La Regression Logistique

HENNANI AYOUB BRAHIM
TSOULI ACHRAF
ZAKRITI Al Hassan

Filière d'ingénieur | Génie Mathématique et Informatique
Presenté devant: M. AZOUAGH Nabil

14 Décembre 2024

Introduction

Définition

La régression logistique est une méthode statistique largement utilisée pour modéliser la relation entre une ou plusieurs variables explicatives (indépendantes) et une variable réponse (dépendante) de nature qualitative.

Types de régression logistique

- Régression logistique binaire
- Régression logistique multinomiale
- Régression logistique ordinale

Types de régression logistique

- Régression logistique binaire
- Régression logistique multinomiale
- Régression logistique ordinale

Types de régression logistique

- Régression logistique binaire
- Régression logistique multinomiale
- Régression logistique ordinale

Les Hypothèses

Hypothèses de la régression logistique

- Indépendance des observations.
- Relation linéaire entre les caractéristiques et la variable logit.
- Absence de multicolinéarité.
- Taille de léchantillon.

Les Hypothèses

Hypothèses de la régression logistique

- Indépendance des observations.
- Relation linéaire entre les caractéristiques et la variable logit.
- Absence de multicolinéarité.
- Taille de léchantillon.

Les Hypothèses

Hypothèses de la régression logistique

- Indépendance des observations.
- Relation linéaire entre les caractéristiques et la variable logit.
- Absence de multicolinéarité.
- Taille de léchantillon.

Les Hypothèses

Hypothèses de la régression logistique

- Indépendance des observations.
- Relation linéaire entre les caractéristiques et la variable logit.
- Absence de multicolinéarité.
- Taille de léchantillon.

Définition

La fonction logistique, aussi appelée fonction sigmoïde, est au cur de la régression logistique. Elle est définie comme suit :

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Graphique de la fonction sigmoïde

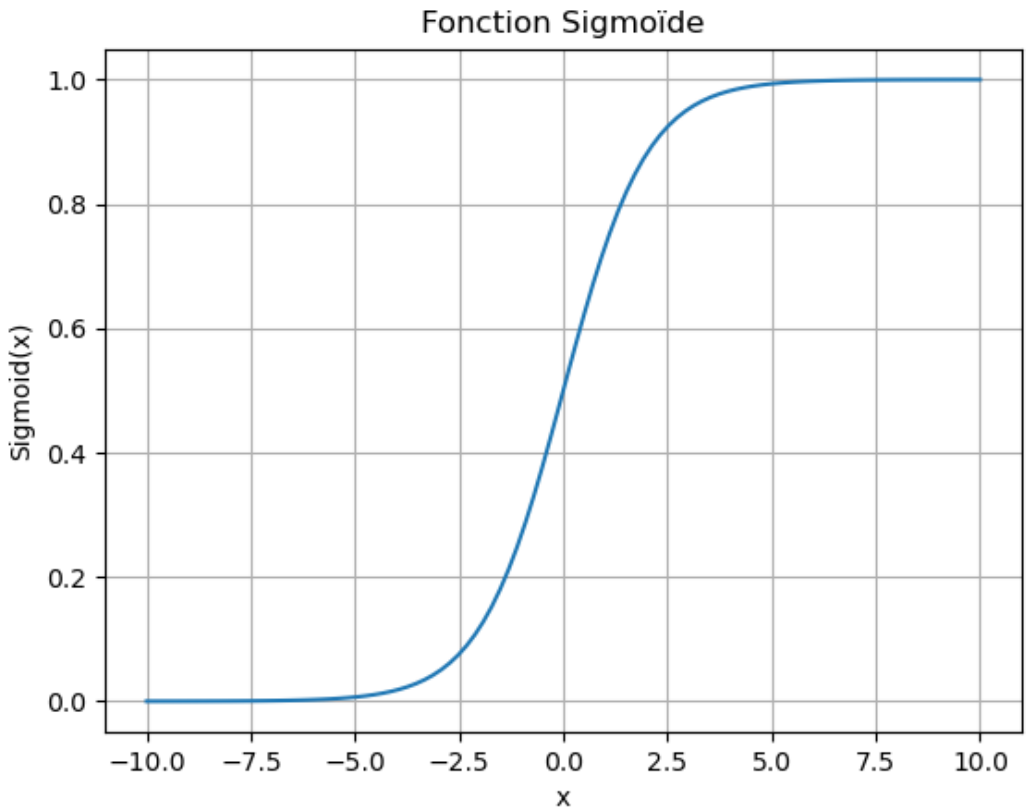

Figure 1: Graphique de la fonction sigmoïde

Modèle mathématique

Modèle mathématique

La probabilité conditionnelle $P(Y = 1 \mid X)$ est modélisée comme suit :

$$P(Y = 1 \mid X) = \pi(X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}}$$

Le *logit*, ou logarithme du rapport des chances, est exprimé comme :

$$\text{logit}(\pi(X)) = \ln\left(\frac{\pi(X)}{1-\pi(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Hypothèses du modèle

Hypothèses du modèle

- Les observations sont indépendantes.
- La relation entre le *logit* de $\pi(X)$ et les variables explicatives X est linéaire.
- Les variables explicatives X peuvent être continues, catégoriques, ou les deux.

Hypothèses du modèle

Hypothèses du modèle

- Les observations sont indépendantes.
- La relation entre le *logit* de $\pi(X)$ et les variables explicatives X est linéaire.
- Les variables explicatives X peuvent être continues, catégoriques, ou les deux.

Hypothèses du modèle

Hypothèses du modèle

- Les observations sont indépendantes.
- La relation entre le *logit* de $\pi(X)$ et les variables explicatives X est linéaire.
- Les variables explicatives X peuvent être continues, catégoriques, ou les deux.

Estimation des paramètres

Estimation des paramètres

Les coefficients $\beta_0, \beta_1, \ldots, \beta_p$ sont estimés par la méthode de la **vraisemblance maximale**, qui maximise la probabilité des données observées.

La fonction de vraisemblance pour n observations est donnée par :

$$L(\beta) = \prod_{i=1}^{n} \pi(X_i)^{y_i} [1 - \pi(X_i)]^{1-y_i}$$

En prenant le logarithme de la vraisemblance, on obtient :

$$\ell(\beta) = \sum_{i=1}^{n} [y_i \ln(\pi(X_i)) + (1 - y_i) \ln(1 - \pi(X_i))]$$

Estimation des paramètres

Estimation des paramètres

Les coefficients $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ sont estimés en maximisant $\ell(\beta)$. Les dérivées partielles pour chaque paramètre β_j $(j = 0, 1, \ldots, p)$ sont :

$$\frac{\partial \ell(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} [y_i - \pi(X_i)] X_{ij}$$

La fonction Log-Loss

La fonction Log-Loss

La fonction de log-loss pour une régression logistique binaire est donnée par :

$$\text{Log-Loss}_{\text{binaire}} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \ln(\pi(X_i)) + (1-y_i)\ln(1-\pi(X_i))\right]$$

où :

- $y_i \in \{0,1\}$ est la classe réelle pour l'observation i.
- $\pi(X_i) = P(Y=1 \mid X_i)$ est la probabilité prédite pour $y_i = 1$.
- n est le nombre total d'observations.

Modèle mathématique

Modèle mathématique

Pour chaque catégorie k $(k = 1, 2, \ldots, K)$, on modélise la probabilité conditionnelle $P(Y = k \mid X)$ comme :

$$P(Y = k \mid X) = \begin{cases} \dfrac{e^{\beta_{k0}+\beta_{k1}X_1+\beta_{k2}X_2+\cdots+\beta_{kp}X_p}}{\sum_{j=1}^{K} e^{\beta_{j0}+\beta_{j1}X_1+\beta_{j2}X_2+\cdots+\beta_{jp}X_p}}, & \text{pour } k = 2, \ldots, K \\ \dfrac{1}{\sum_{j=1}^{K} e^{\beta_{j0}+\beta_{j1}X_1+\beta_{j2}X_2+\cdots+\beta_{jp}X_p}}, & \text{pour } (k = 1). \end{cases}$$

Estimation des paramètres

Fonction de vraisemblance

La fonction de vraisemblance pour n observations est :

$$L(\beta) = \prod_{i=1}^{n} \prod_{k=1}^{K} \pi_k(X_i)^{\mathbb{1}(y_i=k)}$$

où $\mathbb{1}(y_i = k)$ est une fonction indicatrice.

Estimation des paramètres

Log-vraisemblance

La log-vraisemblance associée est donnée par :

$$\ell(\beta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{1}(y_i = k) \ln(\pi_k(X_i))$$

Estimation des paramètres

Estimation des paramètres

Les coefficients β_{kj} sont estimés en maximisant $\ell(\beta)$. Les dérivées partielles par rapport aux paramètres sont :

$$\frac{\partial \ell(\beta)}{\partial \beta_{kj}} = \sum_{i=1}^{n} [\mathbb{1}(y_i = k) - \pi_k(X_i)] X_{ij}$$

Ces équations sont résolues numériquement en utilisant des algorithmes comme Newton-Raphson.

Estimation des paramètres

Probabilités prédites

Après estimation des coefficients, les probabilités pour chaque catégorie sont calculées comme suit :

$$\pi_k(X) = \frac{e^{\beta_{k0}+\beta_{k1}X_1+\cdots+\beta_{kp}X_p}}{\sum_{j=1}^{K} e^{\beta_{j0}+\beta_{j1}X_1+\cdots+\beta_{jp}X_p}}$$

La fonction Log-Loss

Fonction Log-Loss

La fonction de log-loss pour une régression logistique multinomiale est donnée par :

$$\text{Log-Loss}_{\text{multinomiale}} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{1}(y_i = k)\ln(\pi_k(X_i))$$

où :

- $y_i \in \{1, 2, \ldots, K\}$ est la classe réelle pour l'observation i.
- $\pi_k(X_i) = P(Y = k \mid X_i)$ est la probabilité prédite pour la classe k.
- $\mathbb{1}(y_i = k)$ est une fonction indicatrice égale à 1 si $y_i = k$, sinon 0.

Matrice de confusion

Matrice de confusion

La matrice de confusion compare les prédictions du modèle avec les valeurs réelles pour évaluer la classification. Elle contient quatre éléments principaux :

- **Vrais positifs (TP)** : Nombre d'observations correctement prédites comme positives.
- **Faux positifs (FP)** : Nombre d'observations incorrectement prédites comme positives.
- **Vrais négatifs (TN)** : Nombre d'observations correctement prédites comme négatives.
- **Faux négatifs (FN)** : Nombre d'observations incorrectement prédites comme négatives.

Préparation des données

La préparation des données est une étape cruciale dans le processus de modélisation. Elle garantit que les données sont dans un format approprié pour l'algorithme de régression logistique et améliore la performance du modèle. Cela comprend :

- **Normalisation**
- **Encodage des variables catégorielles**

Choix des variables

Le choix des variables explicatives est essentiel pour obtenir un modèle performant. Les caractéristiques doivent être choisies en fonction de leur pertinence pour prédire la variable cible. Les étapes à suivre comprennent :

- **Sélection des variables**
- **Importance des caractéristiques**

Validation

La validation croisée est une technique utilisée pour évaluer la performance d'un modèle et éviter le surapprentissage (overfitting). Voici les étapes essentielles pour une validation croisée efficace :

- **Séparation des données**
- **Validation croisée**

Contexte Generale

Contexte Generale

Ce projet vise à développer un modèle de classification des genres musicaux à partir de données audio, en intégrant l'analyse en composantes principales (PCA) pour la réduction de dimensionnalité. L'objectif principal est d'améliorer l'efficacité computationnelle tout en conservant les informations discriminantes essentielles des données.

Importation de la base de données

Importation de la base de données

On commence par charger le jeu de données à l'aide de la bibliothèque **pandas**

```
# Load the dataset
df = pd.read_csv('music_dataset_mod.csv')

# Make a copy of the DataFrame for analysis
df_copy = df.copy()
```

Figure 2: Importation de la base de données

Résumer sur la base de données

```
6
7  # Get further insights into the DataFrame
8  print(df_copy.info())
9
10 # Check for missing or null values
11 print(df_copy.isnull().sum())
12
13 # Discover the types of music genres
14 unique_genres = df_copy['Genre'].unique()
15 print(f"Unique genres: {unique_genres}")
16
```

Figure 3: summary data base

Résultats

```
RangeIndex: 1000 entries, 0 to 999
Data columns (total 13 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   Tempo                         1000 non-null   float64
 1   Dynamics Range                1000 non-null   float64
 2   Vocal Presence                1000 non-null   float64
 3   Percussion Strength           1000 non-null   float64
 4   String Instrument Detection   1000 non-null   float64
 5   Electronic Element Presence   1000 non-null   float64
 6   Rhythm Complexity             1000 non-null   float64
 7   Drums Influence               1000 non-null   float64
 8   Distorted Guitar              1000 non-null   float64
 9   Metal Frequencies             1000 non-null   float64
 10  Ambient Sound Influence       1000 non-null   float64
 11  Instrumental Overlaps         1000 non-null   float64
 12  Genre                         890 non-null    object
dtypes: float64(12), object(1)
memory usage: 101.7+ KB
```

Figure 4: information sur la base de données

Conclusion

Merci de votre attention.