Classification naive Bayésienne

fiche d'aide

Arts et Metiers

A quoi sert l'algorithme ?

La classification naïve bayésienne est utilisée dans divers domaines pour la classification et la prédiction. Elle est particulièrement efficace pour la classification de textes (comme le filtrage des spams), la reconnaissance de patterns dans les données, et dans des applications de machine learning où la simplicité et la rapidité sont cruciales.

Comment fonctionne la classification naive Bayésienne ?

La méthode repose sur le théorème de Bayes et suppose que les caractéristiques d'une donnée sont indépendantes entre elles. Pour classer une donnée, le modèle calcule la probabilité de chaque classe possible et choisit la classe ayant la probabilité la plus élevée. Cela se fait en analysant les caractéristiques de la donnée et en se basant sur des données d'entraînement préalablement fournies

Avantages et Inconvenients

Avantages

1. Simplicité et Rapidité : Facile à implémenter et rapide dans les calculs.
2. Efficace avec de Grandes Bases de Données : Performe bien même avec de grandes quantités de données.
3. Bon avec les Données Indépendantes : Excellente performance lorsque les caractéristiques des données sont indépendantes

Inconvenients

1. Hypothèse d'Indépendance : Peut être irréaliste dans certains cas, ce qui affecte la performance.
2. Sensibilité à la Donnée d'Entraînement : La qualité de la classification dépend fortement de la qualité des données d'entraînement.

Exemple

Considérons le filtrage des e-mails en tant que "spam" ou "non-spam". La classification naïve bayésienne analyse les mots dans les e-mails et, en se basant sur la fréquence des mots dans les catégories de spam et de non-spam apprises lors de l'entraînement, détermine la probabilité qu'un e-mail donné soit du spam.

1 Pour aller plus loin

Lissage de Laplace : Pour éviter le problème des probabilités nulles (lorsqu'un attribut n'apparaît pas dans le jeu de données d'entraînement), le lissage de Laplace est souvent utilisé. Cette technique ajuste les probabilités pour prendre en compte les caractéristiques non observées.
Modèles de Probabilité : Selon le type de données, différents modèles de probabilité peuvent être appliqués. Par exemple, le modèle de Bernoulli traite les attributs comme des variables binaires, tandis que le modèle multinomial est adapté aux caractéristiques discrètes, comme les fréquences de mots dans le traitement du texte.

Pour aller plus loin

1. Scikit-learn.
2. Wikipedia