

ED- Christel Vrain

Alexandre Masson

14 Janvier 2013

## Table des matières

<b>1</b>	<b>Motivation</b>	<b>3</b>
<b>2</b>	<b>Langage R</b>	<b>4</b>
<b>3</b>	<b>Compléments sur R</b>	<b>6</b>
3.1	Un peu de terminologie . . . . .	6
<b>4</b>	<b>Graphique sous R</b>	<b>7</b>
<b>5</b>	<b>Rappels de probabilité</b>	<b>9</b>
5.1	Notion de base . . . . .	9
5.2	20 Février 2013 . . . . .	10
5.3	Variable aléatoire continue . . . . .	11

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 peut 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, numérique, complex, character (ou raw). la fonction `mode(x)` renvoie le type des valeurs présentent dans x. on utilise `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("chaîne", "chaîne2",...]` 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 opérations 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 <- fonction(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 remplit 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 peut être construit par la fonction `array(vecteur, dimension)`. on a aussi des opérations sur les matrices, soient `a` et `b` deux matrices et `f` une fonction `n` alors `outer` applique la fonction et renvoie une matrice de taille la concaténation des tailles des deux vecteurs.

**Listes** Collection ordonnée d'objets, appelés composants. Un composant peut être désigné 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 mais avec des modes et attributs différents, les chaînes de caractères sont transformées en facteurs.

**Lecture des fichiers** première ligne doit avoir une chaîne de caractère par attribut du dataframe, tout le reste est ensuite lu comme attribut, mais les chaînes 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 côté de l'équation, on remplace le  $b$  par une entrée  $x_0$  toujours égale à 1, et un poids  $w_0$  qui devra être égale à  $b$ .

### apprentissage par correction d'erreurs

```
Entrées : un ensemble d'exemples  $S$  de  $\mathbb{R}^n \times \{0,1\}$ 
Sortie :  $P$  un perceptron défini par  $(w_0, w_1, \dots, w_n)$ 
Initialiser aléatoirement les poids  $w_i$ 
Repeter{
  Pour tout exemple  $(x_1, x_2, \dots, x_n)$  dans  $S$ 
    calculer la sortie  $o$ 
    pour tout  $i$ ,  $w_i \leftarrow w_i + (c - o)x_i$ 
  }
  jusqu'à 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

### 3.1 Un peu de terminologie

- Population : ensemble d'éléments sur lesquels se porte l'analyse
- individu : élément de la population
- variable : sert à décrire la population
- modalité d'une variable : valeur de la variable

**plot** `plot(x,y)`, ou `plot(u)` avec liste de vecteur, ou matrice 2D.

**variables qualitatives** Soucis de commit, penser à regarder le pdf graphique sous R.

## 4 Graphique sous R

6 Février 2013

**Continuons la partie représentation** summary nous donne des infos sur une dataframe telles que le min, le 1st.quarter, la median, la mean, le 3rd. Quarter ainsi que le max. Faire attention, car le choix de découpage influe beaucoup sur la perception, et peut donner une fausse interprétation. faire plusieurs diagramme pour voir les données sous différents angles.

Regarder d'un œil critique

- Mesures de tendances centrales ; moyenne, médiane, etc...

-

**Peut être un cours de rappel de proba** pas des proba pour faire des probas, mais par exemple revoir les lois gaussienne car on utilise des échantillons gaussiens

**Propriété de la moyenne** Toutes les observations ont le même poids : les valeurs extrêmes influencent autant que les autres.

Aussi la somme des écarts à la moyenne est nulle  $\sum(x_i - \text{mean}(x)) = 0$ .

La somme des carrés des distances à la moyenne est plus faible que la somme des carrés des distances à toute autre valeur.

-> la moyenne est la mesure qui minimise la somme des carrés des écarts à elle-même.

**autres moyennes**

- moyenne géométrique : racine  $n_{ieme}$  du produit des n facteurs :  $x_1 * x_2 * ... * x_n$
- moyenne géométrique pondérée racine  $k_{ieme}$  des  $x_n^k$
- moyenne harmonique : k éléments, divisé par la somme des  $1/x^k$
- moyenne quadratique :  $Q = \sqrt{1/k(x_1^2 + x_2^2 + ... + x_n^2)}$

Toutes ces moyennes se généralisent en : le nombre M tel que

$$f(M) = 1/n * (n_1.f(x_1) + ... + n_n.f(x_n))$$

**Médiane** point qui partage la distribution d'une série d'observations en deux parts égales -> partage l'histogramme(densité) en deux parties de surface égales. Ne s'applique que pour des observations pouvant être ordonnées.

Minimise la somme des distances absolues de toutes les observations à ce point.

**Estimation** Classement des observations par ordre croissant

**Calcul** : si le nombre d'observations est impair, la médiane est la  $(k+1)/2$  valeur

Si c'est paire, toute observation entre  $k/2$  et  $(k/2)+1$ , ou la moyenne des observations  $k/2$  et  $(k/2)+1$ .

si les observations sont regroupées par classe, une valeur plus précise sous l'hypothèse d'une répartition uniforme des observations.

**Mode** Valeur qui possède la fréquence la plus élevée  
n'est pas toujours une valeur centrale de la distribution  
une distribution peut avoir plusieurs modes, unimodale, bimodale, asymétrique...

### Comparaison

- moyenne : tiens compte des valeurs de la distribution
- mode : indique une seule valeur de la distribution
- médiane : indique un rang
- si la distribution est unimodale et symétrique, moyenne, mode, et médiane sont confondus.
- Si la distribution est bimodale et symétrique, moyenne et médiane sont confondus.
- si la distribution est asymétrique, ils peuvent avoir des valeurs différentes.
  - tirée à droite, mode, puis médiane, puis moyenne
  - tirée à gauche : moyenne, médiane, mode

**Variance** Elle donne la répartition des valeurs autour de la moyenne, si elle est faible, l'ensemble des valeurs est proche de la moyenne, sinon, cela nous indique que la moyenne n'est pas représentative des valeurs des observations.

### Propriété

la variance est toujours positive

si elle est nulle, toutes les observations sont identiques,

si on ajoute la même constante à toutes les observations, alors la moyenne change, mais pas la variance.

si on multiplie les observations par une même constante positive ou négative, on modifie la variance à un facteur multiplicatif (le carré de la constante, vu que la variance fait un carré)

**Autre façons de calculer la variance** variable quantitative avec  $k$  observations,  $x_1, \dots, x_n = 1/k \sum (x_i^2 - \text{mean}(x)^2)$  Différence entre var de  $R$ , et la vraie variance, car  $R$  se base sur un échantillon pour trouver la moyenne pour effectuer son calcul. On parle d'estimateur sans biais, si quand on change d'échantillon, on retrouve la même valeur qu'avec toute la population, donc la variance ne l'est pas.

**Autres mesures** écart moyen : prend la valeur absolue plutôt que le carré  
écart médian : prend la médiane au lieu de la moyenne pour le calcul de variance.  
intervalle interquartile :

- Intervalle comprenant 50 % des observations le plus au ventre de la distribution.
- Se définit à partir des quantiles qui sont des positions particulières (ex : médiane)
- quartile : coupe en quatre, déciles : amis en 10, centiles : idem mais en 100.



- IQR : intervalle inter quantile : il fait la différence entre le 75% et le 25%

#### **Dispersion pour les variables qualitatives**

- variables dichotomique : même définitions que pour les variables qualitatives
- variable multicatégorielles :  $\sigma^2(X) = p_1 * \dots * p_j$  ou  $p_i$  est la proportion d'apparition de la classe.

#### **Coefficient de Pearson** 13 février 2013

**Moment d'ordre r, en langage R**  $sum[(x - mean(x))^3]/length(x)]^2$

**mesures d'aplatissement** coefficient de Pearson  $\beta_2 = \frac{u^4}{u_2^2}$  d'autant plus faible que la courbe est platicurtique.

coefficient de Fisher  $\gamma_2 = \beta_2 - 3$

## **5 Rappels de probabilité**

### **5.1 Notion de base**

Expérience aléatoire : on connaît l'ensemble des résultats possible, on ne peut prédire le résultat

Ensemble fondamental  $\Omega$  de tout les résultats possibles de l'expérience

Événement : résultat , événement vide =  $\emptyset$

#### **exemple : lancer de deux dés successifs**

- événement vide  $\emptyset$  :somme des deux lancer = 0
- événement certain  $\Omega$  somme < 13

#### **Opérations**

- Négation
- conjonction
- disjonction

#### **Propriétés : Axiome**

- $p(\Omega) = 1$ .
- $p(\emptyset) = 0$
- $p(!A) = 1 - p(A)$
- $p(A \cup B) = p(A) + p(B)$

**Exemple : deux billets pour quatre enfants, 2 filles(b et d), deux garçons (a et c), quelle probabilité de choisir un gars et une fille ?**

- événement  $\Omega$  : on choisit deux personnes a b, a c, a d, b c, b d, c d. Donc chaque couple a une probabilité de  $\frac{1}{6}$
- E = a b, a d, b c, b d. donc  $\frac{4}{6}$

**probabilités conditionnelles** probabilité de B sachant A

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

**Théorème de Bayes**  $P(B|A).P(A) = P(A|B).P(B)$

**Exemple concret** classification supervisée

Utilisation du classifieur bayésien.

Il nous dit que les données sont décrites par des attributs  $A_1, A_2, \dots, A_n$ .

Et sont étiquetées par leur classe.

Étant donnée une observation  $A_1 = v_1, A_2 = v_2, \dots, A_n = v_n$

déterminer la classe.

$$\begin{aligned} \operatorname{argmax}_{C \in \text{Classes}} P(C|A_1 = v_1, A_2 = v_2, \dots, A_n = v_n) &= \\ \operatorname{argmax}_{C \in \text{Classes}} \frac{P(A_1=v_1, A_2=v_2, \dots, A_n=v_n|C) * P_r(C)}{P(A_1=v_1, A_2=v_2, \dots, A_n=v_n)} &= \\ = \operatorname{argmax}_{C \in \text{Classes}} P(A_1 = v_1, A_2 = v_2, \dots, A_n = v_n|C) * P_r(C) \end{aligned}$$

2 classes, 2 attributs  $A_1, A_2$  qui ont 2 valeurs

$2 + 4 * 2 = 10$  probabilités à calculer, c'est beaucoup

**Hypothèse :** attributs indépendants/à la classe

$$P(A_1 = v_1, A_2 = v_2, \dots, A_n = v_n|C) = P(A_1 = v_1|C) * \dots * P(A_n = v_n|C)$$

**Indépendance**  $P(A \cap B) = P(A) * P(B)$

**Indépendance/à la classe**  $P(A \cap B|C) = P(A|C) * P(B|C)$

## 5.2 20 Février 2013

On rappelle que l'ensemble  $\Omega$  est l'ensemble des événements.

**fonction de répartition**  $F(x) = P(X \leq x)$

- F est croissante
- $\forall x, F(x) \in [0, 1]$

**Formule** Espérance :  $E(X) = \sum x_i p(x_i)$

$$\text{Variance } \sigma^2 = \text{Var}(X) = \sum (x_i - \mu)^2 \cdot p(x_i) = E(X - \mu)^2$$

**Propriété de l'espérance et de la variance**

- $E(a.X + b) = a.E(X) + b$
- si X et Y sont indépendants :  $E(X.Y) = E(X).E(Y)$
- $\text{Var}(a.X+b) = a^2.\text{Var}(X)$

**loi conjointe , loi marginale** X : nombre de piles, Y : nombre de pile dans les deux premiers essais. On considère donc les probabilités d'avoir les événements X et Y, à savoir quel(s) événement(s) satisfont Y et X. si x et y sont indépendants  $p(x, y) = p(x) * p(y)$ . la covariance de deux ev indep est 0 , donc on calcule souvent cette donnée sur nos observations pour savoir si des ev sont

indépendants.

$$E(X) = 0 * \frac{1}{8} + 1 * \frac{3}{8} + 2 * \frac{3}{8} + 3 * \frac{1}{8} = 12/8 = 3/2$$

$$E(Y) = 0 * \frac{2}{8} + 1 * \frac{4}{8} + 2 * \frac{2}{8} = \frac{8}{8} = 1$$

loi binomiale : somme d'une série de Bernoulli (ensemble d'év suivant la même loi)

Probabilité d'avoir x succès sur n tirage :  $P(X = x) = C_n^x * p^x * q^{n-x}$  avec  $C_n^x = \frac{n!}{x!(n-x)!}$

$\mu = n * p$  (somme de n variables de Bernoulli indépendantes)

$$\sigma^2 = n * p * q$$

**Exemple avec R**

rXXXX : génération de nombres aléatoires suivant la loi XXXX

dXXXX étant donné un ensemble de valeurs, retourne la hauteur de la probabilité de distribution à chaque point.

pXXXX : probabilité  $P(x \leq x)$  qu'un nombre tiré aléatoirement suivant cette loi soit inférieur à un nombre x donné.

qXXXX : nombre X tel que  $P(X \leq x)$  soit égale à n , qui est donné.

### 5.3 Variable aléatoire continue

Machine avec 24 composants

probabilité qu'un composant tombe en panne : 0.2

la machine fonctionne quand au moins deux tiers des composants sont en marche.

Probabilité que la machine fonctionne.

X nombre de composants en marche.

$$P(X \geq 16) = \sum_{i=16}^{24} P(X = i) = \sum_{i=16}^{24} C_{24}^i (0.8)^i (0.2)^{24-i}$$