Sélection et extraction de caractéristiques

Introduction

- Nos formes sont représentées comme un vecteur de caractéristiques.
  Quelles caractéristiques choisir ?
- Mesures : couleur, poids…
  Calculs : coefficients de corrélation…
- Exemple : prospection de pétrole
  - Creuser des puits
  - Mesures sismiques

Curse of dimensionality

- Fléau de la dimension. Bellman 1961.
- Dans un espace 10-D, combien faut-il d'exemples pour avoir la même densité de points que pour 100 données dans un espace 1-D ?
- Information redondante ou non-significative
- Sélection de caract : sous-ensemble de caract parmi celles disponibles
- Extraction de caract : projection des données dans un espace de plus petite dimension

Sélection de caractéristiques

Approches

- Filter (filtre)
  La sélection se fait à partir de statistiques sur les caract.
- Wrapper (enveloppe)
  La sélection prend en compte le taux de reconnaissance (% reco).
- Integrated (intégrée)
  Intégrée dans l'algo de classif (ex. random forest)

Recherche optimale

- Nombre de sous-ensembles à explorer :
  $q \binom{D}{d} = \frac{D!}{(D-d)!d!}$
  Pour d=10 et D=100 : + de 10^{13} combinaisons
- Branch and bound (PSE)
  Mais le critère (proba d'erreur minimum) n'est pas vraiment monotone

Recherche sous-optimale (1)

- Meilleures caract
  Evaluation individuelle des caract
  Pb : densité = poids / volume
- Sequential Forward Selection (SFS)
  Par ajout successif de la caract dispo qui produit le meilleur sous-ens (critère max)
- Generalized SFS (GSFS(r))
  Au lieu d'1 caract, on ajoute r à la fois.
  A chaque étape ,il faut évaluer $\binom{D-k}{r}$ ens

Recherche sous-optimale (2)

- Sequential Backward Selection (SBS)
  Par retrait successif de la caract dispo qui produit le meilleur sous-ens (critère max)
- Generalized SBS (GSBS(r))
  Au lieu d'1 caract, on en retire r à la fois.
- Plus p – moins m
  Ajout de p caract par SFS puis retrait de m caract par SBS
  Si p > m alors X_0=∅ sinon X_0=Y (toutes les caract)

Recherche sous-optimale (3)

- Plus p – moins m généralisé
  Ajout de p caract par GSFS puis retrait de m caract par GSBS
  On peut aussi décomposer p et m en sous-parties
- Sequential Forward Floating Selection (SFFS)
  Ajout d'1 caract (SFS) puis on tente d'enlever (SBS) des caract tant que cela améliore le critère.
  On continue tant que le critère est amélioré.
- ASFFS
  Comme SFFS mais avec GSFS(r) et GSBS(r) avec r déterminé dynamiquement

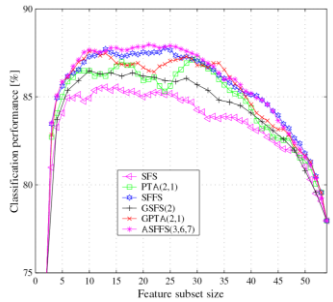
Algorithme génétique

- 1 individu est un tableau de D élts binaires

0	1	1	0	1	1	0	0	0	1

- La fonction objectif est l'évaluation sur la base de validation
- Sélection par tournoi (par exemple)
- Mutation : on inverse un bit d'un individu

Comparaison de méthodes de sélection



Extraction de caractéristiques

Extraction de caractéristiques

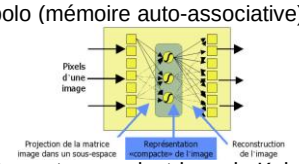
- Calcul de nouvelles caract en effectuant des transformations (linéaires ou non) et des combinaisons des caract originales
- Avantage par rapport à sélection : on peut trouver des caract plus efficaces que le meilleur sous-ens
- Désavantages : les nouvelles caract perdent leur signification et c'est souvent plus difficile

Méthodes statistiques

- Par transformations linéaires
  - ACP
  - ACI (indépendante) : mieux si caract non gaussiennes
  - Séparation de sources indépendantes
    p(A, B) = p(A) . p(B)
  - ACP Kernel
    On transforme (plongement) les données par une fonction non linéaire dans un espace de dim > D puis on applique une ACP

Méthodes neuronales

- MLP (Multi-Layer Perceptron) configuré en diabolo (mémoire auto-associative)



- Cartes auto-organisatrices de Kohonen (SOM)
- Couches cachées d'un MLP
- Réseaux de neurones à couches profondes (CNN par exemple)