

Memo pour l'année

LAURENT Thomas

Master 2 informatique 2018

# Contents

<b>1</b>	<b>Fouille de donnée</b>	<b>1</b>
1.1	Pré traitement des données . . . . .	2
1.1.1	Nettoyage des données . . . . .	2
1.1.2	Normalisation . . . . .	2
1.2	Classification . . . . .	3
1.2.1	Évaluation des classifieurs . . . . .	3
1.3	Arbre de décision . . . . .	4
1.3.1	critères de sélection C4.5 . . . . .	4
<b>2</b>	<b>Recherche Opérationnel</b>	<b>8</b>
<b>3</b>	<b>Apprentissage par le pratique</b>	<b>10</b>
3.1	Rappel . . . . .	11
3.1.1	Matrices et calculs sur les Matrices . . . . .	11
3.2	Algorithms Learn a Mapping From Input to Output . . . . .	12
3.2.1	linear ML algorithms . . . . .	12
3.2.2	Supervised machine learning . . . . .	12
3.2.3	Unsupervised machine learning . . . . .	12
3.2.4	semi-supervised machine leaning . . . . .	12
3.2.5	Overview of bias and variance . . . . .	13
3.3	Overfitting and Underfitting . . . . .	14
3.4	Linear Algorithms . . . . .	15
3.4.1	Régression linéaire . . . . .	15
3.4.2	Least squares linear regression . . . . .	16
<b>4</b>	<b>Représentation des connaissances et raisonnement</b>	<b>17</b>
<b>5</b>	<b>Outils formel</b>	<b>19</b>

<b>6</b>	<b>XML</b>	<b>21</b>
<b>7</b>	<b>Anglais</b>	<b>23</b>

# Chapter 1

## Fouille de donnée

## 1.1 Pré traitement des données

### 1.1.1 Nettoyage des données

#### Caractéristiques descriptives

Objectifs: Résumer, décrire certains aspects (tendances, variation, dispersion...) des données en utilisant certaines mesures :

**Moyenne (espérance)** :  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

**Ecart moyen** :  $\frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$

**Variance** :  $v = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$

**Ecart type** :  $\sigma_x := \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} = \sqrt{\frac{1}{n} (\sum_{i=1}^n x_i^2) - \bar{x}^2}$

**Médiane** : Valeur se trouvant au milieu d'une série de données ordonnées

**Mode** : Valeur la plus fréquente

**Amplitude** : min, max

### 1.1.2 Normalisation

**Min-max** :  $v_n = \frac{v - v_{min}}{v_{max} - v_{min}}$

**Min-max dans l'intervalle [A,B]** :  $v_n = \frac{v - v_{min}}{v_{max} - v_{min}} * (B - A) + A$

**Z-Score** :  $v_n = \frac{v - moyenne}{ecart_{type}}$

**Decimal scaling** :  $v_n = \frac{v}{100^j}$

## 1.2 Classification

### 1.2.1 Évaluation des classifieurs

**Matrice de confusion**

**Percent of correct classification :**

$$\text{PCC}(\%) : = \frac{N_c}{N_t} * 100$$

$N_c$  : nombre d'instances correctement classées

$N_t$  : nombre d'instances testées ( $N_t = |D_{test}|$ )

Exemple:

$$: \begin{pmatrix} - & c1 & c2 & c3 & c4 \\ c1 & 0 & 1 & 0 & 0 \\ c2 & 1 & 60 & 0 & 1 \\ c3 & 0 & 1 & 23 & 0 \\ c4 & 1 & 0 & 7 & 5 \end{pmatrix}$$

**Taux d'erreurs : 100-PCC**

$$\text{PCC}(\%) = \frac{0+60+23+5}{100} * 100 = 88\%$$

## 1.3 Arbre de décision

### 1.3.1 critères de sélection C4.5

Construction d'un arbre de décision C4.5 La construction d'un arbre de décision avec C4.5 passe par deux phases:

**Phase d'expansion** : La construction se fait selon l'approche descendante et laisse croître l'arbre jusqu'à sa taille maximale.

**Phase d'élagage** : Pour optimiser la taille l'arbre et son pouvoir de généralisation, C4.5 procède à l'élagage (pour supprimer les sous-arbres qui ne minimisent pas le taux d'erreurs)

**Approche de construction d'un AD** : Partitionner récursivement les données en sous-ensembles plus homogènes ... jusqu'à obtenir des partitions qui contiennent des objets qui appartiennent majoritairement à la même classe.

=> Théorie de l'information pour caractériser le degré de mélange, homogénéité, impureté, incertitude...

**Théorie de l'information** : Théorie mathématique ayant pour objet l'étude du contenu informationnel d'un message.

Applications en codage, compression, sécurité...

**Entropie** : Mesure la quantité d'incertitude dans une distribution de probabilités.

## Rappel sur les probabilités

**Quelques rappels de probabilités :** Soient X et Y deux variables aléatoires discrètes prenant leurs valeurs dans  $DX=x_1,\dots,x_n$  et  $DY=y_1,\dots,y_m$  respectivement.

$$P(x_i) = \frac{|x_i|}{\sum_{j=1}^n |x_j|}$$

$$\sum_{i=1}^n P(x_i) = 1$$

$$P(x_i|y_i) = \frac{P(x_i, y_i)}{p(y_i)}$$

$P(x_i, y_i) = p(x_i) * p(y_i)$  Si X et Y sont indépendantes

Exemple:

$$: \begin{pmatrix} Anne & Sexe & \# & \% \\ M1 & M & 25 & 25/55 \\ M1 & F & 4 & 4/55 \\ M2 & M & 25 & 25/55 \\ M2 & F & 1 & 1/55 \end{pmatrix}$$

$$P(sexe = M) = P(Sexe = MetAnne = M1) + P(Sexe = MetAnne = M2) = 50/55$$

$$P(Anne = M2|sexe = M) = P(Sexe = MetAnne = M2)/P(Sexe = M) = \frac{25}{55} / \frac{50}{55} = \frac{25}{50} = \frac{1}{2}$$



## Entropie

**Entropie** : Mesure la quantité d'incertitude (manque d'information) dans une distribution de probabilités. Soit  $X$  une variable aléatoire discrète prenant ses valeurs dans  $DX = x_1, \dots, x_n$ . Soit  $P$  la distribution de probabilités associée à  $X$ .

$$H(X) = - \sum_{i=1}^n p(x_i) * \log_2(p(x_i))$$

Par convention, quand  $p(x) = 0, 0 * \log(0) = 0$

Exemple:

X	P(X)
x_1	1/3
x_2	1/3
x_3	1/3

$$H(X) = -p(x_1) * \log_2(p(x_1)) - p(x_2) * \log_2(p(x_2)) - p(x_3) * \log_2(p(x_3))$$

$$H(X) = -3(\frac{1}{3} * \log_2(\frac{1}{3})) = \log_2(3) = 1.58$$

Autre exemples:

$$[\frac{1}{2}, \frac{1}{4}, \frac{1}{4}] : H(X) = 1.5$$

$$[1, 0, 0] : H(X) = 0$$

$$[\frac{1}{2}, \frac{1}{2}] : H(X) = 1$$

Propriétés:

$$H(X) \geq 0$$

$H(X)$  est maximale pour une distribution uniforme (toutes les valeurs sont équiprobables).

**Entropie conjointe** : L'entropie conjointe de deux variables aléatoires  $X$  et  $Y$  est l'incertitude relative à ces deux variables conjointement.

$$H(X, Y) = - \sum_{i,j=1}^n p(x_i, y_j) * \log_2(p(x_i, y_j))$$

**Exemple** :  $[0.2, 0.1, 0.3, 0.4] : H(X, Y) = 1.85$

Critère de sélection: Gain d'information:

$$GAIN(T, A) = Info(T) - Info(T|A)$$

**Avec**  $Info(T)$  : Entropie au niveau de T (avant de partitionner)

$$Info(T) = - \sum_{c_i} freq(c_i, T) * \log_2(freq(c_i, T))$$

$$\textbf{Avec } freq(c_i, T) = p(c_i) = \frac{|c_i|}{|T|}$$

**Avec**  $Info(T|A)$  l'entropie conditionnelle de T une fois partitionné selon les valeurs de l'attribut A.

$$Info(T|A) = \sum_{a_j \in A} freq(a_j, T) * Info(T|a_j)$$

Critère de sélection: Gain Ration:

Le gain d'information favorise les attributs ayant de larges domaines.

Le ratio de gain utilise le gain d'information avec un facteur pénalisant les attributs ayant des domaines trop larges.

$$GainRatio(T, A) = \frac{Gain(T, A)}{SplitInfo(T, A)}$$

$$\textbf{Avec } SplitInfo(T, A) = - \sum_{a_j \in A} freq(a_j, T) * \log_2(freq(a_j, T)) = EntropiedeA$$

## Chapter 2

# Recherche Opérationnel

gggg

## Chapter 3

### Apprentissage par le pratique

## 3.1 Rappel

### 3.1.1 Matrices et calculs sur les Matrices

#### Addition

$$\begin{pmatrix} 1 & 3 \\ 1 & 0 \\ 1 & 2 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 7 & 5 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 1+0 & 3+0 \\ 1+7 & 0+5 \\ 1+2 & 2+1 \end{pmatrix} = \begin{pmatrix} 1 & 3 \\ 8 & 5 \\ 3 & 3 \end{pmatrix}$$

#### Multiplication

$$\begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix}$$
$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 19 & 22 \\ 43 & 50 \end{pmatrix}$$
$$(1 * 5) + (2 * 7) = 19$$

#### Transposer

$$\begin{pmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}$$

#### Inverse

Soit une matrice 2x2 comme :  $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$

Soit Determinant  $D = ad - bc$

Si  $D \neq 0$  alors il existe une matrice inverse égal à :  $\frac{1}{D} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$

## 3.2 Algorithms Learn a Mapping From Input to Output

### 3.2.1 linear ML algorithms

Simplifier les processus d'apprentissage et réduire la fonction sur ce qu'on connaît

**Soit** :  $B_0 + B_1X_1 + B_2X_2 + B_3X_3 = 0$

Où  $B_0, B_1, B_2, B_3$  sont les coefficients présent sur l'axe des ordonnées.

Et  $X_1, X_2, X_3$  sont les valeurs en Input.

### 3.2.2 Supervised machine learning

L'apprentissage supervisé peut se diviser en 2 partis

**Classification** : Quand les variables en sortie sont des Classe (*Blue, Rouge, Vert, Carr, Homme*)

**Regression** : Quand les variables en sortie sont des valeur numérique (*euro, poids, quantits*)

### 3.2.3 Unsupervised machine learning

Les problèmes de l'apprentissage non supervisé sont:

**Clustering** : L'art de faire des paquet d'éléments qui ont des points commun, comme regrouper les clients par paquet de choses qu'ils ont le plus en commun.

**Association** : Associer des règles d'apprentissage pour décrire une portion du data, comme une personne qui a acheté un item A et qui est aussi tenté par acheter un item B

### 3.2.4 semi-supervised machine leaning

L'apprentissage semi supervisé c'est avoir un bonne quantité de données en input X, et un peu de data avec le label Y.

### 3.2.5 Overview of bias and variance

La prédiction des erreurs pour les algorithmes sont regroupé en 3 points:

**Bias Error** : Simplifier l'hypothèse faite par le modèle pour faire une fonction d'apprentissage plus facile.

**Variance Error** : Et la quantité estimée par la fonction visée qui changera via un différent ensemble de données utilisé.

**Irreducible Error** : Ne peut pas être réduit



### 3.3 Overfitting and Underfitting

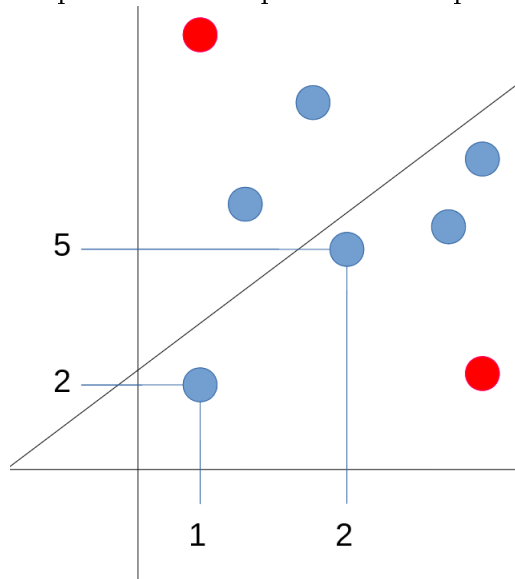
ddddddddd

## 3.4 Linear Algorithms

Soit  $X$  l'ensemble des variables indépendantes sur l'axe des abscisses et  $Y$  l'ensemble des variables dépendantes sur l'axe des ordonnées.

### 3.4.1 Régression linéaire

Étant donné un plan à deux dimensions où l'abscisse contient les points d'entrée  $X$  et l'ordonnée contient les points de sortie  $Y$ , et un nuage de points précédemment acquitté de tout point éloigné du nuage.



*Figureap – linear – regression<sub>1</sub>*

Avec :  $y = \beta_0 + \beta_1 x$

Pour un hyperPlan (3d) :  $y = \beta_0 + \beta_1 x_1 + \dots \beta_n x_n$

Exemple:

$$5 = \beta_0 + 2 * \beta_1$$

$$2 = \beta_0 + 1 * \beta_1$$

### 3.4.2 Least squares linear regression

Calculer la régression linéaire avec la méthode Least squares:

Soit

$\mathbf{X} = [1, 2, 3, 4, 5]$  les variables indépendantes d'axe abscisse

$\mathbf{Y} = [2, 4, 5, 4, 5]$  les variables dépendantes d'axe ordonnée

**Calculons**  $y = \beta_0 + \beta_1 x$

Calcule de la moyenne de X et Y:

$$\mathbf{Xm} = \sum x_i \in X = 3$$

$$\mathbf{Ym} = \sum y_i \in Y = 4$$

Toutes ligne de régression doivent passer par le point (Xm,Ym). Calculer tout les écarts des  $x_i \in X$  par rapport à Xm:

X	Y	$X - Xm$	$Y - Ym$	$(X - Xm)^2$	$(X - Xm)(Y - Ym)$
1	2	-2	-2	4	4
2	4	-1	0	1	0
3	5	0	1	0	0
4	4	1	0	1	0
5	5	2	1	4	2

## Chapter 4

# Représentation des connaissances et raisonnement

ggggg

## Chapter 5

### Outils formel

yyyyyy

# Chapter 6

## XML



uuuuu

# Chapter 7

## Anglais

fffff