

Université de Genève  
—  
Sciences Informatiques



13X005 Intelligence Artificielle  
Projet – Regression Logistique / Naive Bayes  
[Repository Github: 13X005-AI-Project](#)

Gregory Sedykh, Leandre Catogni, Noah Peterschmitt, Noah Munz

Janvier 2024

# Contents

|  |          |
|--|----------|
| <b>1 – Introduction &amp; Rappels théoriques</b>                 | <b>1</b> |
| 1.1 – Régression Logistique . . . . .                            | 1        |
| 1.2 – Naive Bayes . . . . .                                      | 1        |
| <b>2 – Méthodologie</b>  | <b>2</b> |
| 2.0 – Choix du dataset & outils utilisés . . . . .               | 2        |
| 2.1 – Gradient Descent . . . . .                                 | 3        |
| 2.2 – Régression Logistique . . . . .                            | 4        |
| 2.2.1 – Fonction de coût pour la régression logistique . . . . . | 4        |
| 2.2.2 – Apprentissage . . . . .                                  | 8        |
| 2.2.3 – Prédictions . . . . .                                    | 9        |
| 2.2.4 – Résultats . . . . .                                      | 10       |
| 2.3 – Naive Bayes . . . . .                                      | 11       |
| 2.3.1 – Extraction des distributions . . . . .                   | 11       |
| 2.3.2 – Prédictions . . . . .                                    | 11       |
| 2.3.3 – Résultats . . . . .                                      | 12       |
| 3. – Analyse . . . . .   | 13       |

# Projet – Regression Logistique / Naive Bayes

## 1 – Introduction & Rappels théoriques

Dans ce document, nous approfondirons des techniques de regression logistique et “Naive Bayes” comme outils d’apprentissage supervisés.

Dans le cadre de l’intelligence artificielle et de l’apprentissage supervisé, la compréhension et la classification précises des données revêtent une importance capitale. Parmi les diverses méthodologies existantes, la Régression Logistique et “Naive Bayes” se distinguent par leur efficacité et leur applicabilité dans de nombreux contextes. Ce document se propose d’étudier ces deux techniques, en mettant l’accent sur leur mise en œuvre pratique, et leur efficacité comparative dans divers scénarios.

### 1.1 – Régression Logistique

En statistiques, la régression logistique, s’inscrit dans le cadre des modèles de régression pour les variables binaires. Bien qu’elle soit quasiment exclusivement utilisée en tant que méthode de classification.

En effet, c’est l’ajout d’un seuil, à la probabilité continue donnée par le model de regression qui nous permet de l’utiliser pour la classification.

Ce type de modèle vise à expliquer de manière optimale une variable binaire, qui représente la présence ou l’absence d’une caractéristique spécifique, à l’aide d’un ensemble conséquent de données réelles et d’un modèle mathématique.

Autrement dit, il s’agit de relier une variable aléatoire de Bernoulli, généralement notée  $y$ , aussi appelé “label” à un vecteur constitué de plusieurs variables aléatoires,  $(x_1, \dots, x_K)$ , aussi appelés “features”. [5].

La régression logistique s’appuie sur un classifieur linéaire [1] i.e. un classifieur dont la sortie (pour un vecteur de feature  $x \in \mathbb{R}^n$ ) est donnée par:

$$g(x) = f(\langle w, x \rangle + b)$$

où  $w \in \mathbb{R}^n$  est le vecteur de poids,  $b \in \mathbb{R}$  le biais et  $\langle \cdot, \cdot \rangle$  le produit scalaire usuel.  $f$  est une fonction dite de seuillage qui va séparer nos résultats. Un choix commun pour  $f$  est la sigmoïde ou la fonction signe [1].

Par exemple, dans le cas de la regression logistique binaire, on suppose le modèle suivant:

$$y_i \sim \text{Bernoulli}(p_i), \quad p_i = \sigma(\langle w, x_i \rangle + b), \quad \sigma(z) = \frac{1}{1 + e^{-z}}$$

où  $x_i \in \mathbb{R}^K$  représente un vecteur (ligne) de  $K$  valeurs pour les  $K$  features (aussi appelé un *sample*), et  $y_i$  la variable aléatoire qui représente le label qui leur est associé.

Cependant, dans notre dataset (voir [section 2.0](#)) nous avons 3 classes (3 espèces d’iris),  $y$  ne suit donc, évidemment, plus une loi de Bernoulli.

**À modifier ?** La sigmoïde étant continue, nous avons simplement modifié la manière dont nous lui appliquons le seuillage, pour distinguer 3 cas au lieu de 2. i.e. Au lieu de séparer le domaine en 2 ( $\sigma(z) \leq 0.5$ ,  $\sigma(z) > 0.5$ ), nous l’avons séparé en  $N$  (ici  $N = 3$ ). On a donc que  $y_i = k \Leftrightarrow \frac{k}{N} \leq \sigma(z) < \frac{k+1}{N}$ , ce qui a donné des résultats plus que satisfaisants comme nous le verrons en [section 2.2](#).

### 1.2 – Naive Bayes

“Naive Bayes” se présente comme une méthode de classification probabiliste basée sur le [théorème de Bayes](#), caractérisée par l’adoption d’une hypothèse d’indépendance forte entre les features (attributs), qualifiée de “naïve”.

Plus simplement, le classifieur est classifié de “naïf” car il part du principe que chaque feature (attribut) est indépendante des autres et a un poids égal quant à la probabilité qu’un point appartienne à une classe.

Ce model est dit génératif contrairement à la regression logistique étant considéré comme “méthode discriminante” [1] et consiste à modéliser les probabilités conditionnelles  $P(x|classe)$  pour chaque classe  $y$  et *sample*  $x$  afin de trouver celle qui maximise cette probabilité.



En d'autres termes, le problème revient à trouver, pour des attributs  $x_1, \dots, x_k$ , la classe  $\tilde{y}$  telle que:

$$\tilde{y} = \arg \max_{y \in \mathcal{Y}} \left[ P(y) \prod_{k=1}^K P(x_k|Y) \right]$$

## 2 – Méthodologie

### 2.0 – Choix du dataset & outils utilisés


Pour la suite de ce projet les outils suivants ont été utilisés dans chaque parties:

- [python](#)
- [poetry](#)
- [Make](#)
- [numpy](#)
- [pandas](#)
- [sklearn](#)
- [matplotlib](#)
- [ucmilrepo](#)
- [pytest](#)

Le package `ucmilrepo` a été utilisé pour charger les données de notre dataset depuis la base de donnée du [UC Irvine Machine Learning Repository](#).

Le dataset que nous avons choisi est le fameux dataset “Iris” [3], un des plus anciens et connus dataset de classification. Il contient 150 observations de 3 espèces différentes d’iris (Iris setosa, Iris virginica et Iris versicolor) avec  $K = 4$  features (longueur et largeur des sépales et pétales).

Voici un aperçu des points-clés du dataset:



# Iris

Donated on 6/30/1988

A small classic dataset from Fisher, 1936. One of the earliest known datasets used for evaluating classification methods.

Dataset Characteristics

Tabular

Subject Area

Biology

Associated Tasks

Classification

Variables Table

| Variable Name | Role    | Type        | Demographic | Description   | Units |
|---------------|---------|-------------|-------------|---|-------|
| sepal length  | Feature | Continuous  |             |   | cm    |
| sepal width   | Feature | Continuous  |             |   | cm    |
| petal length  | Feature | Continuous  |             |   | cm    |
| petal width   | Feature | Continuous  |             |   | cm    |
| class         | Target  | Categorical |             | class of iris plant: Iris Setosa, Iris Versicolour, or Iris Virginica |       |

Figure 1: Iris descriptive table

Le label que nous allons prédire sera donc *class*, i.e. l’espèce de l’iris.

## 2.1 – Gradient Descent

Dans cette section, une implémentation de la “descente en gradient” a été réalisée. La fonction a la signature suivante

```
1 def gradient_descent(df, params: NDArray, alpha: float, num_iters: int) -> NDArray:
```

Elle calcule de manière itérative le(s) paramètre(s) `params` qui minimisent la fonction dont `df` est le gradient avec un “taux de convergence” `alpha`.

La fonction a été testé avec la fonction `scipy.optimize.fmin` [6] de la librairie `scipy` sur la fonction suivante:

$$f(x) = x \cdot \cos(\pi(x + 1))$$

avec différents  $x_0 \in \{-\pi, 0, \pi\}$  (valeur initiale de `params`, i.e. `NDArray` avec `D=0`).

Les minimas locaux trouvés par les deux fonctions sont les suivants:

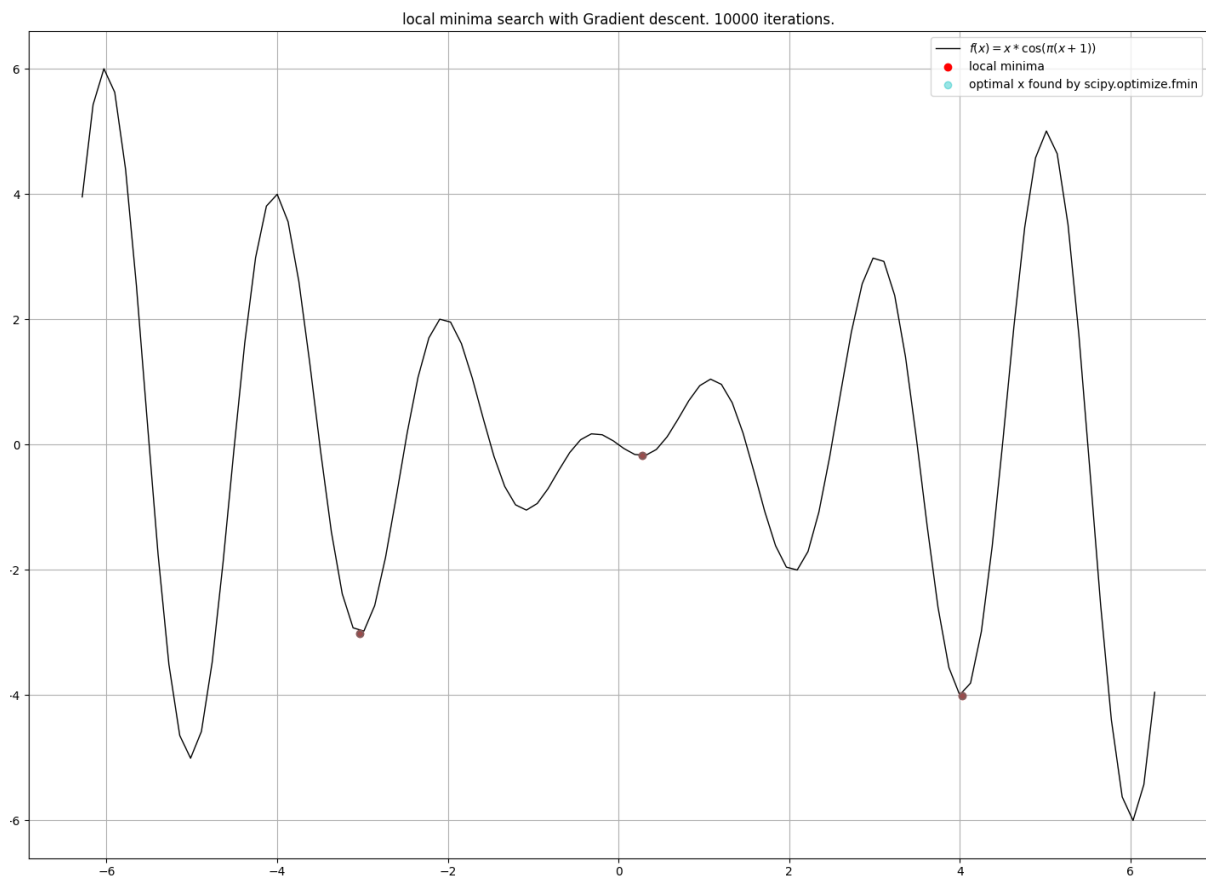


Figure 2: minimas locaux\_gradient descent

Ce résultat illustre bien 2 choses: la première est que l’implémentation de la descente en gradient fonctionne correctement puisque pour chaque points trouvé par notre fonction est confondu avec celui trouvé par la fonction de scipy (c’est ce qui donne cette teinte “grise”). La deuxième est que la “qualité” du minima local (i.e. la distance avec le minima globale) dépend fortement de la valeur initiale et ce pour les deux fonctions.

## 2.2 – Régression Logistique

### 2.2.1 – Fonction de coût pour la régression logistique

Afin d'entraîner les paramètres de la régression logistique, il faut pouvoir comparer les résultats obtenus par la régression avec les résultats attendus.

On souhaite définir une fonction à minimiser permettant de trouver les paramètres optimaux de la régression logistique.

Notre classification se base sur la fonction sigmoïde  $\sigma(z) = \frac{1}{1+e^{-z}}$ .

Comme la fonction exponentielle est toujours positive, on a bien que  $\sigma(z) \in [0, 1]$ .

La fonction sigmoïde nous donne la probabilité que l'élément donné appartienne à un label.

Autrement dit, la fonction sigmoïde est la fonction de répartition de la régression logistique.

Soit  $Y \in \{0, 1\}$  les différents labels que peut prendre l'élément que l'on considère et soit  $X$  l'ensemble des caractéristiques connues de l'élément, dont on cherche à déterminer dans quelle classe le mettre, donc quel label on doit lui attribuer. Soit  $\theta$  le vecteur des poids des covariables, indiquant à quel point les covariables influencent sur la décision du label. On a donc:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(X_1 \cdot w_1 + X_2 \cdot w_2 + \dots + b)}}$$

et

$$P(Y = 0|X) = 1 - \frac{1}{1 + e^{-(X_1 \cdot w_1 + \dots + b)}}$$

Pour plus de simplicité, on va considérer que le biais est compris dans les poids: au lieu d'écrire  $z = wX + b$ , on écrit  $z = \hat{X}\theta$  avec  $\hat{X} = [X \ 1]$  modifié ou on a ajouté une colonne avec que des 1 à la fin de la matrice  $X$  et  $\theta = [w \ b]$  afin d'avoir une bonne cohérence avec le rapport et le code. (On a trouvé cela plus facile d'avoir pour chaque labels les poids et biais sur une ligne, donc d'avoir  $\theta_1$  pour le label 1 etc...) Ainsi, on a:  $\hat{X}\theta^T = X_1 \cdot w_1 + X_2 \cdot w_2 + \dots + b$

Pour la suite, on va noter  $X = \hat{X}$

Notre régression logistique binaire peut donc s'écrire comme:

$$P(Y = 1|X) = \frac{1}{1 + e^{X\theta^T}} = \sigma(X\theta^T)$$

et

$$P(Y = 0|X) = 1 - \sigma(X\theta^T)$$

**2.2.1.1 – Généralisation** On désire donc trouver une nouvelle distribution  $\phi(z)$  tel que:

$$\phi(z) \in [0, 1] \ \forall z$$

est une généralisation de la fonction  $\sigma(z)$

On veut donc que pour une régression logistique binaire, on ait  $\sigma(z) = \phi(z)$ .

On peut remarquer que:

$$\begin{aligned} P(Y = 1|X) &= \frac{1}{1 + e^{-X\theta^T}} \\ &= \frac{1}{1 + e^{-X\theta^T}} \cdot \frac{e^{X\theta^T}}{e^{X\theta^T}} \\ &= \frac{e^{X\theta^T}}{e^{X\theta^T} + e^{X\theta^T - X\theta^T}} \end{aligned}$$

$$\begin{aligned}
 &= \frac{e^{X\theta^T}}{e^{X\theta^T} + e^0} \\
 &= \frac{e^{X\theta^T}}{e^{X\theta^T} + 1}
 \end{aligned}$$

On peut considérer que nous avons un vecteur de poids pour chaque label.

Ainsi, on a  $\theta_0 = [w_0 \ b_0]$  pour le label 0 et  $\theta_1 = [w_1 \ b_1]$  pour le label 1.

Comme on a besoin seulement d'un vecteur de poids pour déterminer le label de nouveaux éléments avec leurs caractéristiques, on peut considérer que  $\theta_0 = [0 \ \dots \ 0]$ .

Ainsi, la formule précédente nous donne:

$$\begin{aligned}
 P(Y = 1|X) &= \frac{e^{X\theta_1^T}}{e^{X\theta_1^T} + 1} \\
 &= \frac{e^{X\theta_1^T}}{e^{X\theta_1^T} + e^0} \\
 &= \frac{e^{X\theta_1^T}}{e^{X\theta_1^T} + e^{0 \cdot X}} \\
 &= \frac{e^{X\theta_1^T}}{e^{X\theta_1^T} + e^{X\theta_0^T}} \\
 &= \frac{e^{X\theta_1^T}}{\sum_{i=0}^1 e^{X\theta_i^T}}
 \end{aligned}$$

On peut donc généraliser cette formule pour  $K$  labels.

Cela nous donne:

$$P(Y = k|X) = \frac{e^{X\theta_k^T}}{\sum_{i=0}^K e^{X\theta_i^T}}$$

Comme la fonction exponentielle est toujours positive, on a bien que:

$$\begin{aligned}
 0 &\leq e^{X\theta_k^T} \leq e^{X\theta_k^T} + \sum_{i \neq k}^K e^{X\theta_i^T} \\
 \Leftrightarrow 0 &\leq e^{X\theta_k^T} \leq \sum_i^K e^{X\theta_i^T} \\
 \Leftrightarrow 0 &\leq \frac{e^{X\theta_k^T}}{\sum_i^K e^{X\theta_i^T}} \leq 1 \\
 \Leftrightarrow 0 &\leq \phi(z) \leq 1
 \end{aligned}$$

De plus, on a que:

$$\begin{aligned}
 &\sum_k^K P(Y = k|X) \\
 &= \sum_k^K \frac{e^{X\theta_k^T}}{\sum_i^K e^{X\theta_i^T}} \\
 &= \frac{\sum_k^K e^{X\theta_k^T}}{\sum_i^K e^{X\theta_i^T}}
 \end{aligned}$$



$$= \frac{\sum_i^K e^{X\theta_i^T}}{\sum_i^K e^{X\theta_i^T}} = 1$$

Donc la fonction  $\phi(z)$  est bien une fonction de distribution de probabilité qui généralise la fonction sigmoïde pour des problèmes à plusieurs labels.

Cette fonction est couramment appelée fonction **softmax**.

**2.2.1.2 – Fonction de coût** Notre objectif est donc de trouver une fonction de coût pour pouvoir entraîner les paramètres de la régression multinomiale. On cherche à maximiser la vraisemblance des données. Donc pour un label  $Y$  donné, on veut maximiser:

$$\sum_k^K f(Y, k) P(Y = k|X)$$

avec  $f(Y, k)$  la fonction qui vaut 1 si  $Y = k$  et 0 sinon.

Comme on a plusieurs couples de données  $(X_i, Y_i)$ , on peut écrire la fonction précédente comme:

$$\sum_i^n \sum_k^K f(Y_i, k) P(Y_i = k|X_i)$$

En maximisant cette fonction, on fait en sorte que le paramètre  $\theta_k$  permette d'obtenir la prédiction que le label soit égal à  $k$  avec la somme des probabilités où  $Y_i = k$  est la plus grande possible.

Afin de pouvoir utiliser un algorithme comme la descente en gradient, il faut non pas maximiser une fonction, mais minimiser une fonction.

Tout d'abord, comme on travaille avec des exponentielles, on a intérêt à prendre un logarithme pour éviter d'avoir à travailler avec de trop grandes valeurs. Cette modification n'aura pas d'impact sur la convexité car la fonction logarithme est une fonction strictement croissante.

Enfin, comme on cherche une fonction à minimiser et non pas à maximiser pour pouvoir utiliser la descente en gradient, on va prendre l'inverse de la fonction.

Cela s'appelle couramment le **negative logarithm likelihood**.

Cela nous donne une fonction de coût comme suit:

$$\begin{aligned} & \sum_i^n \sum_k^K f(Y_i, k) \log\left(\frac{1}{P(Y_i = k|X_i)}\right) \\ &= \sum_i^n \sum_k^K f(Y_i, k) (\log(1) - \log(P(Y_i = k|X_i))) \\ &= - \sum_i^n \sum_k^K f(Y_i, k) \log(P(Y_i = k|X_i)) \end{aligned}$$

On peut minimiser cette fonction de coût grâce à une descente en gradient.

**2.2.1.3 – Dérivée de la fonction de coût** On va calculer la dérivée de la fonction de coût.

On a:

$$\begin{aligned} & \log(P(Y = k|X)) \\ &= \log\left(\frac{e^{X\theta_k^T}}{\sum_i^K e^{X\theta_i^T}}\right) \end{aligned}$$





$$= X\theta_k^T - \log \left( \sum_i^K e^{X\theta_i^T} \right)$$

Donc:

$$\frac{\partial}{\partial \theta_j} \sum_i^K f(Y, i) \log(P(Y = i|X))$$

(NB: On considère que  $Y = k$ , tous les autres termes étant annulés car  $f = 0$ )

$$\begin{aligned} &= \frac{\partial}{\partial \theta_j} f(Y, k) \log(P(Y = k|X)) \\ &= \frac{\partial}{\partial \theta_j} \left( X\theta_k^T - \log \left( \sum_i^K e^{X\theta_i^T} \right) \right) \end{aligned}$$

Supposons que  $j = k$ .

$$\begin{aligned} &= X - \frac{\partial}{\partial \theta_j} \log \left( \sum_i^K e^{X\theta_i^T} \right) \\ &= X - \frac{1}{\sum_i^K e^{X\theta_i^T}} \frac{\partial}{\partial \theta_j} \sum_i^K e^{X\theta_i^T} \\ &= X - \frac{1}{\sum_i^K e^{X\theta_i^T}} \frac{\partial}{\partial \theta_j} e^{X\theta_j^T} \\ &= X - \frac{X e^{X\theta_j^T}}{\sum_i^K e^{X\theta_i^T}} \\ &= X - X P(Y = j|X) \\ &= X(1 - P(Y = j|X)) \end{aligned}$$

Supposons que  $j \neq k$ .

$$\begin{aligned} &\frac{\partial}{\partial \theta_j} \left( X\theta_k^T - \log \left( \sum_i^K e^{X\theta_i^T} \right) \right) \\ &= -\frac{\partial}{\partial \theta_j} \log \left( \sum_i^K e^{X\theta_i^T} \right) \\ &= -\frac{1}{\sum_i^K e^{X\theta_i^T}} \frac{\partial}{\partial \theta_j} \sum_i^K e^{X\theta_i^T} \\ &= -\frac{1}{\sum_i^K e^{X\theta_i^T}} \frac{\partial}{\partial \theta_j} e^{X\theta_j^T} \\ &= -\frac{X e^{X\theta_j^T}}{\sum_i^K e^{X\theta_i^T}} \\ &= -X P(Y = j|X) \end{aligned}$$

On a donc:

$$\frac{\partial}{\partial \theta_j} \sum_i^K f(Y, i) \log(P(Y = i|X)) = X(f(Y, j) - P(Y = j|X))$$

car  $f(Y, k)$  est égal à 1 si  $Y = k$  et 0 sinon.

Donc pour  $n$  données, cela nous donne:

$$\frac{\partial}{\partial \theta_j} \sum_m^n \sum_i^K f(Y_m, i) \log(P(Y_m = i|X_m)) = \sum_m^n X_m(f(Y_m, j) - P(Y_m = j|X_m))$$

Maintenant, on est prêt pour entraîner notre régression logistique multinomiale !



## 2.2.2 – Apprentissage

Maintenant que nous avons une fonction de coût permettant de quantifier (en moyenne) à quel point un set de  $N$  prédictions est correct/incorrect à un point de l'apprentissage donné. Il ne reste plus qu'à chercher les paramètres optimaux qui minimisent cette fonction de coût. Ce que l'on va réaliser à l'aide de la descente en gradient. C'est le processus d'apprentissage.

En effet, lors de l'apprentissage, on va chercher de manière itérative les  $\mathbf{w}$  et  $b$  qui respectent les critères mentionnés ci-dessus en calculant le gradient de la fonction de coût à chaque itérations et en allant dans la direction opposé.

Concrètement cela revient à appliquer l'algorithme suivant:

---

### Algorithm 1 gradient descent

---

```

function GRADIENTDESCENT( $f, \mathbf{w}_{init}, b_0, \alpha, \text{num\_iters}$ )
     $\mathbf{w} \leftarrow \mathbf{w}_{init}$ 
     $b \leftarrow b_0$ 
    for 1 to num_iters do
         $\mathbf{dw}, db \leftarrow \nabla f(w, b)$ 
         $\mathbf{w} \leftarrow \mathbf{w} - \alpha \cdot \mathbf{dw}$ 
         $b \leftarrow b - \alpha \cdot db$ 
    end for
    return  $w, b$ 
end function

```

---

En pratique, il est plus simple de passer directement la fonction qui calcul le gradient en argument, que d'essayer de le calculer dynamiquement, c'est pourquoi la signature de notre implémentation prend un  $\mathbf{df}$  en argument plutôt que la fonction de coût elle même.

Où le calcul des dérivées partielles a été défini comme ci-dessous.

Soit  $\nabla C(\mathbf{w}, b) = (\frac{\partial C(\mathbf{w}, b)}{\partial \mathbf{w}}, \frac{\partial C(\mathbf{w}, b)}{\partial b})$ , pour un sample  $\mathbf{x}_i$  et sa classe  $y_i$ , on obtient:

$$\frac{\partial \log(y_i | \mathbf{x}_i; \mathbf{w}, b)}{\partial b} = y_i - \sigma(z_i) = y_i - \sigma(\mathbf{w}^T X_i + b)$$

$$\frac{\partial \log(y_i | \mathbf{x}_i; \mathbf{w}, b)}{\partial w_j} = x_{ij} \cdot (y_i - \sigma(z_i)) = (y_i - \sigma(\mathbf{w}^T X_i + b)) \cdot x_{ij}$$

Or le  $\mathbf{db}$  dans l'algorithme ci-dessus se réfère à la moyenne (pour tout  $i$ ) de ces valeurs (i.e. distance moyenne *classes prédites* – *“vrai” classes*).

On l'obtient donc comme suit: (la somme des dérivées est la dérivée de la somme, linéarité de la dérivée)

$$\nabla_b C = \frac{1}{N} \sum_{i=1}^N \frac{\partial \log(y_i | \mathbf{x}_i; \mathbf{w}, b)}{\partial b} = \frac{1}{N} \sum_{i=1}^N y_i - \sigma(\mathbf{w}^T X_i + b)$$

De même pour  $\mathbf{dw}$ :

$$\begin{aligned} \nabla_{\mathbf{w}} C &= \frac{1}{N} \sum_{i=1}^N (x_{ij}(y_i - p_i))_{1 \leq j \leq k} = \frac{1}{N} \sum_{i=1}^N (y_i - \sigma(z_i)) \cdot (x_{ij})_{1 \leq j \leq k} \\ &= \frac{1}{N} \sum_{i=1}^N (y_i - \sigma(\mathbf{w}^T \mathbf{x}_i + b)) \mathbf{x}_i \end{aligned}$$

On retrouve ainsi, le calcul effectué dans la fonction `grad` de `log_reg.py` de signature suivante:

---

```

1  def grad(X: NDArray, y: NDArray, w: NDArray, b: float) -> tuple:

```

---

Etant donné que pour le calcul du gradient il est nécessaire d'avoir une matrice de feature  $X$  et vecteur de label  $y$ , une version “modifiée” de la descente en gradient a été implémenté.

---

```
1 def grad_desc_ml(features: NDArray, labels: NDArray, df, w: NDArray, b: float, alpha: float,
    num_iters: int) -> tuple[NDArray, float]:
```

---

Cette fonction se comporte exactement de la même manière que celle décrite en [section 2.1](#). La seule différence est qu'elle passe `features` et `labels` comme `X` et `y` à la fonction `df` (dans notre cas `df` est toujours la fonction `grad`), i.e. on a `df(features, labels, w, b)` au lieu de `df(params)`.

### 2.2.3 – Prédiction

Pour la prédiction, nous avons utilisé la fonction suivante:

---

```
1 def predict_log_reg(X: NDArray, w: NDArray, b):
```

---

qui prend simplement  $\sigma(w^T X + b)$  et seuil la sortie du sigmoïde de manière à retourner un nombre entre 0 et 2 (avec les poids et biais entraînés).

## 2.2.4 – Résultats

Suite à l'apprentissage, nous avons obtenu les résultats suivants:

$$w = [0.53452349, 0.36463584, 1.16132476, 1.08204578]$$

$$b = 0.45146791$$

N.B.:

L'apprentissage peut être ré-effectué de manière efficient si besoin est à l'aide du jupyter notebook [training\\_test.ipynb](#) disponible sur la branche [gpu-training](#) du repository github. Le code de l'entraînement (uniquement sur cette branche) a été "porté" sur cuda / gpgpu à l'aide de la librairie [cupy](#) [2].

A noter qu'il utilise des fonctions des sklearn alors que nous devions les implémenter nous mêmes, (telles que les metrics f1-score...). Ces fonctions ont bien été implémenté mais pour une raison de simplicité, elle n'ont pas été utilisée pour l'entraînement. Le code de cette branche ne fera donc pas partie du rendu mais reste publiquement accessible sur github.

Comme dit en section 1.1, ces paramètres sont, en effet, plus que satisfaisants, comme on peut le voir sur l'output de `pytest` suivant:

```
1 src/log_reg.py::test_log_reg_f1score
2 weights & biases: [0.53452349, 0.36463584, 1.16132476, 1.08204578], 0.45146791
3 { 'accuracy': 1.0, 'f1_score': 1.0, 'precision': 1.0, 'recall': 1.0 }
4 PASSED
5
6 src/naive_bayes.py::test_predict_bayes_f1score_all
7 { 'accuracy': 0.97, 'f1_score': 0.975, 'precision': 0.976, 'recall': 0.974 }
8 PASSED
```

NB: pour reproduire cette output, lancer `make test_model`.

Ce résultat a été obtenu avec une séparation 70/30 de training/test data. Lorsque l'on essaye de changer la portion qui est prise aléatoirement dans chaque catégorie, on obtient un F1-score qui varie entre 0.93 et 1.0 (avec, dans de rares exceptions 0.91 ou 0.89).

De plus, l'on voit que les performances que nous avons obtenus rentrent tout à fait dans le cadre de celles annoncées par le UCI ML Repository:

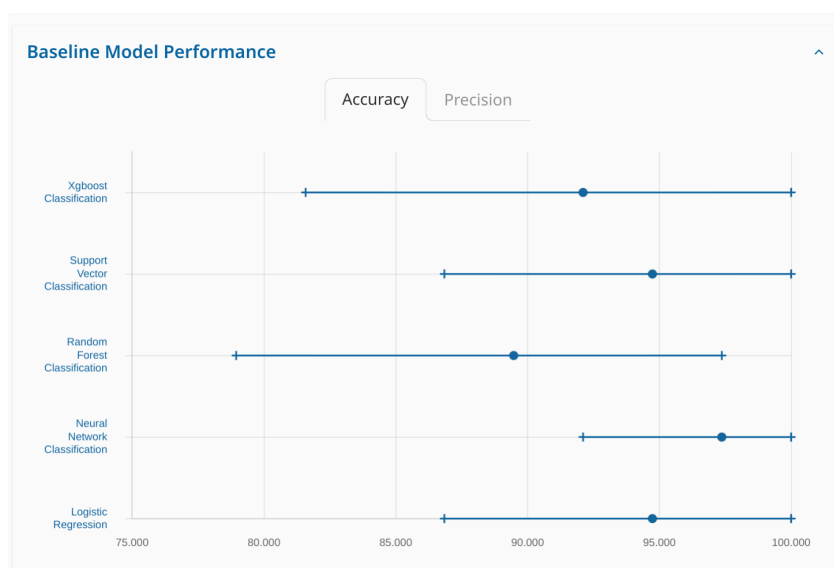


Figure 3: performances attendu d'après le UCI ML Repo [4]

Ce résultat illustre bien que notre démarche est correcte et que nos 2 modèles sont efficaces, avec un penchant pour la régression logistique qui semble être plus efficace que Naive Bayes.

## 2.3 – Naive Bayes

Dans cette section, une implémentation d'un classifieur linéaire bayésien (naive bayes) a été réalisée.

### 2.3.1 – Extraction des distributions

Dans cette implémentation, étant données que toutes nos features sont continues, nous avons considéré que *sepal length*, *sepal width*, *petal length* et *petal width* seront représenté comme 4 variables aléatoires  $X_0, \dots, X_3$  suivant 4 lois normales normales de paramètre  $(\mu_k, \sigma_k)$ .

C'est à dire:

$$X_k \sim \mathcal{N}(\mu_k, \sigma_k) \quad k \in \llbracket 0, 3 \rrbracket$$

Elles peuvent être récupérées à l'aide de la fonction suivante:

---

```
1 def get_distrib_parameters(features: DataFrame, labels) -> dict[Any, list[tuple[f1, f1]]]:
```

---

qui va retourner un dict mappant chaque classe à une liste contenant les paramètres des distributions conditionnelles (normales) des features pour cette classe.

### 2.3.2 – Prédiction

Deux fonctions de prédictions ont été implémenté,

1. Prennant un sample et prédisant sa classe
2. Une deuxième qui prend tous les samples et applique, en parallèle, la première fonction à chacun d'eux.

Elles ont les signatures suivantes:

---

```
1 def predict_bayes(x, params_by_class: dict[Any, list[tuple[f1, f1]]]) -> Any:
2 def predict_bayes_all(X: DataFrame, params_by_class: dict[Any, list[tuple[f1, f1]]] | None
   = None) -> list[Any]:
```

---

Comme dit précédemment, pour pouvoir prédire la classe d'un sample, il faut calculer les probabilité conditionnelle  $P(\mathbf{x}|classe)$  pour chaque classe  $y$  et sample  $\mathbf{x}$  et prendre la classe qui maximise cette dernière.

Cela revient à chercher le  $\tilde{y}$  défini en [section 1.2](#), développons le calcul qui nous amené à cette formule:

$$\tilde{y} = \arg \max_{y \in \mathcal{Y}} P(y|\mathbf{x}) = \arg \max_{y \in \mathcal{Y}} \frac{P(\mathbf{x}|y)P(y)}{P(\mathbf{x})} = \arg \max_{y \in \mathcal{Y}} P(\mathbf{x}|y)P(y)$$

Or

$$P(\mathbf{x}|y) = P(x_1|y) \prod_{i=2}^n P(x_i|x_{i-1}, \dots, x_1, y)$$

Avec l'hypothèse que les  $\{X_i\}_{i \leq n}$  sont indépendants, on obtient que:

$$P(x_i|x_{i-1}, \dots, x_1, y) = P(x_i|y)$$

Donc

$$P(\mathbf{x}|y) = P(x_1|y) \prod_{k=2}^K P(x_k|y) = \prod_{k=1}^K P(x_k|y)$$

En conclusion:

$$\tilde{y} = \arg \max_{y \in \mathcal{Y}} \left[ P(y) \prod_{k=1}^K P(x_k|y) \right]$$

(où  $K$  reste le nombre de features.)



Où au début on cherche à maximiser  $P(y|x)$  car idéalement on voudrait savoir la probabilité que  $y$  soit le bon label pour n'importe quel sample  $\mathbf{x}$ . Cependant, on aimerait pouvoir effectuer cette prédictions pour des  $\mathbf{x}$  qui n'appartiennent pas à notre dataset d'apprentissage, i.e. on ne doit pas avoir besoin d'avoir déjà vu exactement ce sample. On a donc besoin d'une généralisation, c'est ainsi que l'on fini par retomber sur

$$\tilde{y} = \arg \max_{y \in \mathcal{Y}} \left[ P(y) \prod_{k=1}^K P(x_k|y) \right]$$

qui est ce que calculent les fonctions dont on a donné la signature ci-dessus.

### 2.3.3 – Résultats

Dans cette section, nous allons simplement reprendre ce qui a été fait dit dans la [section 2.2.4](#) et remonter les mêmes tests.

Voici l'output du test `pytest` pour les rapports de performances du model bayesien:

---

```
1 src/log_reg.py::test_log_reg_f1score
2 weights & biases: [0.53452349, 0.36463584, 1.16132476, 1.08204578], 0.45146791
3 { 'accuracy': 1.0, 'f1_score': 1.0, 'precision': 1.0, 'recall': 1.0 }
4 PASSED
5
6 src/naive_bayes.py::test_predict_bayes_f1score_all
7 { 'accuracy': 0.97, 'f1_score': 0.975, 'precision': 0.976, 'recall': 0.974 }
8 PASSED
```

---

Ce résultat a été obtenu avec une séparation 70/30 de training/test data.

Ces résultats illustrent bien que notre démarche est correcte et que nos 2 modèles sont efficaces, avec un penchant pour la régression logistique qui semble être plus efficace que Naive Bayes.

Cependant, un f1-score de  $> 0.95$  reste excellent.

### 3. – Analyse

Pour chaque classe  $y$ , on peut tracer les fonctions de distribution de probabilité pour chaque donnée  $X_k$  sachant la classe  $y$  afin d'analyser la structure des données.

Pour la classe  $Y=0$ , on obtient le graphe suivant :

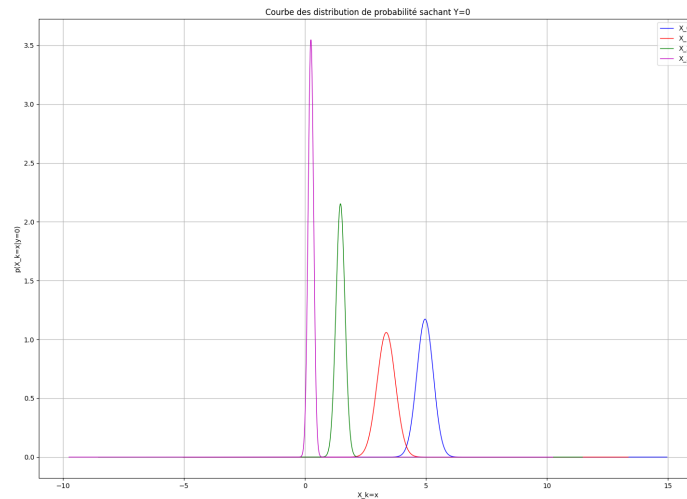


Figure 4: graphe des fonctions de distribution sachant  $Y=0$

On peut voir tout d'abord que pour cette classe, les pics des courbes bleue et rouge sont bien inférieurs aux pics des courbes vert et magenta. Ainsi, on en conclut que les variables  $X_0$  et  $X_1$  ont moins d'influence dans la prédiction de cette classe. Alors que le pic de la courbe magenta est bien supérieur aux autres, indiquant que la variable  $X_3$  a une forte influence sur la prédiction de cette classe. De plus, on observe que seules les courbes bleue et rouge ont un chevauchement perceptible mais quand même assez petit, on en conclut que les variables sont pour cette classe très peu indépendantes les unes des autres.

Pour la classe  $Y=1$ , on obtient le graphe suivant :

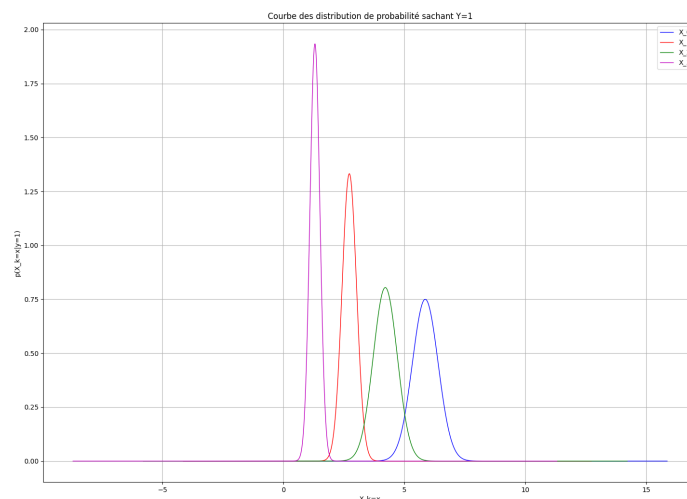


Figure 5: graphe des fonctions de distribution sachant  $Y=1$

On peut voir tout d'abord que pour cette classe, les pics des courbes bleue et verte sont bien inférieurs aux pics des courbes rouge et magenta. Ainsi, on en conclut que les variables  $X_0$  et  $X_2$  ont moins d'influence dans la prédiction de cette classe. Alors que le pic de la courbe magenta est bien supérieur aux autres, indiquant que la variable  $X_3$  a une forte influence sur la prédiction de cette classe. De plus, on observe que les courbes rouge et magenta ont un faible chevauchement indiquant une faible interdépendance entre

$X_3$  et  $X_1$  alors que les courbes rouge et verte ainsi que verte et bleue ont un chevauchement assez élevé montrant une certaine interdépendance entre les variables  $X_1$  et  $X_2$  ainsi qu'entre les variables  $X_0$  et  $X_2$ .

Enfin pour la classe  $Y=2$ , on obtient le graphe suivant :

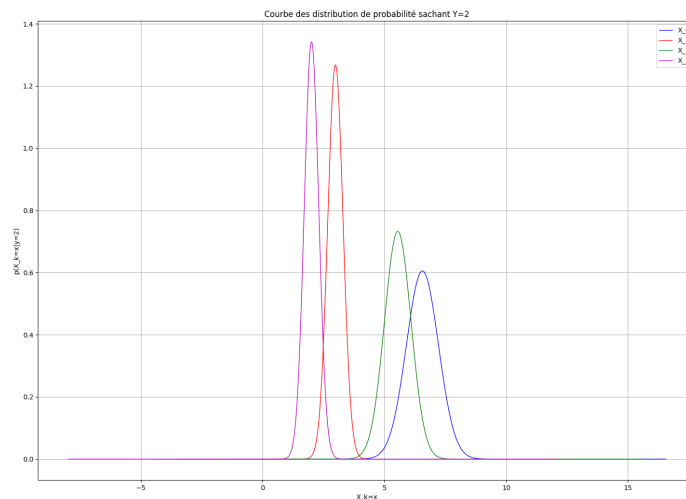


Figure 6: graphe des fonctions de distribution sachant  $Y=0$

On observe que les pics des courbes rouge et magenta sont presque deux fois plus grand que ceux des courbes bleue et verte, de plus les courbes rouge se chevauchent fortement. Ainsi les variables  $X_3$  et  $X_1$  ont une très forte influence sur cette classe et sont assez interdépendant alors que les variables  $X_0$  et  $X_2$  ont très peu d'impacte sur la prédiction de cette classe. Les courbes bleue et verte se chevauchent aussi énormément montrant aussi une forte interdépendance entre les variables  $X_0$  et  $X_2$ .

Ainsi on peut remarquer que globalement la variable  $X_3$  a une forte influence sur la classification alors que la variable  $X_0$  a une plus faible. De plus, les variables indépendantes ne sont pas séparables les unes des autres.

## Références

- [1] *Classifieur linéaire*. In: *Wikipédia*. URL: [https://fr.wikipedia.org/w/index.php?title=Classifieur\\_lin%C3%A9aire&oldid=189518969](https://fr.wikipedia.org/w/index.php?title=Classifieur_lin%C3%A9aire&oldid=189518969) (visited on 01/05/2024).
- [2] *NumPy & SciPy for GPU*. CuPy. URL: <https://cupy.dev/> (visited on 01/11/2024).
- [3] R. A. Fisher. *Iris*. 1936. DOI: [10.24432/C56C76](https://doi.org/10.24432/C56C76). URL: <https://archive.ics.uci.edu/dataset/53> (visited on 01/07/2024).
- [4] R. A. Fisher. *Iris*. 1936, *UCI ML repository website*. URL: <https://archive.ics.uci.edu/dataset/53/iris>.
- [5] *Régression logistique*. In: *Wikipédia*. URL: [https://fr.wikipedia.org/w/index.php?title=R%C3%A9gression\\_logistique&oldid=210479759](https://fr.wikipedia.org/w/index.php?title=R%C3%A9gression_logistique&oldid=210479759) (visited on 01/05/2024).
- [6] *Scipy.Optimize.Fmin — SciPy v1.11.4 Manual*. URL: <https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.fmin.html#scipy.optimize.fmin>.