.

Il est recommandé de **supprimer les mots apparaissant trop peu ou trop souvent** qui peuvent poser problème ou l'induire en erreur. Il peut classer les documents dans plusieurs catégories, le cas binaire du spam étant un cas particulier.

# CONCLUSIONS SUR NAIVE BAYES

---

## Avantages :

Modèle très simple.

Rapide car très peu de calculs.

Efficace dans certains cas.

## Inconvénients :

Sa naïveté ne s'applique pas à tous les problèmes.

Il ne dispense pas (au contraire!) de préprocesser le texte (comme tous les algorithmes).

## En Python :

```
from sklearn.naive_bayes import BernoulliNB
model = BernoulliNB()
model.fit(Xtrain, Ytrain)
predictions = model.predict(Xtest)
```

# Régression linéaire



# RÉGRESSION LINÉAIRE SIMPLE

---

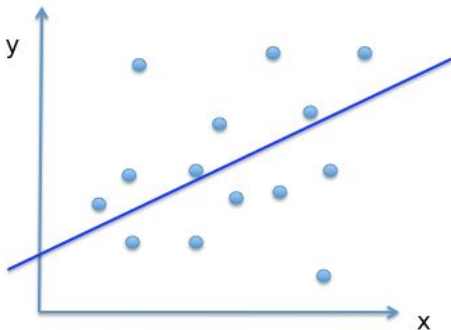
A partir de n exemples, trouver la droite qui passe en leur "centre".

$\beta_0$  : biais (valeur en  $x = 0$ )

$\beta_1$  : pente

réponse =  $\beta_0 + \beta_1 \times$  (variable dépendante) + bruit

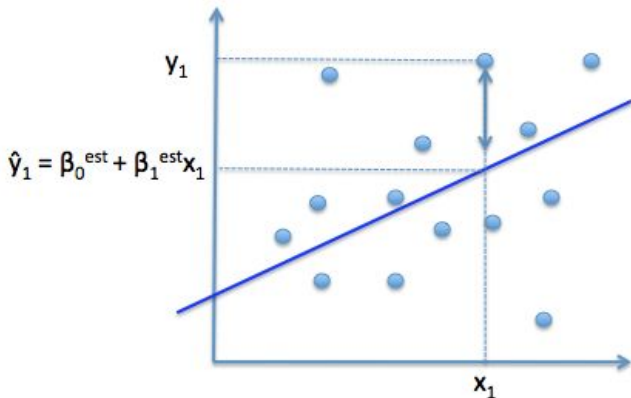
$$Y = \beta_0 + \beta_1 x + \varepsilon$$



# ESTIMATION

Objectif : minimiser la **variance résiduelle** (i.e.  $\varepsilon$ ) trouver la meilleure droite :

$$\min \sum_{i=1}^n (y_i - \hat{y}_i)^2 \Leftrightarrow \min_f \sum_{i=1}^n (y_i - f(x_i))^2 \Leftrightarrow \min_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$



# REGRESSION LINEAIRE MULTIPLE

---

En ajoutant de nouvelles variables :

réponse =  $\beta_0$  +  $\beta_1 \times \text{variable}_1$  +  $\beta_2 \times \text{variable}_2$  +  $\dots$  + bruit

$$Y = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \dots + \varepsilon$$

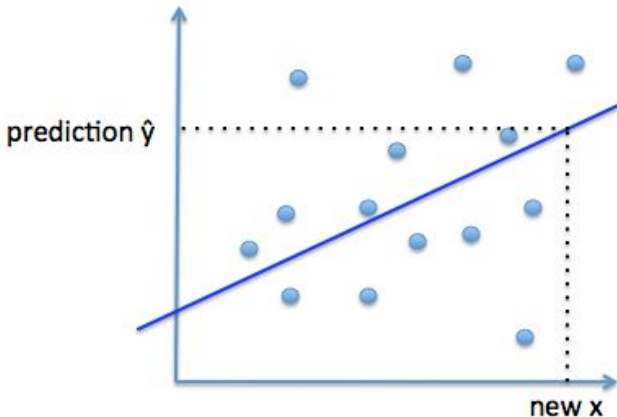
$$Y = \beta_0 + \sum_{j=1}^p \beta_j x_j + \varepsilon = X\beta + \varepsilon$$

Solution :

$$(\hat{\beta}_0, \hat{\beta}) = \underbrace{\arg \min_{\beta_0 \in \mathbf{R}, \beta \in \mathbf{R}^p}}_{\text{Trouver les valeurs qui minimisent}} \underbrace{\sum_{i=1}^n (y_i - (\beta_0 + \beta x_i))^2}_{\text{la fonction de coût}}$$

# PHASE DE TEST

Quand un nouveau point  $x$  est donné, on peut **prédire** le  $\hat{y}$  correspondant:



# INTERPRÉTATION

## Supposons un modèle :

$$\text{salaire} = \beta_0 + \beta_1 \times \text{expérience} + \beta_2 \times \text{études} + \text{bruit}$$

## Et supposons que les résultats soient :

$$\beta_0 = 900$$

$$\beta_1 = 100$$

$$\beta_2 = 200$$

## On interprète que :

Quelqu'un qui n'a ni études ni expérience touchera 900 euros en moyenne.

Quelqu'un qui a un bac+5 et 5 ans d'expérience touchera 2400 euros en moyenne.

Une année d'expérience supplémentaire rapporte en moyenne 100 euros de plus.

Une année d'études supplémentaire rapporte en moyenne 200 euros de plus.

# ÉVALUATION

On cherche à expliquer la **variance totale** (total sum of squares) :

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$$

On calcule la **variance expliquée** (explained sum of squares) :

$$ESS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

On veut minimiser la **variance résiduelle** (residual sum of squares) :

$$RSS = \sum_{i=1}^n (\hat{y}_i - y_i)^2$$

Et on a :  $TSS = ESS + RSS$

Le **pourcentage de variance expliquée** donne donc une indication sur la qualité de la régression qu'on appelle  $R^2$  (à calculer également sur l'ensemble de test) :

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

Le  $R^2$  se trouve entre 0 et 1 :

0 : rien n'est expliqué par les variables

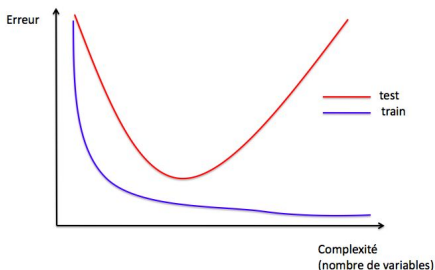
1 : tout est expliqué par les variables

# LIMITES

---

Le modèle linéaire est séduisant par sa simplicité mais vite limité :

- La solution demande l'inversion de la matrice  $X^T X$  qui n'est pas inversible si l'on a **plus de variables que d'observations**.
- Si des variables sont **corrélées**, cela va perturber le modèle : deux variables très ressemblantes n'auront pas forcément les mêmes poids.
- Plus on ajoute de variables, plus le modèle devient **instable** et on risque donc le **sur-apprentissage**.



# CONCLUSION SUR LA RÉGRESSION LINÉAIRE

---

## Avantages :

- Modèle simple et facile à estimer, solution “exacte”.

- Interprétable.

- Adapté pour les problèmes simples.

## Inconvénients :

- Impossible en grande dimension.

- On se trouve rarement dans des situations où ce modèle est pertinent.

## En Python :

```
from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(Xtrain,Ytrain)
```



# Régression logistique

# RÉGRESSION LOGISTIQUE BINAIRE

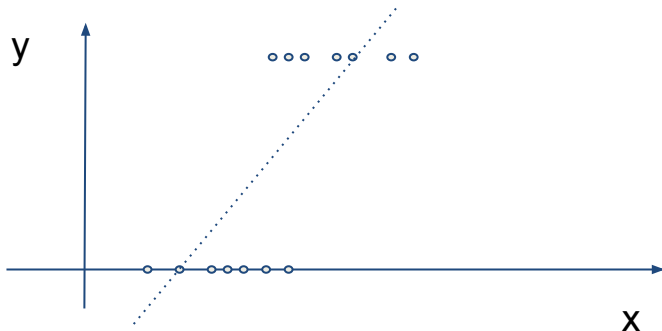
---

Dans la régression linéaire, on prédit une variable  $Y$  continue  
La régression logistique permet en fait de faire de la **classification**.  
On considère ici la classification binaire :  $Y \in \{0, 1\}$ .

# PRINCIPE

---

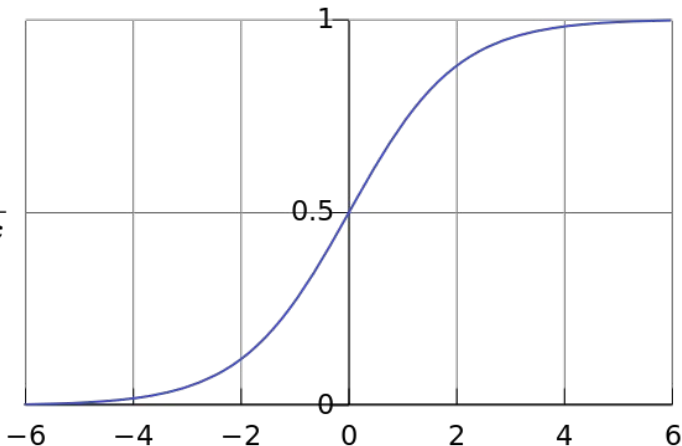
Puisque l'on a une variable à expliquer qui ne prend que les valeurs 0 et 1, un modèle linéaire n'a pas de sens.



# LA FONCTION LOGIT

---

$$f(x) = \frac{e^x}{1 + e^x}$$



Source : [wikipedia.com](https://fr.wikipedia.org/wiki/Fonction_logit)

**La régression logistique revient à appliquer cette fonction à une régression linéaire simple pour "pousser" les valeurs vers 0 et 1.**

# PRINCIPE

---

Comme d'habitude, X est l'input, Y l'output. La régression logistique binaire repose sur le modèle suivant :

$$P(Y = 1|X) = \frac{e^{b_0 + b_1 x_1 + \dots + b_p x_p}}{1 + e^{b_0 + b_1 x_1 + \dots + b_p x_p}}$$

toujours compris entre 0 et 1

On réfléchit en termes de **rapport de probabilités** ou encore de **odd ratios** :

$$\ln \frac{P(Y = 1|X)}{P(Y = 0|X)} = b_0 + b_1 x_1 + \dots + b_p x_p$$

# INTERPRÉTATION

---

La régression logistique permet d'interpréter l'effet de chaque variable sur ce qu'on essaie d'expliquer.

Exemple: on cherche à expliquer si un étudiant va valider une matière.

$$\ln \frac{P(Valider = 1 | \text{travail, cadeaux, video})}{P(Valider = 0 | \text{travail, cadeaux, video})} = b_0 + b_1 \text{travail} + b_2 \text{cadeaux} + b_3 \text{video}$$

On trouve  $\hat{b}_1 = 0.5$ ,  $\hat{b}_2 = 1$ ,  $\hat{b}_3 = -0.5$ . On en déduit que:

- $\exp(0.5) = 1.65$  : chaque unité de travail en plus multiplie par 1.65 les chances de valider.
- $\exp(1) = 2.7$  : chaque cadeau supplémentaire fait au professeur multiplie par 2.7 les chances de valider.
- $\exp(-0.5) = 0.6$  : chaque vidéo regardée pendant le cours divise par  $1/0.6 = 1.65$  les chances de valider.

# CONCLUSIONS SUR LA RÉGRESSION LOGISTIQUE

---

## Avantages

- Une vision probabiliste de la classification
- Interprétation de chaque variable de manière probabiliste
- Simple et rapide à estimer

## Inconvénients

- Mêmes que la régression linéaire
- En particulier, mauvais résultats avec beaucoup de variables.

## En Python :

```
from sklearn.linear_model import LogisticRegression
model = LogisticRegression()
model.fit(Xtrain, Ytrain)
model.predict(Xtest)
```