Apprentissage Artificiel

Fiche

0.1 Introduction

0.2 Classification supervisée :

Def
•Étant donné une collection d'enregistrements (jeu d'apprentissage) - Chaque enregistrement contient un ensemble d'attributs, l'un des attributs étant la classe.
• Trouver un modèle pour l'attribut de classe en fonction des valeurs d'autres attributs.
• Objectif: les enregistrements non encore vus doivent se voir attribuer une classe aussi précisément que possible. - Un ensemble de test est utilisé pour déterminer la précision du modèle. Habituellement, le jeu de données donné est divisé en ensembles de formation et de test, le jeu de formation étant utilisé pour construire le modèle et le jeu de tests utilisé pour le valider.

Utilité:
• Prédire que les cellules tumorales sont bénignes ou malignes
• Classer les transactions par carte de crédit comme légitimes ou frauduleuses
• Classer les structures secondaires de la protéine en hélice alpha, feuille bêta ou bobine aléatoire
• Classer les reportages comme les finances, la météo, les divertissements, les sports, etc.

Techniques de classification: • Méthodes basées sur un arbre de décision
• Méthodes basées sur des règles
• Raisonnement basé sur la mémoire
• Réseaux de neurones
• Réseaux de Bayes naïfs et de croyances bayésiennes
• Machines à vecteurs de support

Induction basée sur arbres de décision
Plusieurs algorithmes:
– Algorithme de Hunt (un des premiers)
– CART
– ID3, C4.5
– SLIQ,SPRINT

Algorithme de Hunt:
• Soit Dt l'ensemble des enregistrements d'apprentissage qui atteignent un nœud t
• Procédure générale:
- Si Dt contient des enregistrements appartenant à la même classe yt, t est un nœud feuille libellé yt
- Si Dt est un ensemble vide, t est un nœud feuille libellé par la classe par défaut, yd
- Si Dt contient des enregistrements appartenant à plusieurs classes, utilisez un test d'attribut pour fractionner les données en sous-ensembles. Appliquez récursivement la procédure à chaque sous-ensemble.

Induction d'arbres:
• Stratégie gourmande. - Fractionner les enregistrements en fonction d'un test d'attribut permettant d'optimiser certains critères.
• Problèmes - Déterminez comment diviser les enregistrements uComment spécifier la condition de test d'attribut? uComment déterminer le meilleur partage?
- Déterminer quand arrêter le fractionnement

Comment spécifier une condition de test?
• Dépend des types d'attribut - Nominal - Ordinal - Continu
• Dépend du nombre de façons de fractionner - Séparation à deux
voies - Séparation à plusieurs voies

Fractionnement basé sur les attributs nominaux
• Multi-way split: utilisez autant de partitions que de valeurs distinctes.
SCHEMA
• Division binaire: divise les valeurs en deux sous-ensembles. Besoin de trouver un partitionnement optimal.
SCHEMA
Qu'en est-il de cette scission? SCHEMA

• Différentes méthodes de traitement - Discrétisation pour former un attribut catégoriel ordinal. U Statique - Discrétion une fois au début. X Les plages dynamiques peuvent être trouvées par intervalle égal, intervalle égal en fréquence (centiles) ou en cluster.
- Décision binaire: (A ¡ v) ou (A $\geq$ v) u considère toutes les divisions possibles et trouve la meilleure coupe u peut être plus intensif en calcul

SCHEMA

Comment déterminer le meilleur partage
• Approche gourmande: - Les nœuds avec une distribution de classe homogène sont préférables.
• Besoin d'une mesure de l'impureté du noeud:
C0: 5
C1: 5 →Non homogène, degré d'impureté élevé
C0: 9
C1: 1→Homogène, faible degré d'impureté

Mesures d'impureté de nœud
• Index de Gini
• Entropie
• Erreur de classification

Problèmes pratiques de classification
• Insuffisance et overfitting
• Valeurs manquantes
• Coûts de classification

Autres issues
• Fragmentation des données
• Stratégie de recherche
• expressivité
• Réplication de l'arborescence

Evaluation de modèles

• Mesures d'évaluation des performances - Comment évaluer les performances d'un modèle?
• Méthodes d'évaluation des performances - Comment obtenir des estimations fiables?
• Méthodes de comparaison des modèles - Comment comparer les performances relatives des modèles concurrents?

tab titre

...	...	...

0.3 Data Mining Classification: Alternative Techniques

Classificateurs basés sur des instances

Exemples:

–Rote-apprenant! Mémorise l'ensemble des données d'apprentissage et n'effectue la classification que si les attributs de l'enregistrement correspondent exactement à l'un des exemples d'apprentissage

-Voisin le plus proche ! Utilise k points les plus proches (voisins les plus proches) pour effectuer le classement

Classificateurs de voisin le plus proche

Idée fondamentale:

- Si ça marche comme un canard, les charlatans comme un canard, alors c'est probablement un canard

Requiert trois éléments

- L'ensemble d'enregistrements stockés
- Distance Métrique permettant de calculer la distance entre les enregistrements
- La valeur de k, le nombre de voisins les plus proches à extraire

Pour classer un enregistrement inconnu:

- Calculez la distance par rapport aux autres enregistrements d'entraînement - Identifiez les k voisins les plus proches
- Utilisez les étiquettes de classe des voisins les plus proches pour déterminer l'étiquette de classe d'un enregistrement inconnu (par exemple, en prenant un vote à la majorité)

...

Example: PEBLS

Classificateur Bayésien

Classificateur Naïf Bayes

...

Réseaux de neurones artificiels(ANN)

...