

Projet SINF2275

« Data mining and decision making »

Projet classification et reconnaissance de caractères manuscrits

Année académique 2006-2007

Professeurs : Marco Saerens

Adresse : Université catholique de Louvain
Information Systems Research Unit (ISYS)
Institut d'Administration et de Gestion
Place des Doyens 1
B-1348 Louvain-la-Neuve
Belgique

Téléphone : 010 47.92.46.

Fax : 010 47.83.24.

Courriel : saerens@isys.ucl.ac.be

Objectif

L'objectif de ce travail est la mise en pratique concrète d'un certain nombre de techniques d'analyse de données quantitatives, à travers l'étude d'un cas pratique nécessitant l'utilisation de logiciels de traitement statistique de données (SAS/JMP, SPSS, R, S-Plus, Matlab). L'application visée est l'implémentation d'un programme de reconnaissance de caractères manuscrits.

Contexte

Les données fournies sont des « scans » de chiffres écrits de différentes manières, puis numérisés (données MNIST). Les prétraitements de bases sont déjà réalisés de manière à obtenir en final des images de $28 \times 28 = 784$ pixels sur 256 niveaux de gris, sans bruit et centrées. 1000 images sont disponibles pour chaque chiffre allant de 0 à 9 pour un total de 10.000 images.

Pour faciliter les manipulations, toutes les images sont regroupées en une méga matrice *digit* de 10.000×784 , où la ligne *i* contient les 784 pixels de l'image *i*. Les catégories (chiffres tracés) sont stockées dans un vecteur 10.000×1 appelé *classe*. Les fichiers *digit.mat*, *classe.mat* ainsi que le fichier texte *digit.txt* et *classe.txt* (pour ceux qui préfèrent travailler avec un autre logiciel que Matlab), seront disponibles à la première séance de TP consacrée à la classification.

A partir de ces données, vous devrez les diviser en une partie *training set* et une partie *test set*. Les données du *training set* vont servir à adapter les méthodes de classification choisies à notre cas de reconnaissance de caractère. Les données du *test set* vont, quand à elles, servir à évaluer la qualité de prédiction de vos modèles. Ces données du test set ne devront pas participer au processus d'entraînement sous peine de sous-évaluer l'erreur de classification. Il est recommandé de recourir à la méthode de "leave-one-out" ou de "cross-validation" pour ne pas gaspiller les données et avoir une meilleure estimation de l'erreur.

Modalité pratique

Il existe une large littérature sur la reconnaissance de caractères manuscrits, notamment pour la lecture automatique de code postale dans le tri des courriers. Pour ce travail nous allons employer une approche simple qui consiste à traiter chaque pixel des images (qui ont la même taille) comme étant une variable, sans nous préoccuper des notions de voisinages ou de l'orientation des traits. Ensuite, nous allons extraire des traits caractéristiques de l'image (features) à partir de ces pixels.

Avant toute chose il est conseillé de visualiser quelques échantillons d'images pour se familiariser avec les données. On peut constater que, malgré les prétraitements déjà réalisés, les données ne sont pas encore utilisables tel quel. Les pré-traitements qui doivent encore être réalisés sont :

- L'importance de la dimension de l'espace de caractéristique (dimensionality curse) : les images sont composées de 28x28 pixels, soit 784 variables en tout. Ce qui risque de fausser le résultat de la classification malgré la disponibilité de 10.000 échantillons d'apprentissage (soit à peine 1000 par classe). Une extraction de caractéristique (feature extraction) sera nécessaire pour réduire le nombre de variables.
- L'inclinaison des lettres tracées : chaque personne a ses propres habitudes d'écriture, et cela peut se traduire par la légère inclinaison des chiffres au niveau des échantillons. Pour rendre le système de classification invariant à l'orientation des chiffres, il vous est demandé dans ce travail de redresser les chiffres grâce à une analyse en composantes principales en prétraitement.
- L'épaisseur des traits peut varier aussi en fonction de la hauteur des chiffres ou de l'instrument d'écriture utilisé. Ce problème peut être surmonté en effectuant une squelettisation de l'image mais l'opération peut supprimer aussi par la même occasion des informations essentielles. Une analyse avec et sans opération de squelettisation vous sera demandée.

Vous êtes évidemment libre d'employer les méthodes de prétraitement et d'extraction de caractéristique plus sophistiquées que celles proposées dans cet énoncé.

Après le prétraitement, différents modèles de classification seront évalués sur ces données. Le nombre de modèles à implémenter devra être plus grand ou égal au nombre de membres de votre groupe. Toutes les comparaisons de modèles seront effectuées sur base d'une procédure de type « leave-n-out » ou de type « cross-validation ». Vous choisirez, sur cette base, votre meilleur modèle.

Outils logiciels

Nous demandons l'utilisation d'un outil logiciel spécialisé d'analyse des données.

Cinq outils (*SAS*, *SPSS*, *Matlab*, *R*, *S-Plus*) sont disponibles en salle informatique. Notons également que SAS/JMP et SPSS peuvent s'acheter à un prix modique (environ 25€, si vous souhaitez travailler à domicile). R, quant à lui, est gratuit, et peut être téléchargé à partir du site : <http://www.R-project.org>. L'utilisation de R, SAS ou Matlab est conseillée.

Notons que la version étudiant de S-Plus, l'équivalent commercial de R, est gratuite et peut être téléchargée à partir de www.insightful.com.

Rapport

Un seul rapport final comprenant les résultats des trois projets vous sera demandé à la fin du quadrimestre. Pour la partie concernant ce projet, un petit rappel théorique sera demandé. Chaque méthode d'analyse utilisée devra être décrite: les hypothèses seront clairement énoncées et les principes théoriques synthétisés. L'objectif est de démontrer votre compréhension des mécanismes des modèles utilisés et non la retranscription aveugle des formules.