Introduction à R

Web-Mining - Apprentissage statistique

Lundi 17 novembre 2011

Découverte de R

Le but de cette séance est d'apprendre à utiliser efficacement le langage R et son environnement de travail. Après avoir appris à effectuer les commandes de base, vous apprendrez les rudiments de la programmation en langage R (qui est très proche de nombreux autres langages de calculs).

La séance sera d'autant plus profitable que chacun essayera de faire le maximum en autonomie. Avant de poser une question, essayez de résoudre votre problème tout seul. Vous pourrez vous aider de l'aide en ligne, ainsi que du manuel disponible à l'adresse :

http ://cran.r-project.org/doc/manuals/R-intro.html

Un tutorial R en français, très bien fait, écrit par E. Paradis, est disponible en ligne à l'adresse suivante :

http ://cran.r-project.org/doc/contrib/Paradis-rdebuts_fr.pdf

Découverte de R

Ligne de commande, aide en ligne et manuel

Basiquement, R a toutes les fonctionnalités d'une calculatrice moderne. Bien noter l'opérateur d'affectation ← un peu inhabituel.

1. Essayer quelques commandes de bases, comme par exemple :
   `a <- runif(1) ; M <- a*matrix(c(1 :3,rep(4, 3)), ncol = 3, nrow = 2) ; ls()`
2. A l'aide de l'aide en ligne help('ls'), trouver comment effacer toutes les variables à la fois.
3. Trouver à quoi sert la fonction `mosaicplot`, et lancer les exemples fournis par l'aide en ligne.

Utilisation d'un éditeur

En matière d'ergonomie, la ligne de commande R montre vite ses limites. Il est indispensable de taper son code dans un autre éditeur, puis de les exécuter grâce à la commande source.

4. Récupérer sur le site

   http ://perso.telecom-paristech.fr/ ∼ garivier/centrale/

   le fichier relatif aux "Lois dérivées de la Gaussienne", et le lancer dans R.

Fonctions

Les fonctions sont en R des objets comme les autres, qui sont affectées de même façon.

5. Programmer la fonction factorielle récursivement, puis en utilisant une boucle.

Graphiques

La commande de base pour les représentations graphiques est plot. Par défaut, elle ne relie pas les points entre eux.

6. Regarder ce que fait la commande lines. Représenter deux fonctions usuelles sur le même graphe.
7. Illustrer graphiquement la loi forte des grands nombres pour les variables de Bernoulli de paramètre 3/4.

Gestion des données

En plus des classiques tableaux, vecteurs et matrices, R possède deux structures de données particulièrement utiles pour manipulées des données numériques : `list` et surtout `dataframe`. Ces données peuvent être chargés simplement grâce à la commande `read.table`. R contient aussi dans sa distribution quelques jeux de données que l'on utilisera pour illustrer les algorithmes vus en cours. Par ailleurs, les données auxquelles il est fait référence dans HTF sont disponible sur le site indiqué plus haut.

8. Regarder dans le manuel ce que sont les data frames.
9. Exécuter les commandes suivantes, et comprendre ce qui se passe :
   `data() ; attach(cars) ; plot(speed,dist)`
10. Grâce à la commande
    `read.table("http ://www-stat.stanford.edu/~tibs/ElemStatLearn/datasets/SAheart.data",...`
    `sep=",",head=T,row.names=1)`
    récupérer des données, et regarder de quoi elles sont constituées. Quels graphiques est-il pertinent de faire pour les représenter ?

Gestion des packages

Le chargement d'un package (=module complémentaire, qui ajoute des fonctionnalités au noyau de base de R) se fait par la commande require.

11. Récupérer sur le site
    `http ://perso.telecom-paristech.fr/~garivier/centrale/` le fichier relatif à la régression linéaire simple, et regarder ce qu'il contient.

Programmation en R

Régression linéaire

Les régressions linéaires sont gérées par dans logiciel R par la commande lm.

12. Executer les commandes suivante, comprendre ce qu'elles font et ce qu'elles renvoient.

```
Code 1 - Régression linéaire simple
1 : attach(cars)
2 : summary(cars)
3 : plot(speed, dist)
4 : reglin <- lm(dist ~ speed)
5 : summary(reglin)
```

13. Retrouver sans son aide tous les résultats (ou en tous cas une bonne partie) fournis par la commande `lm` sur cet exemple.

Estimation de densité par histogrammes

Soit B une variable de Bernoulli de paramètre $p = 3/4$, soit Z une variable gaussienne centrée réduite, et soit Y la variable $2B - 1 + Z$. On appelle P la loi de Z.

14. Simuler un échantillon de $n = 100$ réalisations indépendantes de la loi P et représenter graphiquement le nuage de point correspondant sur une droite.

15. Dessiner des histogrammes de cet échantillon, en faisant varier le nombre k de catégories ("bins"). Que remarque-t-on quand k est trop faible ? Quand il est trop grand ? Pour quelle valeur de k estime-t-on le mieux la densité de P ?

Perceptron

16. Ecrire une procédure qui ajuste perceptron monocouche comme décrit dans la section 11.3 du HTF.
17. Générer des données séparables dans $\mathbb{R}^2$, et illustrer graphiquement la convergence de l'algorithme. Essayer avec un exemple de données non séparables.