ED- Christel Vrain

Alexandre Masson

14 Janvier 2013

Table des matières

les TDs sont en anglais.

1 Motivation

Plein de données, expansion de l'univers digital. différentes étapes :
- statistique descriptive
- Régression : analyser la relation d'une var vec d'autres
- Fouille de données : Classification supervisée ou non, prédiction , recherche de motifs fréquents.

univ lyon enseignements R

2 Langage R

Langage et environnement pour le calcul statistique et les graphiques.
il tes libre.

commande de bases **Affectation** -> u <- ou assign("x",valeur).

Des Variables : pas de chiffre ou caractère spécial en premier dans le nom, sensible à la casse, certains noms sont déjà pris.

structure de données : vecteur numérique la structure élémentaire est i, vecteur, un nombre est un vecteur, on affecte a une variable un vecteur, avec l'opérateur c(...) en mettant dans c, le contenu du vecteur. l'opérateur [] donne l'élément du tableau , on peux passer à [] un vecteur de données, il renvoi les valeurs placées à tous les index présents dans le tableau entre [].

Arithmétique vectorielle toutes les opérations telles que sin, et cos, sont appliquées point à point sur tous les variables d'un vecteur. les tableaux sont indexés de 1 à N. la fonction sort(vecteur) est implémentée, ainsi que min et max, et range, length, sum, prod, mean.

NaN signifie not a number.

Vecteur Logique les valeurs sont TRUE,FALSE,NA(not available). Nous avons a disposition des opérateurs logiques tels que <, <=, >, >=, ==, !=. nous avons aussi c1&c2, c1|c2 et !c1. is.na() teste si X est NA ou NAN, is.nan(x) teste uniquement le nan.

Vecteur de caractères les valeurs d'un vecteur doivent etre de meme type : logical , numéric, complex, character (ou raw). la fonction mode(x) renvoie le type des valeurs présentent dans x. on utlise as.character(x) pour transformer les valeurs de x en characters.

Index des vecteurs un vecteur, l'opérateur x :y construit un vecteur remplis avec toutes les valeurs entre x et y. l'opérateur - sur un vecteur renvoie le complémentaire du vecteur. il est possible avec names(vecteur) <- c("chaine", "chaine2",...] pour nommer les index des tableaux.

Séquences fonction seq avec au plus 5 arguments :
- from =
- to =
- by =
- length =
- along = vecteur(seul argument si utilisé, créer une séquence de 1 à length(vecteur).

Factor Utilisé pour étiqueter les données de vecteurs de même longueur., il est ensuite possible d'effectuer des opération en discriminant les valeurs par les étiquettes., tapply permet d'appliquer une fonction (3em param) sur un vecteur(1er param), en utilisant les étiquettes un vecteur d'étiquettes (2nd param).

Fonctions définies de la forme nom <- function(arg1,...,argn). les instructions sont séparées par des ;. On a aussi accès aux conditionnelles, et au boucle, mais il faut éviter les boucles, les boucles for, repeat, while(condition) expr, et le break pour terminer les boucles.

Tableau - Matrices Un vecteur peut être utilisé comme un tableau à plusieurs dimensions si on lui associe un vecteur de dimension. La fonction dim(vect) <- c(x,y) ; vect, transforme le vecteur en tableau de lignes et de y colonnes, il remplie ensuite le tableau avec les données du vecteur, en remplissant la première colonne avec autant de valeurs que de lignes, puis la seconde colonne avec la suite, etc...

Tableaux d'indices il peux être construit par la fonction array(vecteur, dimension). on a aussi des opérations sur les matrices, soient a et b deux matrice et f une fonction n alors outer applique la fonction et renvoi une matrice de taille la concaténation des taille des deux vecteurs.

Listes Collection ordonnée d'objets , appelés composant. Un composant peut être désigne par :

- son numéro : Liste [[x]]
- un nom : List $name
- ATTENTION : List[i :J] retourne une sous liste (avec les noms des composants).

Les listes sont extensibles, il est possible de rajouter des champs en mettant : List$nomChamp<-valeur.

Data frame c'est comme une matrice lais avec des modes et attributs différents, les chaines de caractères sont transformées en facteurs.

Lecture des fichiers première ligne doit avoir une chaine de caractère par attribut du dataframe, tout les reste est ensuite lu comme attribut, mais les chaines sont considérées comme facteur.

Réseau de neurones Introduit dans les 60's. Les observations sont décrites par n variables et une étiquette 1-,1 ou 0,1. Les entrées sont reliées au neurones avec des poids, et l'étiquette c'est le produit scalaire des entrées et des poids, et si elle est plus grande qu'un poids on retourne un sinon zéro. Le neurone sépare donc l'espace en deux demi plan, un ou c'est vrai et un ou c'est faux. on note O pour output. Le perceptron est donc un ensemble de neurones qui sont connectés et où les sorties de neurones sont les entrées d'autres neurones.

Apprendre le perceptron c'est apprendre les poids, on constate que les poids et le seuil ne sont pas du même coté de l'équation, on remplace le b par une entrée x0 toujours égale à 1, et un poids w0 qui devra être égale à b.

apprentissage par correction d'erreurs

```
Entrées : un ensemble d'exemples S de R^{n}*0.1
Sortie : P un perceptron défini par (w0,w1,...,wn)
Initialiser aléatoirement les poids wi
Repeter{
Pour tout exemple (x1,x2,...,xn) dans S{
calculer la sortie o
pour tout i, wi <- wi + (c-o)xi
}
jusqu'a convergence }
```

Première question : converge-t-il ? il a été montré que si les exemples sont linéairement séparable, il existe donc un hyperplan, il va le trouver. Dans la pratique, on va limiter le nombre de répétition des tours de boucle.
2 critiques pour cet algorithme :

- : algorithme ne converge que si les données sont linéairement séparables. Donc il converge s'il doit converger.
- : pas de garantie sur la droite que l'on trouve, pas de distance maximum.
- Cet algorithme s'appelle algorithme par descente de gradient.

3 Compléments sur R