ENSA
ÉCOLE NATIONALE DES SCIENCES APPLIQUÉES
KHOURIBGA

جامعة السلطان مولاي سليمان
Université Sultan Moulay Slimane

ENSA KHOURIBGA ÉCOLE NATIONALE DES SCIENCES APPLIQUÉES

Projet de Fin de module
Data mining

Master Big Data et Aide à la Décision

# LE CLASSIFICATEUR BAYÉSIEN NAÏF

Présenté par

CHAMAKH CHAIMAA
HAFIANE BOUTAYNA
LOUDGHIRI OUMAYMA

# Remerciement

Nous tenons à remercier dans un premier temps, toute l'équipe pédagogique de l'Ecole Nationale des Sciences Appliquées de Khouribga.

Avant d'entamer ce rapport, nous profitons de l'occasion pour remercier tout d'abord notre professeur Madame Ourdou qui n'a pas cessé de nous encourager pendant la durée du projet, ainsi pour sa générosité en matière de formation et d'encadrement. Nous le remercions également pour l'aide et les conseils concernant les missions évoquées dans ce rapport, qu'il nous a apporté lors des différents suivis, et la confiance qu'il nous a témoigné.

Nous remercions aussi tous les professeurs du département informatiques et tous les professeurs de l'Ecole Nationale des Sciences Appliquées de Khouribga. pour tous leurs efforts qui ont consacré pour assurer notre enseignement.

# Table des matières

# Table des figures

# Introduction générale

L'intelligence artificielle doit fonctionner avec des données qui, dans de nombreux cas, sont volumineuses mais incomplètes. Tout comme les humains, l'ordinateur doit prendre des risques et penser à l'avenir qui n'est pas certain.

L'incertitude est difficile à supporter pour les êtres humains. Mais dans l'apprentissage automatique, certains algorithmes vous aident à contourner cette limitation. L'algorithme d'apprentissage automatique Naive Bayes est l'un des outils permettant de gérer l'incertitude à l'aide de méthodes probabilistes.

La probabilité est un domaine des mathématiques qui nous permet de raisonner sur l'incertitude et d'évaluer la probabilité de certains résultats ou événements. Lorsque vous travaillez avec la modélisation ML prédictive, vous devez prédire un avenir incertain. Par exemple, vous pouvez essayer de prédire la performance d'un champion olympique lors des prochains Jeux Olympiques en vous basant sur les résultats passés. Même s'ils ont déjà gagné, cela ne veut pas dire qu'ils gagneront cette fois. Des facteurs imprévisibles, comme une dispute avec leur partenaire ou le fait de ne pas avoir le temps de déjeuner, peuvent influencer leurs résultats.

Par conséquent, l'incertitude fait partie intégrante de la modélisation de l'apprentissage automatique, car la vie est compliquée et rien n'est parfait. Les trois principales sources d'incertitude dans l'apprentissage automatique sont les données bruyantes, la couverture incomplète du problème et les modèles imparfaits.

# Chapitre 1

# Apprentissage automatique

## 1.1 Introduction

Ce chapitre introduit le vocabulaire de l'apprentissage automatique (machine learning dans la littérature anglo-saxonne). La discipline étant relativement récente et en mutation constante, le vocabulaire évolue et est sujet à des abus de langage, en particulier lorsqu'on francise des termes techniques issus de la littérature scientifique en langue anglaise. L'objectif de cette introduction est également de dresser un panorama de l'apprentissage et d'expliciter l'articulation entre les chapitres du cours.

## 1.2 Qu'est-ce que l'apprentissage automatique ?

l'apprentissage automatique est apprentissage artificiel ou apprentissage statistique est un champ d'étude de l'intelligence artificielle qui se fonde sur des approches mathématiques et statistiques pour donner aux ordinateurs la capacité d' « apprendre » à partir de données, c'est-à-dire d'améliorer leurs performances à résoudre des tâches sans être explicitement programmés pour chacune. Plus largement, il concerne la conception, l'analyse, l'optimisation, le développement et l'implémentation de telles méthodes. On parle d'apprentissage statistique car l'apprentissage consiste à créer un modèle dont l'erreur statistique moyenne est la plus faible possible.
Trois grandes approches relèvent de l'apprentissage automatique : l'apprentissage supervisé, l'apprentissage non-supervisé, et l'apprentissage par renforcement. Bien entendu, cette classification est sujette à discussion, l'apprentissage semi-supervisé ou l'apprentissage faiblement supervisé (par exemple) apparaissant aux interfaces de ces approches. Ce rapport a destinne a une methode de l'apprentissage supervise le classifieur naif bayesien.
Avant de détailler cette methode nous esseyons d'aborder la definition des deux grand branches apprentissage supervisé et non-supervisé.

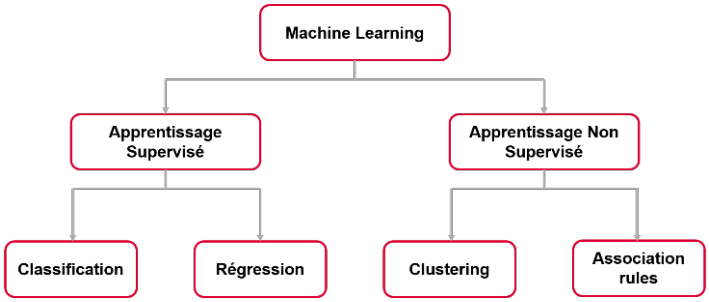


FIGURE 1.1 – Décomposition du machine learning

### 1.2.1 Apprentissage non-supervisé

l'apprentissage non supervisé désigne la situation d'apprentissage automatique où les données ne sont pas étiquetées (par exemple étiquetées comme « balle » ou « poisson »). Il s'agit donc de découvrir les structures sous-jacentes à ces données non étiquetées. Puisque les données ne sont pas étiquetées, il est impossible à l'algorithme de calculer de façon certaine un score de réussite. Ainsi, les méthodes non supervisées présentent une auto-organisation qui capture les modèles comme des densités de probabilité ou, dans le cas des réseaux de neurones, comme combinaison de préférences de caractéristiques neuronales encodées dans les poids et les activations de la machine.
Les autres niveaux du spectre de supervision sont l'apprentissage par renforcement où la machine ne reçoit qu'un score de performance numérique comme guide, et l'apprentissage semi-supervisé où une petite partie des données est étiquetée.
L'introduction dans un système d'une approche d'apprentissage non supervisé est un moyen d'expérimenter l'intelligence artificielle. En général, des systèmes d'apprentissage non supervisé permettent d'exécuter des tâches plus complexes que les systèmes d'apprentissage supervisé, mais ils peuvent aussi être plus imprévisibles. Même si un système d'IA d'apprentissage non supervisé parvient tout seul, par exemple, à faire le tri entre des chats et des chiens, il peut aussi ajouter des catégories inattendues et non désirées, et classer des races inhabituelles, introduisant plus de bruit que d'ordre

### 1.2.2 Apprentissage supervisé

L'apprentissage supervisé (supervised learning en anglais) est une tâche d'apprentissage automatique consistant à apprendre une fonction de prédiction à partir d'exemples annotés, au contraire de l'apprentissage non supervisé. On distingue les problèmes de régression des problèmes de classement1. Ainsi, on considère que les problèmes de prédiction d'une variable quantitative sont des problèmes de régression tandis que les problèmes de prédiction d'une variable qualitative sont des problèmes de classification.
Les exemples annotés constituent une base d'apprentissage, et la fonction de prédiction apprise peut aussi être appelée « hypothèse » ou « modèle ». On suppose cette base d'apprentissage représentative d'une population d'échantillons plus large et le but des méthodes d'apprentissage supervisé est de bien généraliser, c'est-à-dire d'apprendre une fonction qui

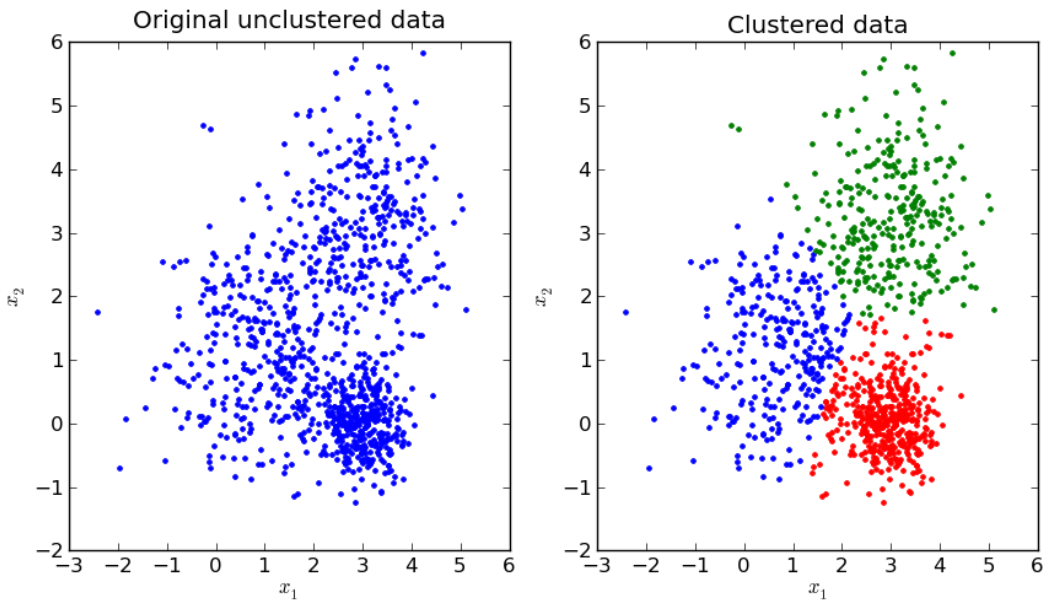


FIGURE 1.2 – Un échantillon de clusters trouvés à partir de données non étiquetées

fasse des prédictions correctes sur des données non présentes dans l'ensemble d'apprentissage.

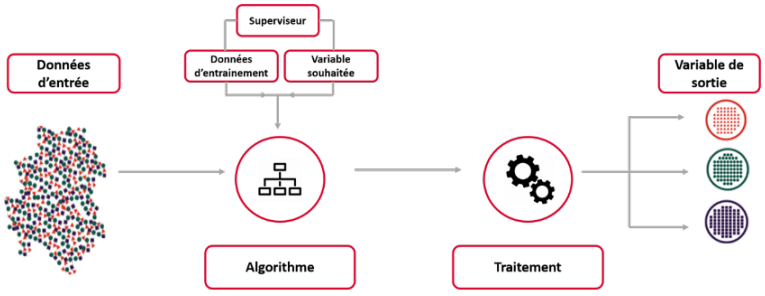


FIGURE 1.3 – Apprentissage automatique supervisé

# Chapitre 2

# Classifieur naive bayesienne

## 2.1 théorème de Bayes

Le théorème de Bayes est l'un des principaux théorèmes de la théorie des probabilités. Il est aussi utilisé en statistiques du fait de son application, qui permet de déterminer la probabilité qu'un événement arrive à partir d'un autre évènement qui s'est réalisé, notamment quand ces deux évènements sont interdépendants.
En d'autres termes, à partir de ce théorème, il est possible de calculer précisément la probabilité d'un évènement en tenant compte à la fois des informations déjà connues et des données provenant de nouvelles observations. La formule de Bayes peut être dérivée des axiomes de base de la théorie des probabilités, en particulier de la probabilité conditionnelle.
La particularité du théorème de Bayes est que son application pratique nécessite un grand nombre de calculs, c'est pourquoi les estimations bayésiennes n'ont commencé à être utilisées activement qu'après la révolution des technologies informatiques et de réseau.
Le théorème de Bayes est un corollaire du théorème de probabilité totale. Il énonce des probabilités conditionnelles de plusieurs évènements. Par exemple, pour les évènements A et B, il permet de déterminer la probabilité de A sachant B, si l'on connaît les probabilités de A , de B et de B sachant A, à condition que la probabilité de B ne soit pas égale à 0.

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

à condition que P(B) != 0 et où :

A et B sont deux évènements ;

— P(A) et P(B) sont la probabilité des deux évènements.
— P(A | B) est la probabilité conditionnelle que l'évènement A se réalise étant donné que l'évènement B s'est réalisé.
— P(B | A) est la probabilité conditionnelle que l'évènement B se réalise étant donné que l'évènement A s'est réalisé.

## 2.2 Le classifier bayésien naïf

La classification naïve bayésienne est un type de classification bayésienne probabiliste simple basée sur le théorème de Bayes avec une forte indépendance (dite naïve) des hypothèses. Elle met en œuvre un classifieur bayésien naïf, ou classifieur naïf de Bayes, appartenant à la famille des classifieurs linéaires.
Un terme plus approprié pour le modèle probabiliste sous-jacent pourrait être « modèle à caractéristiques statistiquement indépendantes ».
En termes simples, un classifieur bayésien naïf suppose que l'existence d'une caractéristique pour une classe, est indépendante de l'existence d'autres caractéristiques. Un fruit peut être considéré comme une pomme s'il est rouge, arrondi, et fait une dizaine de centimètres. Même si ces caractéristiques sont liées dans la réalité, un classifieur bayésien naïf déterminera que le fruit est une pomme en considérant indépendamment ces caractéristiques de couleur, de forme et de taille.
Selon la nature de chaque modèle probabiliste, les classifieurs bayésiens naïfs peuvent être entraînés efficacement dans un contexte d'apprentissage supervisé. Dans beaucoup d'applications pratiques, l'estimation des paramètres pour les modèles bayésiens naïfs repose sur le maximum de vraisemblance. Autrement dit, il est possible de travailler avec le modèle bayésien naïf sans se préoccuper de probabilité bayésienne ou utili :ser les méthodes bayésiennes.

Le modèle probabiliste pour un classifieur est le modèle conditionnel :

$$P(C|F_1, \ldots\ldots\ldots, F_n)$$

où C est une variable de classe dépendante dont les instances ou classes sont peu nombreuses, conditionnée par plusieurs variables caractéristiques

$$F_1, ..., F_n$$

Lorsque le nombre de caractéristiques n est grand, ou lorsque ces caractéristiques peuvent prendre un grand nombre de valeurs, baser ce modèle sur des tableaux de probabilités devient impossible. Par conséquent, nous le dérivons pour qu'il soit plus facilement soluble. À l'aide du théorème de Bayes, nous écrivons :

$$P(C|F_1, \ldots\ldots\ldots, F_n) = \frac{P(C) * P(F_1, \ldots\ldots\ldots F_n|C)}{P(F_1, \ldots\ldots\ldots, F_n)}$$

En pratique, seul le numérateur nous intéresse, puisque le dénominateur ne dépend pas de C et les valeurs des caractéristiques Fi sont données.
Le dénominateur est donc en réalité constant. Le numérateur est soumis à la loi de probabilité à plusieurs variables et peut être factorisé de la façon suivante, en utilisant plusieurs fois la définition de la probabilité conditionnelle :

$$P(C|F_1, \ldots, F_n) \tag{2.1}$$
$$= P(C)P(F_1, \ldots, F_n|C) \tag{2.2}$$
$$= P(C)P(F_1|C)P(F_2, \ldots, F_n|C, F_1) \tag{2.3}$$
$$= P(C)P(F_1|C)P(F_2|C, F_1)P(F_3, \ldots, F_n|C, F_1, F_2) \tag{2.4}$$
$$= P(C)P(F_1|C)P(F_2|C, F_1)P(F_3|C, F_1, F_2)P(F_4, \ldots, F_n|C, F_1, F_2, F_3) \tag{2.5}$$
$$= P(C)P(F_1|C)P(F_2|C, F_1)P(F_3|C, F_1, F_2)\ldots P(F_n|C, F_1, F_2, F_3, \ldots, F_(n-1)) \tag{2.6}$$
$$\tag{2.7}$$

C'est là que nous faisons intervenir l'hypothèse naïve : si chaque Fi est indépendant des autres caractéristiques Fj != i, conditionnellement à C alors

$$P(F_i|C, F_j) = P(F_i|C)$$

pour tout j != i, par conséquent la probabilité conditionnelle peut s'écrire

$$P(F_1, \ldots, F_n|C) = P(F_1|C)P(F_2|C)P(F_3|C)\ldots P(F_n|C) = \prod_{i=1}^{n} P(F_i|C)$$

Par conséquent, en tenant compte de l'hypothèse d'indépendance ci-dessus, la probabilité conditionnelle de la variable de classe C peut être exprimée par :

$$P(C|F_1, .., F_n) = \frac{1}{Z}P(C)\prod_{i=1}^{n} P(F_i|C)$$

où (appelé « évidence ») est un facteur d'échelle qui dépend uniquement de F1,...,Fn, à savoir une constante dans la mesure où les valeurs des variables caractéristiques sont connues.
Les modèles probabilistes ainsi décrits sont plus faciles à manipuler, puisqu'ils peuvent être factorisés par l'antérieure P(C) (probabilité apriori de C) et les lois de probabilité indépendantes P(Fi|C).
S'il existe k classes pour C et si le modèle pour chaque fonction P(Fi|C =c) peut être exprimé selon r paramètres, alors le modèle bayésien naïf correspondant dépend de (k-1) + nrk paramètres.
En effet, pour C = c et un i donné, p(Fi|C) nécessite r paramètres,donc nr paramètres pour tous les F ind, et nrk paramètres pour toutes les classes C=c. Reste à déterminer les paramètres .
On peut en fixer (k-1), sachant que $\sum_c P(C = c) = 1$

## 2.3 Classification naïve bayésienne vs régression logistique

La classification ou régression logistique est une autre méthode classification, également très utilisée en machine learning. Ce modèle est facilement interprétable, à l'instar de la classification naïve bayésienne. Ces deux modèles ont pour point commun d'être des classifieurs

linéaires.
Concrètement, la régression logistique est modèle linéaire généralisé recourant à une fonction logistique pour faire le lien entre un ensemble de variables (par exemple des médicaments prescrits à certaines doses) et une variable qualitative (l'état de santé de patients dans notre cas). Après avoir ingéré un historique de données, le modèle a pour objectif de prédire la probabilité qu'un événement survienne, soit la guérison dans notre exemple. Quand la valeur prédite est supérieure à un certain seuil (toujours situé entre 0 et 1), cela signifie que l'événement peut se produire.

# Chapitre 3

# Exemple d'application

## 3.1 Classification multi-classes avec Naive Bayes

Pour mieux comprendre le Naive Bayes Classifier, déroulons le sur un exemple simple.

Note : l'exemple ci-dessus est inspiré d'une discussion sur le site StackOverflow.

Supposons qu'on ait un jeu de données sur 1000 fruits. On dispose de trois types : Banane, Orange, et "autre". Pour chaque fruit, on a 3 caractéristiques :

- ◇ Si le fruit est long ou non
- ◇ S'il est est sucré ou non
- ◇ Si sa couleur est jaune ou non

Notre jeu de données se présente comme suit :

| Type | Long | petit ( ¬ long) | sucré | ¬ sucré | Jaune | Pas jaune | Total |
|---|---|---|---|---|---|---|---|
| Banane | 400 | 100 | 350 | 150 | 450 | 50 | 500 |
| Orange | 0 | 300 | 150 | 150 | 300 | 0 | 300 |
| Autre fruit | 100 | 100 | 150 | 50 | 50 | 150 | 200 |
| Total | 500 | 500 | 650 | 350 | 800 | 200 | 1000 |

FIGURE 3.1 – jeu de donnees

Note :
$lesymbole\neg signifielangation(\neg froid = chaud)$

L'idée du jeu est de prédire le type d'un fruit (orange, banane ou autre) qu'on n'a pas encore vu. Ceci en se basant sur ses caractéristiques.
Supposons que quelqu'un nous demande de lui donner le type d'un fruit qu'il a. Ses caractéristiques sont les suivantes :

— Il est jaune
— Il est long
— Il est sucré

Pour savoir s'il s'agit d'une banane, ou d'une orange ou d'un autre fruit, il faut qu'on calcule les trois probabilités suivantes :

— P(Banane|Long,jaune,sucre) :La probabilité qu'il s'agisse d'une banane sachant que le fruit est long, jaune et sucré.

— P(Orange|Long,jaune,sucre) :La probabilité qu'il s'agisse d'une orange sachant que le fruit est long, jaune et sucré.

— P(Autre fruit|long,jaune,sucre) :La probabilité qu'il s'agisse d'un autre fruit sachant que ce dernier est long, jaune et sucré.

Le type du fruit "inconnu" qu'on cherche à classifier sera celui où on a **la plus grande probabilité.**
Selon la formule de Bayes, on a :

$$P(Banane|Long, jaune, sucre) = \frac{P(Long|Banane)*P(Sucre|Banane)*P(Jaune|Banane)*P(Banane}{P(Long)*P(Sucre)*P(jaune)}$$

Avec notre jeu de données, on peut calculer facilement un certain nombre de probabilités :

$$P(Banane) = \frac{cardinale(Banane)}{cardinale(Touslesfruits)} = \frac{50}{100} = 0.5$$

De même :

— P(Orange)= 0.3
— P(Autre fruits)= 0.2
— P(Long) = 0.5
— P(Sucre)= 0.65
— P(Jaune)= 0.8

Calculons maintenant le terme : P(Long | Banane) : Probabilité que le fruit est long sachant qu'il s'agit d'une banane.

$$P(Long|Banane) = \frac{cardinale(BananeEtLong)}{cardinale(Banane)} = \frac{400}{500} = 0.8$$

Avec la même logique, on peut calculer les probabilités suivantes :

$\rightarrow P(Sucre|Banane)$
$\rightarrow P(Jaune|Banane)$

Maintenant, qu'on a toutes nos probabilités utiles pour notre calcul, on peut calculer la probabilité suivante :

$$P(Banane|Long, jaune, sucre) = \frac{0.8 * 0.7 * 0.9 * 0.5}{0.5 * 0.65 * 0.8} = \frac{0.252}{0.26} = 0.969$$

Avec la même logique on obtient :

— P(Orange|Long,jaune,sucre)=0

— P(Autre fruit|long,jaune,sucre)=0.072

On remarque que la probabilité que notre fruit soit une banane P(Banane | long, jaune, sucre) est largement plus grande que celle des autres probabilités. **On classifie notre fruit inconnu comme étant une Banane.**

# Chapitre 4

# Cas d'utilisation, Avantages et inconvénients

## 4.1 Cas d'utilisation du classifieur bayesien naif

Les applications courantes de Naive Bayes pour des tâches réelles sont :

- Classification des documents : Cet algorithme peut vous aider à déterminer à quelle catégorie appartient un document donné. Il peut être utilisé pour classer des textes dans différentes langues, genres ou sujets (grâce à la présence de mots-clés).
- Filtrage du spam : Naive Bayes trie facilement le spam à l'aide de mots-clés. Par exemple, dans le spam, vous pouvez voir le mot «Viagra» beaucoup plus souvent que dans le courrier ordinaire. L'algorithme doit être formé pour reconnaître ces probabilités, puis il peut les appliquer efficacement pour le filtrage du spam.
- Analyse des sentiments : En fonction des émotions exprimées par les mots d'un texte, Naive Bayes peut calculer la probabilité qu'elle soit positive ou négative. Par exemple, dans les avis clients, «bon» ou «bon marché» signifie généralement que le client est satisfait. Cependant, Naive Bayes n'est pas sensible au sarcasme.
- Classification d'image : À des fins personnelles et de recherche, il est facile de créer un classificateur bayésien naïf. Il peut être formé pour reconnaître les chiffres écrits à la main ou mettre des images dans des catégories grâce à un apprentissage automatique supervisé.

L'empoisonnement bayésien est une technique utilisée par les spammeurs de courrier électronique pour tenter de réduire l'efficacité des filtres anti-spam qui utilisent la règle de Bayes. Ils espèrent augmenter le taux de faux positifs du filtre anti-spam en transformant des mots auparavant innocents en mots de spam dans une base de données bayésienne. L'ajout de mots qui étaient plus susceptibles d'apparaître dans les e-mails non-spam est efficace contre un filtre bayésien naïf et permet au spam de passer.
Cependant, le recyclage du filtre empêche efficacement tous les types d'attaques. C'est pourquoi Naive Bayes est toujours utilisé pour la détection de spam avec certaines heuristiques, telles que la liste noire.

## 4.2 Avantages et Inconvénients

◇ Avantages

Le classifieur bayésien naïf est un algorithme de classification supervisée qui présente plusieurs avantages. Tout d'abord, sa facilité d'utilisation le rend accessible même aux débutants en apprentissage automatique. L'algorithme est relativement simple à implémenter et à comprendre, ce qui permet une utilisation aisée.
De plus, le classifieur bayésien naïf est très rapide et peut traiter de grandes quantités de données en un temps raisonnable. Cela est particulièrement utile pour les applications où le traitement de données volumineuses est nécessaire.
En outre, le classifieur bayésien naïf ne nécessite pas beaucoup de données d'entraînement pour produire des résultats précis. Cette faible complexité est un avantage pour les cas où il y a peu de données d'entraînement disponibles.
Enfin, le classifieur bayésien naïf présente une bonne performance dans de nombreuses applications, notamment la classification de textes, la reconnaissance d'images, la classification de courriers indésirables (spam) et la catégorisation de documents.

◇ Inconvénients

Quant aux points les plus faibles de Naive Bayes, il fonctionne mieux avec des valeurs catégoriques qu'avec des valeurs numériques. Il assume automatiquement la distribution de la courbe en cloche, ce qui n'est pas toujours correct. De plus, si une variable catégorielle a une catégorie dans l'ensemble de données de test qui n'était pas incluse dans l'ensemble de données d'apprentissage, le modèle lui attribuera une probabilité 0 et ne pourra pas faire de prédiction. C'est ce qu'on appelle le problème de la fréquence zéro .
Pour résoudre ce problème, nous devrons utiliser la technique du lissage. Et, bien sûr, son principal inconvénient est qu'il est rare dans la vraie vie que les événements soient complètement indépendants. Vous devez appliquer d'autres algorithmes pour suivre la dépendance causale.

# Conclusion

En conclusion, le classificateur bayésien naïf est une technique de classification simple mais puissante, basée sur le théorème de Bayes. Il suppose que les caractéristiques d'une instance sont indépendantes les unes des autres et utilise cette hypothèse pour calculer la probabilité qu'une instance appartienne à une classe donnée.

Bien que le classificateur bayésien naïf soit considéré comme naïf en raison de son hypothèse simpliste d'indépendance des caractéristiques, il est souvent utilisé dans des applications pratiques telles que la classification de textes, la reconnaissance d'images et la détection de spam. Il est également rapide à entraîner et à utiliser, ce qui en fait un choix attrayant pour les problèmes de classification de grande taille.

Enfin, le classificateur bayésien naïf est un outil utile pour la classification dans de nombreux domaines, mais il convient de prendre en compte ses limitations et ses hypothèses avant de l'utiliser.