Projet – **reconnaissance de chiffres manuscrits**

{Yann.Esposito}@lif.univ-mrs.fr

20 octobre 2005

☞ Le but de ce projet est la réalisation d'un processus qui permettra de reconnaitre les chiffres manuscrits.

Les données vous sont données sous la forme d'un fichier arff contenant 2000 instances. Chaque instance étant caractérisée par 257 attributs : 256 attributs correspondant à des images de dimension 8×8. La valeur de chacun de ces attributs ayant été renormalisé, c'est-à-dire que le minimum de la valeur de tous les attributs en considérant toutes les instances est -1 que le maximum est 1.

Les données sont accessibles sur le site : http://www.lif.univ-mrs.fr/~esposito/ Dans la section enseignement, téléchargez tous les fichier nécessaire au projet.

Pour vous aider à visualiser les choses, sur mon site, vous pouvez aussi trouver les images numérotées de l'ensemble d'apprentissage.

Il vous est demandé de créer une méthode de sélection. Vous pouvez soit utiliser directement SAS ou Weka, soit faire un programme externe par exemple en utilisant le code C généré par un nœud score de SAS soit en utilisant du code Java qui utilise l'API de Weka.

Vous devez vous limiter aux algorithmes que vous avez vu en cours. En aucun cas vous ne devez utiliser les SVM ou le boosting par exemple.

Il vous est demandé de rendre un rapport. Il devra contenir la liste des étapes par lesquelles vous êtes passés pour résoudre ce problème ainsi que les choix que vous avez fait.

Lorsque l'on fait de la fouille de données, un grande liberté de choix vous est laissée. C'est à vous de faire transparaître dans le rapport les raisons de vos choix. En particulier essayer de justifier la plupart de vos choix, même dans les détails ;

- Pourquoi choisir de la cross-validation ?
- Pourquoi 10 blocs et pas 20 ?
- Pourquoi séparer à 66% et pas à 90% ?
- Pourquoi avoir supprimé un attribut ?
- Pourquoi avoir fait seulement une dizaine d'essai et pas plus ?
- Pourquoi avoir utilisé un certaine méthode de sélection d'attributs ?
- Comment avez-vous visualisés les données ?
- Pourquoi avoir filtré ou non les données ?
- Pourquoi un filtre en particulier ?
- Est-ce qu'un filtre aura un effet sensible sur l'algorithme d'apprentissage ?
- Quelles pistes n'avez-vous pas suivi faute de temps ?
- ...

Pour résumé, le temps vous manquera forcément pour faire tous les essais possibles. Justifiez vos choix et préférez vous concentrer dans peu de domaines plutôt que vous éparpiller dans d'innombrables essais. Dans l'idéal, vous aurez développé des méthodes, voir créer vos propres outils qu'il s'agissent de scripts de filtrages ou de visualisation.

C'est à vous d'organiser votre document en fonction de chaque étape de la création du modèle. Utilisez les méthodes vues en cours et en TP si vous les estimées justifiées.

En fin de compte, votre programme ou votre classifieur devra classer 2000 instances dans un fichier test que vous ne possédez pas. La qualité des résultats obtenus aura nettement moins d'influences que la qualité de votre réflexion visible à travers vos choix.

✷