Introduction au data mining
L3 MIS - STA 1616 - 2010
V. Monbet

Introduction à l'utilisation de SAS Enterprise Miner
Un exemple de problème de data mining

Les objectifs de ce TD sont les suivants :

- s'initier à l'utilisation de SAS Enterprise Miner pour la modélisation en classification supervisée ;
- appréhender et discuter le problème de la comparaison de modèles (ce point sera approfondi plus tard dans le cours).

Le TD sera illustré par le problème suivant : il s'agit de construire un modèle permettant de diagnostiquer une maladie coronarienne sachant un certain nombre d'informations sur le patient listées ci-dessous. On cherche à obtenir un modèle qui soit en priorité performant en prédiction, son interprétabilité étant secondaire. On dispose d'une base de données collectant certains facteurs qui sont susceptibles d'influencer la maladie cardiaque d'hommes du Western Cape, Afrique de Sud
(voir `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`, South African Heart Disease).

Les variables renseignées sont les suivantes :

- `sbp` : pression sanguine systolique *(systolic blood pressure)*
- `tobacco` : la quantité de tabac consommé cumulée *(cumulative tobacco (kg))*
- `ldl` : taux de cholestérol dans le sang *(low density lipoprotein cholesterol)*
- `adiposity` : adiposité
- `famhist` : antécédents familiaux : `Present` s'il y a eu des antécédants familiaux *(family history of heart disease (Present, Absent))*
- `typea` : comportement de type A *(type-A behavior)*
- `obesity` : obésité
- `alcohol` : consommation courante d'alcool *(current alcohol consumption)*
- `age` : age au moment de l'attaque cardiaque *(age at onset)*
- `chd` : variable réponseăcodée 1 si la maladie coronaire est présente, 0 sinon *(response, coronary heart disease)*

Nous proposons le plan de travail suivant.

1. Lire les données, les charger dans SAS EM et visualiser les distributions marginales, en particulier celle de la variable réponse afin de vérifier si elle est équilibrée ou non. Rédéfinir, si besoin, le type des variables (nominal/ordinal, target).

```
libname TPsasem 'h:/DATAMINING/TP1' ;
data TPsasem.heart ;
infile 'h:/DATAMINING/TP1/HeartData.txt' dlm='','';
input sbp tobacco ldl adiposity famhist Present typea
      obesity alcohol age chd ;
run ;
```

2. Partionner les données en un échantillon d'apprentissage et un échantillon de validation. On pourra par exemple choisir une répartition en 2/3-1/3, ce qui correspond à un choix usuel si la distribution de la variable à prédire est équilibrée. Il est préférable de faire un échantillonnage stratifié de façon à conserver la même répartition de la variable réponse dans les deux échantillons.

3. Ajuster les modèles suivants :
   - un modèle linéaire généralisé de type régression logistique ;
   - un arbre de décision de type CART ; on utilisera par exemple le critère de Gini avec un seuil de test à 20% ;
   - un perceptron (réseau de neurones) à une couche cachée comportant 3 neurones.

   On utilisera pour ceci les noeuds appropriés de SAS EM. On choisira comme critère d'optimisation le taux de mauvais classements ou le maximum de vraisemblance.

4. Comparer les modèles obtenus en utilisant les critères suivants :
   - les matrices de confusion,
   - la courbe de taux de réponse

   et interpréter les résultats.