Ecole Centrale de Lyon
Data mining et Big Data

BE- apprentissage statitique
C.HELBERT

Exercice 1 : Bitume - approches PLS, PCR , Lasso

Dans cet exercice on s'intéresse à un jeu de données constitué de 35 Bitumes caractérisés par deux éléments : une mesure de pénétrabilité (colonne PENE des fichiers) et leur spectre infra-rouge, c'est-à-dire par les mesures des absorbances pour 3000 longueurs d'onde distinctes . On dispose également d'un échantillon de test de 9 individus. La mesure de pénétrabilité est très couteuse à obtenir contrairement à l'obtention du spectre IR. L'objectif de l'exercice est de proposer un modèle de prédiction de la pénétrabilité en fonction du spectre.

1. Lire les données «bitume.train.txt» et «bitume.test.txt».
2. Visualiser sur le même graphes avec des couleurs différentes les 35 spectres de l'échantillon d'apprentissage. Tracer l'histogramme des pénétrabilités correspondantes.
3. Faire de même avec l'échantillon test.

BONUS Faire une classification pour identifier des typologies différentes de spectres (routines *kmeans* ou *hclust*). Tracer les pénétrabilités en fonction du numéro de classe. Y-a-t-il un lien ?

4. Ajuster un modèle PCR et PLS (fonction *pcr* et *plsr* du package *pls*) puis un modèle *lasso*. Expliquer les différentes étapes de la sélection des hyperparamètres des méthodes. Interpréter les modèles obtenus. Pour les modèles PCR et PLS on visualisera notamment les premières fonctions propres. Comparer la qualité prédictive sur l'ensemble test des modèles ainsi "calibrés".
5. Pour les 3 méthodes, représenter les pénétrabilités prédites en fonction des pénétrabilités observées sur les 2 ensembles : apprentissage et test.
6. Pourrait-on essayer d'ajuster un modèle linéaire ?

Exercice 2 : Carseats

Dans cet exercice on s'intéresse à un jeu de données constitué de sièges auto pour bébé de 400 magasins décrits par les 11 variables suivantes :

- **Sales** Quantité vendue (en millier) dans chaque magasin
- **CompPrice** Prix du concurrent
- **Income** Niveau de revenu des habitants (en milliers de dollars)
- **Advertising** Budget de publicité du fabriquant dans chaque magasin (en milliers de dollars)
- **Population** Taille de la population locale (in thousands)

— **Price** Charges du fabriquant dans le magasin
— **ShelveLoc** Variable qualitative indiquant le niveau mauvais, bon et moyen selon la qualité de l'emplacement des sièges auto au sein de chaque magasin
— **Age** Age moyen de la population locale
— **Education** Niveau d'éducation de la population locale
— **Urban** Variable qualitative (Oui ou Non) indiquant si le magasin est en zone rurale ou urbaine
— **US** Variable qualitative (Oui ou Non) indiquant si le magasin est aux USA ou non.

On cherche à expliquer la variable **Sales** en fonction des autres.

1. Séparer les données en un ensemble d'apprentissage (70%) et un ensemble test (30%).
2. Mettre en place un modèle CART, RF, bagging et boosting. Expliquer précisément les différentes étapes de la mise en oeuvre. Donner les avantages et les inconvénients de chacun de ces modèles. Illustrer cela sur les données à disposition.
3. Quelle est l'approche qui donne les meilleurs résultats sur l'ensemble test ?
4. Si on décide de mettre en place un modèle linéaire avec une sélection backward sur les tests d'influence des variables, comment faut-il procéder ?