

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

2

Cours RF - Di 5

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

3

Cours RF - Di 5

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)

5

Cours RF - Di 5

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

6

Cours RF - Di 5

Recherche sous-optimale (1)

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

7

Cours RF - Di 5

Recherche sous-optimale (2)

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

8

Cours RF - Di 5

Recherche sous-optimale (3)

- Plus p – moins m généralisé
Ajout de p caractéristiques par GSFS puis retrait de m caractéristiques par GSBS
On peut aussi décomposer p et m en sous-parties
- Sequential Forward Floating Selection (SFFS)
Ajout d'1 caractéristique (SFS) puis on tente d'enlever (SBS) des caractéristiques 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

9

Cours RF - Di 5

Algorithme génétique

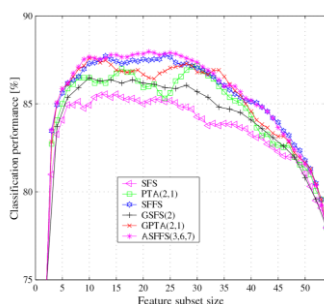
- 1 individu est un tableau de D éléments binaires

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

10

Cours RF - Di 5

Comparaison de méthodes de sélection



11

Cours RF - Di 5

Extraction de caractéristiques

Extraction de caractéristiques

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

13

Cours RF - Di 5

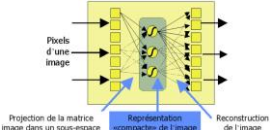
Méthodes statistiques

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

14

Cours RF - Di 5

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)

15

Cours RF - Di 5