Cours : Machine Learning

Chapitre 2 : Apprentissage supervisé

Dr. Asma Ouertatani

Université Centrale

October 15, 2024

Plan du Chapitre II

Algorithmes d'apprentissage Supervisé

- **Classification**
  - **Naïve Bayes**
  - Arbres de décision
  - Forêts aléatoires
  - Support Vector Machines (SVM)
  - K plus proches voisins (K-NN)
- **Régression**
  - Régression linéaire
  - Régression logistique

Introduction au Classifieur Naïve de Bayes

- Le classifieur Naïve de Bayes est un modèle probabiliste de classification basé sur le **théorème de Bayes**.
- Il est basé sur le théorème de Bayes, qui permet de calculer la probabilité d'une classe donnée les caractéristiques observées.
- Il est appelé "naïve" car il fait l'hypothèse d'indépendance entre chaque paire de caractéristiques (features), conditionnellement à la classe.

Théorème de Bayes

Formulation du théorème de Bayes

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

Où :

- $P(C|X)$: La probabilité a posteriori de la classe C donnée les caractéristiques X.
- $P(X|C)$: La probabilité des caractéristiques X sous la classe C.
- $P(C)$: La probabilité a priori de la classe C.
- $P(X)$: La probabilité des caractéristiques X (constante pour toutes les classes).

Indépendance Naïve

- L'hypothèse d'indépendance naïve postule que toutes les caractéristiques $X_1, X_2, \ldots, X_n$ sont indépendantes conditionnellement à la classe C.
- Cette hypothèse permet de simplifier le calcul de $P(X \mid C)$:

$$P(X \mid C) = P(X_1 \mid C) \cdot P(X_2 \mid C) \cdots P(X_n \mid C)$$

Exemple Instructif : Classifieur Naive Bayes

Contexte :
Nous voulons classifier si un email est "spam" ou "non-spam" en utilisant les mots contenus dans l'email.

Données d'entraînement :

- Email 1 : "Achetez des médicaments bon marché" $\Rightarrow$ **Spam**
- Email 2 : "Mise à jour de votre compte bancaire" $\Rightarrow$ **Spam**
- Email 3 : "Invitation à une réunion" $\Rightarrow$ **Non-Spam**

Nouvel email à classifier : "Achetez des médicaments maintenant"

Étape 1 : Calcul des probabilités a priori

Probabilité a priori (des classes) :

- $P(\text{Spam}) = \frac{2}{3}$:
  Proportion d'emails classés comme "Spam".
- $P(\text{Non-Spam}) = \frac{1}{3}$:
  Proportion d'emails classés comme "Non-Spam".

Étape 2 : Calcul des probabilités conditionnelle (dans chaque classe)

- $P(\text{Achetez} \mid \text{Spam}) = \frac{1}{2}$:
  La probabilité que le mot "Achetez" apparaisse dans un email "Spam".
- $P(\text{Achetez} \mid \text{Non-Spam}) = 0$:
  La probabilité que le mot "Achetez" apparaisse dans un email "Non-Spam".

Etapes 3 : Application de la Théorème de Bayes

Pour calculer les probabilités d'appartenance à une classe donnée :

- $P(\text{Spam} \mid \text{Achetez}) = \frac{P(\text{Achetez}|\text{Spam}) \times P(\text{Spam})}{P(\text{Achetez})}$
- $P(\text{Non-Spam} \mid \text{Achetez}) = \frac{P(\text{Achetez}|\text{Non-Spam}) \times P(\text{Non-Spam})}{P(\text{Achetez})}$

Calcul de $P(\text{Achetez})$

Calcul du terme $P($Achetez$)$:

$$P(\text{Achetez}) = P(\text{Achetez} \mid \text{Spam}) \times P(\text{Spam}) + P(\text{Achetez} \mid \text{N-Spam}) \times P(\text{N-Spam})$$

Dans cet exemple :

$$P(\text{Achetez}) = \left(\frac{1}{2} \times \frac{2}{3}\right) + \left(0 \times \frac{1}{3}\right) = \frac{1}{3}$$

Pourquoi $P(x)$ est constant ?

Le terme $P(\text{Achetez})$ est constant dans le processus de comparaison entre les classes, car il ne dépend que de l'observation (le mot "Achetez") et pas de la classe.
Cela signifie que lors de la comparaison entre $P(\text{Spam} \mid \text{Achetez})$ et $P(\text{Non-Spam} \mid \text{Achetez})$, ce terme n'a aucun impact. Cela simplifie la maximisation des probabilités et permet d'utiliser :

$$P(\text{Classe} \mid x) \propto P(x \mid \text{Classe}) \times P(\text{Classe})$$

Le symbole $\propto$ signifie que nous faisons une comparaison proportionnelle des probabilités, sans prendre en compte $P(x)$, puisque ce terme est constant.

Conclusion : Classification proportionnelle

Étape 1 : Calcul pour la classe Spam

$$P(\text{Spam} \mid \text{Achetez}) \propto P(\text{Achetez} \mid \text{Spam}) \times P(\text{Spam}) = \frac{1}{2} \times \frac{2}{3} = \frac{1}{3}$$

Étape 2 : Calcul pour la classe Non-Spam

$$P(\text{N-Spam} \mid \text{Achetez}) \propto P(\text{Achetez} \mid \text{N-Spam}) \times P(\text{N-Spam}) = 0 \times \frac{1}{3} = 0$$

Conclusion : L'email est classé comme **Spam**, car :

$$P(\text{Spam} \mid \text{Achetez}) > P(\text{Non-Spam} \mid \text{Achetez})$$

Exercice : Classification Naïve Bayes

Vous disposez d'une collection d'emails classifiés comme Spam ou Non-Spam. Vous devez utiliser le classifieur Naïve Bayes pour classer un nouvel email.

Données d'entraînement :

- Email 1 : "Achetez des produits pas chers" $\Rightarrow$ **Spam**
- Email 2 : "Mise à jour de votre compte" $\Rightarrow$ **Non-Spam**
- Email 3 : "Promotion spéciale à ne pas manquer" $\Rightarrow$ **Spam**
- Email 4 : "Réunion prévue ce vendredi" $\Rightarrow$ **Non-Spam**
- Email 5 : "Urgent : problème sur votre compte" $\Rightarrow$ **Spam**

Nouvel email à classer : "Achetez des produits en promotion"

1. Calculez les probabilités a priori pour les classes Spam et Non-Spam.
2. Calculez les probabilités conditionnelles des mots présents dans le nouvel email.
3. Appliquez le théorème de Bayes pour déterminer si l'email est Spam ou Non-Spam.

Correction : Probabilités a priori

1. Probabilités a priori :

- Nombre total d'emails : 5
- Nombre d'emails Spam : 3
- Nombre d'emails Non-Spam : 2

$$P(\text{Spam}) = \frac{3}{5} = 0.6, \quad P(\text{Non-Spam}) = \frac{2}{5} = 0.4$$

Correction : Probabilités conditionnelles

2. Probabilités conditionnelles :

- Mots dans le nouvel email : "Achetez", "produits", "promotion"
  - **"Achetez"** :
    - Dans Spam : 1 occurrence (Email 1). Dans Non-Spam : 0 occurrence
  - **"Produits"** :
    - Dans Spam : 1 occurrence (Email 1). Dans Non-Spam : 0 occurrence
  - **"Promotion"** :
    - Dans Spam : 1 occurrence (Email 3). Dans Non-Spam : 0 occurrence
- $P(\text{Achetez} \mid \text{Spam}) = \frac{1}{3}$
- $P(\text{Achetez} \mid \text{Non-Spam}) = 0$
- $P(\text{Produits} \mid \text{Spam}) = \frac{1}{3}$
- $P(\text{Produits} \mid \text{Non-Spam}) = 0$
- $P(\text{Promotion} \mid \text{Spam}) = \frac{1}{3}$
- $P(\text{Promotion} \mid \text{Non-Spam}) = 0$

Correction : Application du théorème de Bayes

3. Application du théorème de Bayes :
Calcul de $P($Spam $\mid$ Nouvel email$)$:

$$P(\text{Spam} \mid \text{Achetez, Produits, Promotion}) \propto$$

$$\text{P(Achetez} \mid \text{Spam)} \times P(\text{Produits} \mid \text{Spam}) \times P(\text{Promotion} \mid \text{Spam}) \times P(\text{Spam})$$

$$= \left(\frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} \times 0.6\right) = 0.02178$$

Calcul de $P($Non-Spam $\mid$ Nouvel email$)$:

$$P(\text{Non-Spam} \mid \text{Achetez, Produits, Promotion}) \propto$$

$$\text{P(Achetez} \mid \text{Non-Spam)} \times P(\text{Produits} \mid \text{Non-Spam}) \times P(\text{Promotion} \mid \text{Non-Spam}) \times P(\text{Non-Spam})$$

$$= (0 \times 0 \times 0 \times 0.4) = 0$$

Conclusion :

$$P(\text{Spam} \mid \text{Nouvel email}) > P(\text{Non-Spam} \mid \text{Nouvel email})$$

Exercice : Classification Naïve Bayes – Fraude Bancaire

Vous travaillez sur un système de détection de fraudes bancaires. Vous disposez de données d'entraînement pour deux classes : **Transaction normale** et **Fraude**. Vous devez utiliser le classifieur Naïve Bayes pour prédire si une nouvelle transaction est frauduleuse ou non.

- Trans 1 : "Montant élevé, Localisation différente, Compte nouveau" ⇒ **Fraude**
- Trans 2 : "Montant bas, Localisation habituelle, Compte ancien" ⇒ **Transaction normale**
- Trans 3 : "Montant élevé, Localisation habituelle, Compte ancien" ⇒ **Transaction normale**
- Trans 4 : "Montant bas, Localisation différente, Compte nouveau" ⇒ **Fraude**
- Trans 5 : "Montant élevé, Localisation différente, Compte ancien" ⇒ **Fraude**

Nouvelle transaction à analyser : "Montant élevé, Localisation différente, Compte nouveau"

Questions

1. **Probabilités a priori** :
   - Calculez les probabilités a priori pour les classes **Transaction normale** et **Fraude**.
2. **Probabilités conditionnelles** :
   - Calculez les probabilités conditionnelles des caractéristiques (**Montant**, **Localisation**, **Compte**) pour chaque classe.
3. **Application du théorème de Bayes** :
   - Appliquez le théorème de Bayes pour déterminer si la transaction est une **Transaction normale** ou une **Fraude**.

Corrigé : Probabilités a priori

Calcul des probabilités a priori :

$$P(Fraude) = \frac{3}{5}, \quad P(Tnormale) = \frac{2}{5}$$

Corrigé : Probabilités conditionnelles

Calcul des probabilités conditionnelles :

- Pour chaque classe, on calcule les probabilités conditionnelles des caractéristiques observées :

$$P(\text{Montant élevé}|Fraude) = \frac{2}{3}, \quad P(\text{Montant élevé}|Tnormale) = \frac{1}{2}$$

$$P(\text{Localisation-diff}|Fraude) = \frac{2}{3}, \quad P(\text{Localisation-diff}|Tnormale) = \frac{0}{2}$$

$$P(\text{Compte nouveau}|Fraude) = \frac{2}{3}, \quad P(\text{Compte nouveau}|Tnormale) = \frac{0}{2}$$

Corrigé : Application du théorème de Bayes

Application du théorème de Bayes :

- Calculez les probabilités a posteriori pour chaque classe :

$$P(\textit{Fraude}|\text{Montant élevé, Localisation différente, Compte nouveau}) \propto$$

$$\text{P(Fraude)} \times P(\text{Montant élevé}|\textit{Fraude}) \times P(\text{Localisation différente}|\textit{Fraude}) \times P(\text{Compte nouveau}|\textit{Fraude}) = \frac{3}{5} \times \frac{2}{3} \times \frac{2}{3} \times \frac{2}{3} = \frac{48}{135} \approx 0.355$$

$$P(\textit{Tnormale}|\text{Montant élevé, Localisation-diff, Compte nouveau}) \propto$$

$$\text{P(Tnormale)} \times P(\text{Montant élevé}|\textit{Tnormale}) \times P(\text{Localisation-diff}|\textit{TTnormale}) \times P(\text{Compte nouveau}|\textit{Tnormale}) = \frac{2}{5} \times \frac{1}{2} \times 0 \times 0 = 0$$

- La transaction est donc plus probablement une **Fraude**.

Types des Classificateurs Naive Bayes

- Modèles de classification basés sur le théorème de Bayes.
- Hypothèse d'indépendance entre les caractéristiques.
- Trois types principaux :
  - Naive Bayes Gaussien
  - Naive Bayes Multinomial
  - Naive Bayes Bernoulli

1. Naive Bayes Gaussien

- **Type de données** : Données continues des valeurs numériques qui peuvent prendre n'importe quelle valeur dans un intervalle donné.
- **Hypothèse** : Les caractéristiques suivent une distribution normale.
- **Calcul des probabilités** :
  - Modélise chaque caractéristique comme une variable aléatoire gaussienne.
  - Paramètres : moyenne et écart-type estimés à partir des données d'entraînement.

2. Naive Bayes Multinomial

- **Type de données** : Données discrètes (comptages) : ne peuvent prendre que des valeurs spécifiques (comme des entiers), les données continues peuvent avoir une infinité de valeurs possibles.
- **Hypothèse** : Caractéristiques représentent des comptages ou des fréquences.
- **Calcul des probabilités** :
  - Modélise les événements avec une distribution multinomiale.
  - Probabilités des classes calculées à partir des comptages.

3. Naive Bayes Bernoulli

- **Type de données** : Données binaires (présence/absence).
- **Hypothèse** : Chaque caractéristique est une variable binaire (0 ou 1).
- **Calcul des probabilités** :
  - Modélise la présence/absence de chaque caractéristique.
  - Fonction de probabilité conditionnelle calculée pour chaque classe.

Comparaison des Types de Naive Bayes

Caractéristique	Naive Bayes Gaussien	Naive Bayes Multinomial	Naive Bayes Bernoulli
Type de données	Continue	Discrète (comptages)	Binaire
Hypothèse de distribution	Normale (gaussienne)	Multinomiale	Binaire
Utilisation typique	Problèmes de régression et classification avec données continues	Classification de texte et documents	Classification de texte binaire (spam/non-spam)
Calcul des probabilités	Basé sur la moyenne et l'écart-type	Basé sur les comptages	Basé sur la présence/absence d'événements

Conclusion

- Chaque classificateur Naive Bayes est adapté à des types de données et des problèmes spécifiques.
- Le choix dépend de la nature des données et des objectifs de classification.

Mesures d'évaluation

- Les mesures d'évaluation sont essentielles pour comprendre la performance d'un modèle de classification.
- Elles permettent de quantifier la précision, la sensibilité, et d'autres aspects du modèle.

Accuracy (Précision)

- L'accuracy est la proportion de prédictions correctes parmi le nombre total d'exemples.
- **Formule :**

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Exemple d'application :**
  - Dans un modèle de détection de spam, si 90 e-mails sont correctement classés sur 100, l'accuracy est de 90%.

Precision (Précision)

- La précision est la proportion de vraies prédictions positives parmi toutes les prédictions positives.
- **Formule :**

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Exemple d'application :**
  - Dans un modèle de détection de maladies, si 80 patients sont diagnostiqués positifs, mais que 20 ne le sont pas, la précision est de 80%.

Recall (Rappel)

- Le rappel est la proportion de vraies prédictions positives parmi toutes les vraies instances positives.
- **Formule :**

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **Exemple d'application :**
  - Dans un système de détection de fraudes, si 50 fraudes sont détectées sur 100, le rappel est de 50%.

F-score (F1-score)

- **Définition** : Le F1-score est une mesure de performance d'un modèle de classification qui prend en compte à la fois la précision et le rappel.
- **Formule** :

$$F1 = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

- **Interprétation** : Le F1-score varie entre 0 et 1, où 1 représente une performance parfaite et 0 une performance médiocre.
- **Utilisation** : Particulièrement utile lorsque les classes sont déséquilibrées, car il combine les deux métriques clés.

Matrice de Confusion

- La matrice de confusion est un tableau qui montre les performances d'un algorithme de classification.
- **Structure :**

	Prédiction Positive	Prédiction Négative
Réalité Positive	*TP*	*FN*
Réalité Négative	*FP*	*TN*

- **Exemple d'application :**
  - Dans un modèle de reconnaissance d'image, la matrice de confusion peut montrer combien d'images de chats et de chiens ont été correctement ou incorrectement classées.

Conclusion

- L'accuracy, la précision, le rappel et la matrice de confusion sont des mesures essentielles pour évaluer les modèles de classification.
- Chaque métrique donne un aperçu différent des performances et peut être utilisée selon le contexte spécifique du problème.