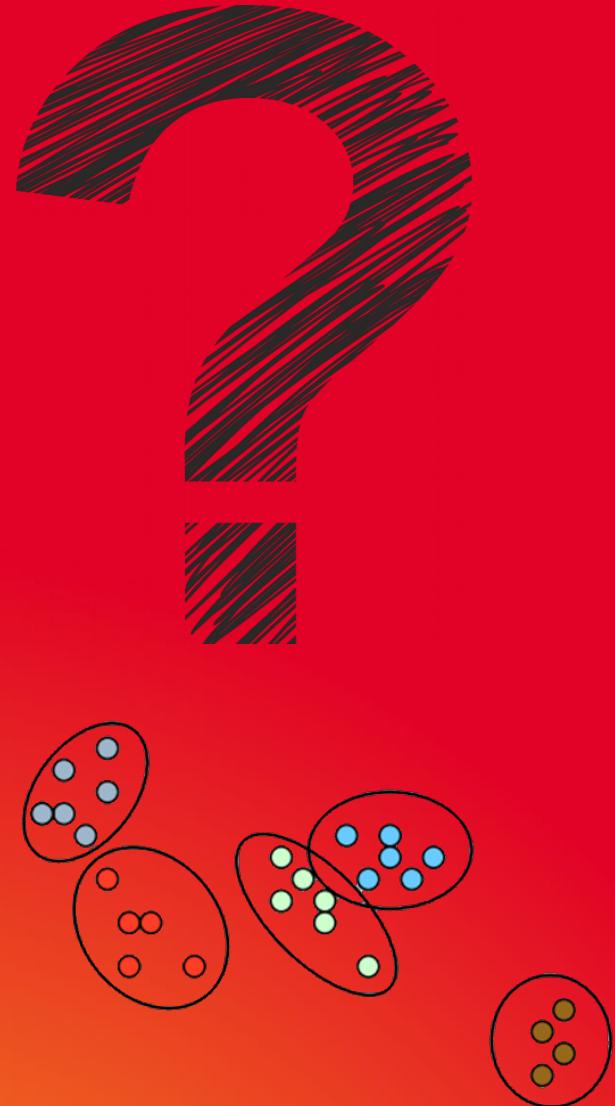




Inria
INVENTEURS DU MONDE NUMÉRIQUE

La classification, quésaco ?



Classification

nom féminin

Action de ranger par classes, par catégories des choses présentant des critères en commun.

Classification

nom féminin

Action de ranger par classes, par catégories des choses présentant des critères en commun.

Variables

Individus : personnes, clients, objets, tickets de caisse, ...

Variables : âge, prix, note,
salaire, longueur, largeur,
statut marital, ...

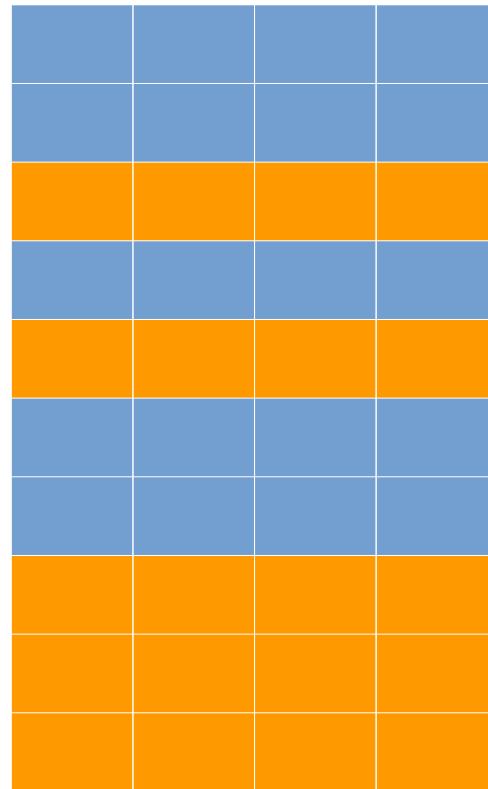
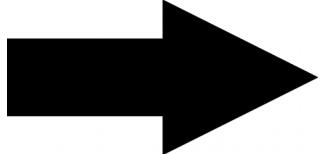
Classification

nom féminin

Action de ranger par classes, par catégories des choses présentant des critères en commun.

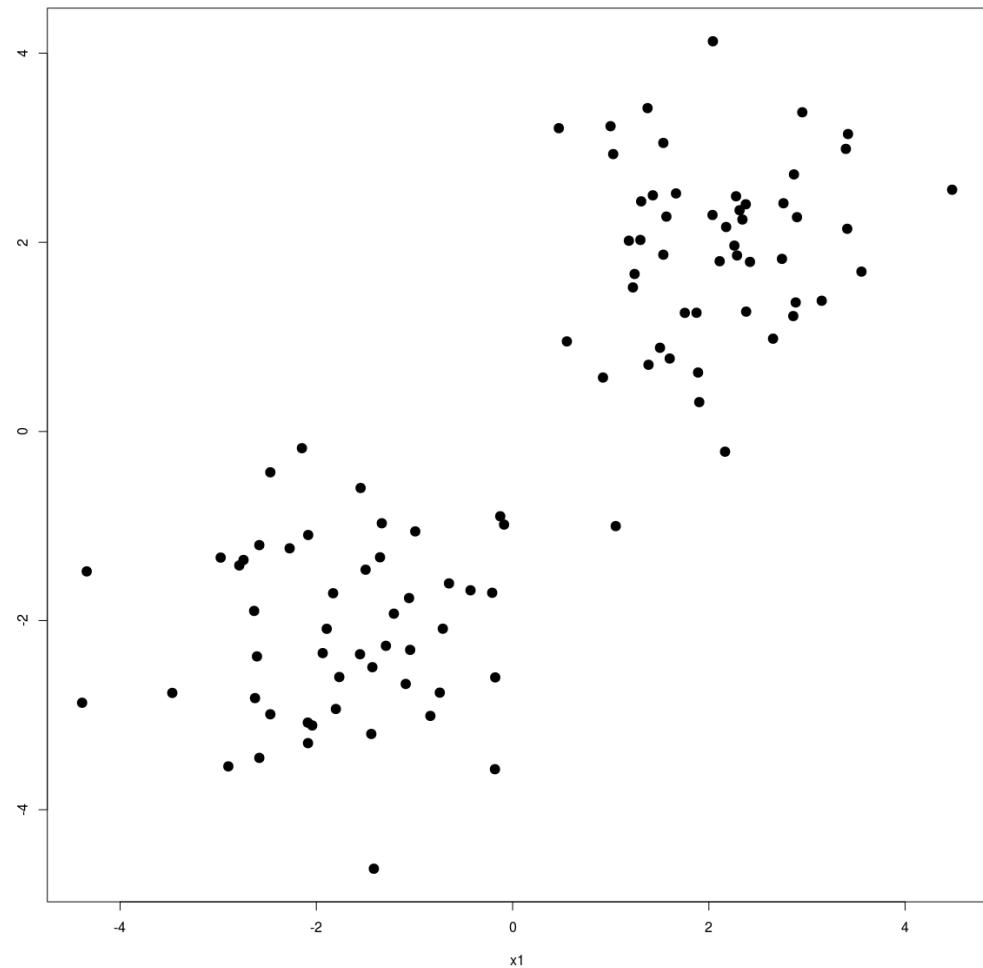
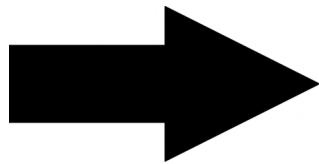
Variables

Individus

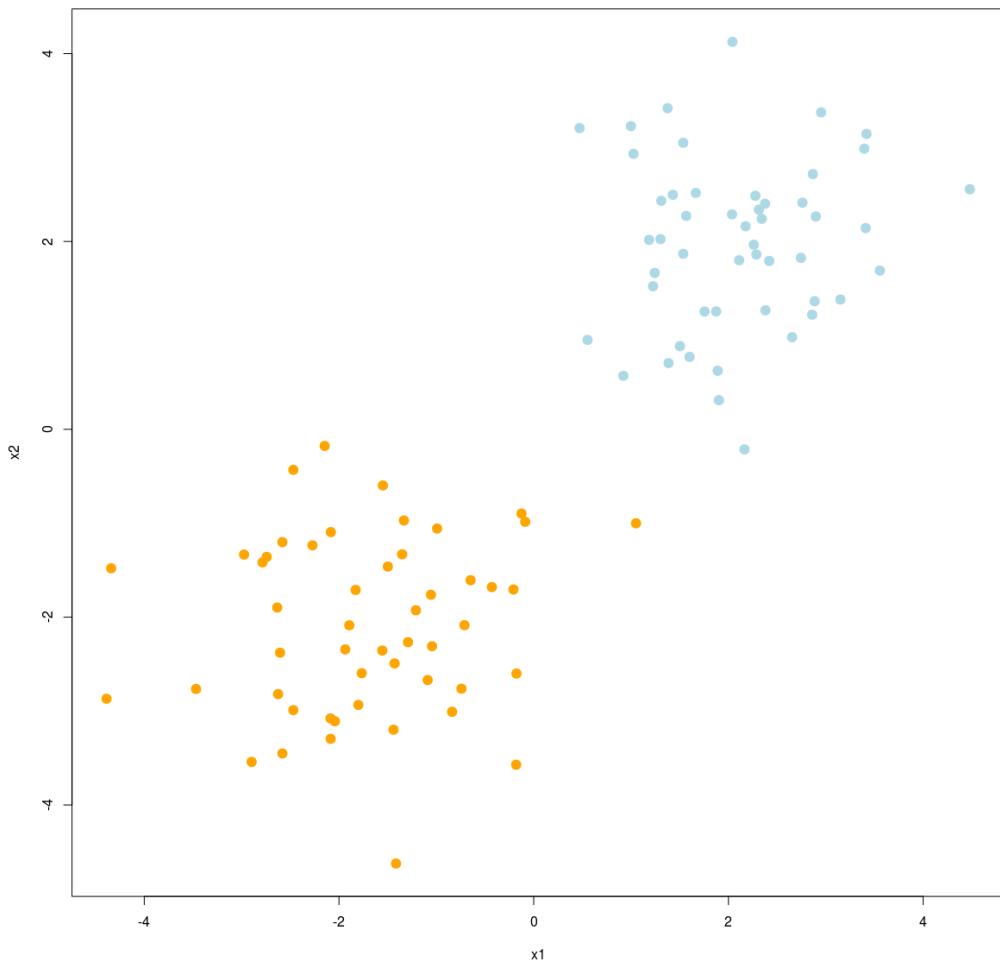
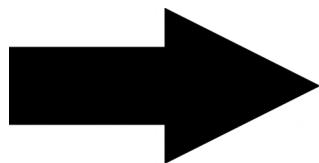


x1	x2
2.283	1.589
2.461	2.693
1.84	1.228
-3.408	-2.089
2.971	2.433
-3.132	-3.686
2.266	2.387
-1.498	-2.757
2.369	1.188
-1.319	-1.132
...	...
-0.457	-3.259
0.144	-1.228
-3.218	-1.385
-2.084	-1.372
-0.496	-2.815
3.548	2.858
-2.191	-2.122
-0.747	-2.848
1.814	2.044
2.975	1.103

x1	x2
2.283	1.589
2.461	2.693
1.84	1.228
-3.408	-2.089
2.971	2.433
-3.132	-3.686
2.266	2.387
-1.498	-2.757
2.369	1.188
-1.319	-1.132
...	...
-0.457	-3.259
0.144	-1.228
-3.218	-1.385
-2.084	-1.372
-0.496	-2.815
3.548	2.858
-2.191	-2.122
-0.747	-2.848
1.814	2.044
2.975	1.103



x1	x2
2.283	1.589
2.461	2.693
1.84	1.228
-3.408	-2.089
2.971	2.433
-3.132	-3.686
2.266	2.387
-1.498	-2.757
2.369	1.188
-1.319	-1.132
...	...
-0.457	-3.259
0.144	-1.228
-3.218	-1.385
-2.084	-1.372
-0.496	-2.815
3.548	2.858
-2.191	-2.122
-0.747	-2.848
1.814	2.044
2.975	1.103

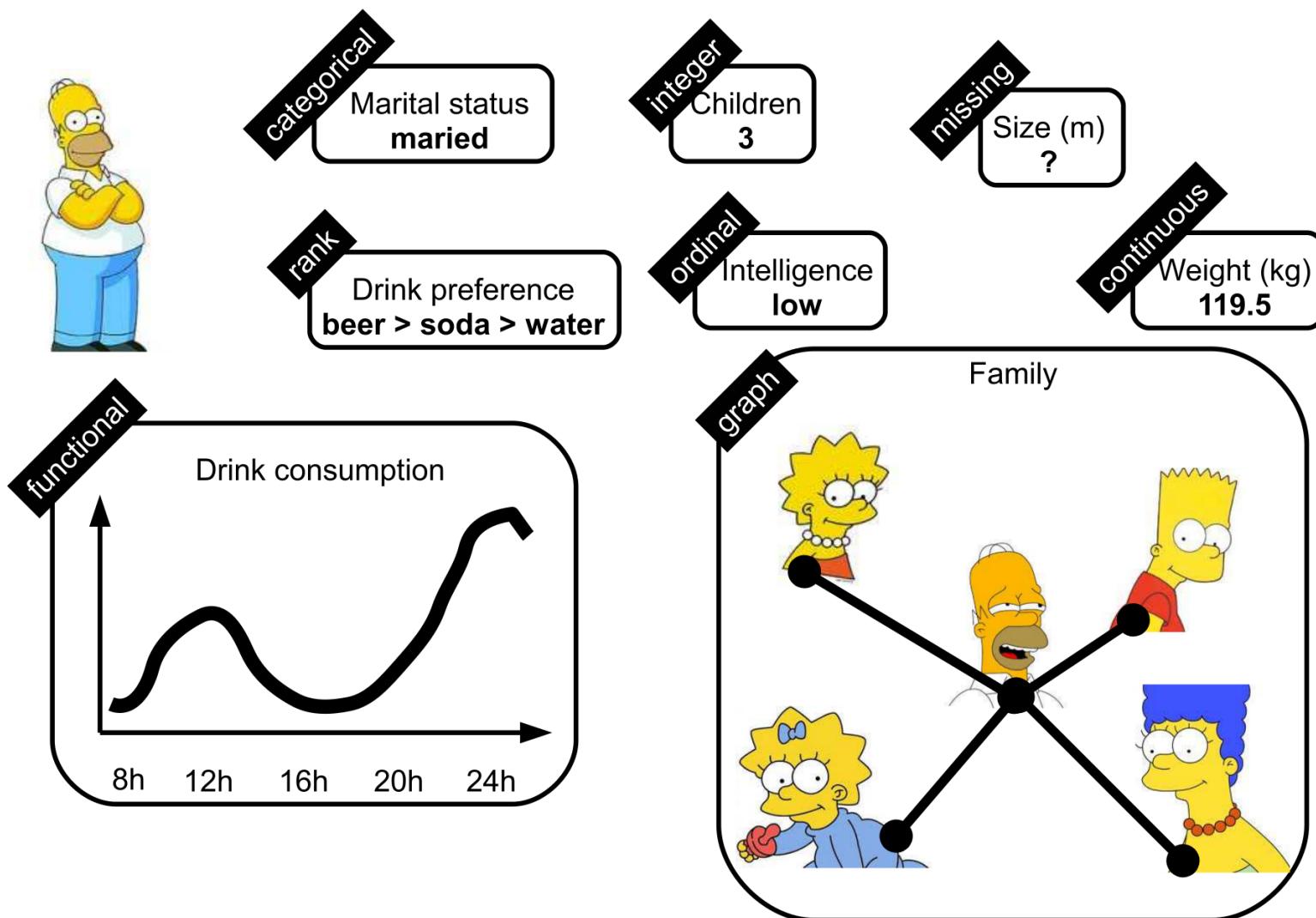


Facile, n'est-ce pas ?

Mais s'il y a :

- des dizaines de milliers d'individus
- des milliers de variables
- des variables de types différents
- des valeurs manquantes

Variables de types différents



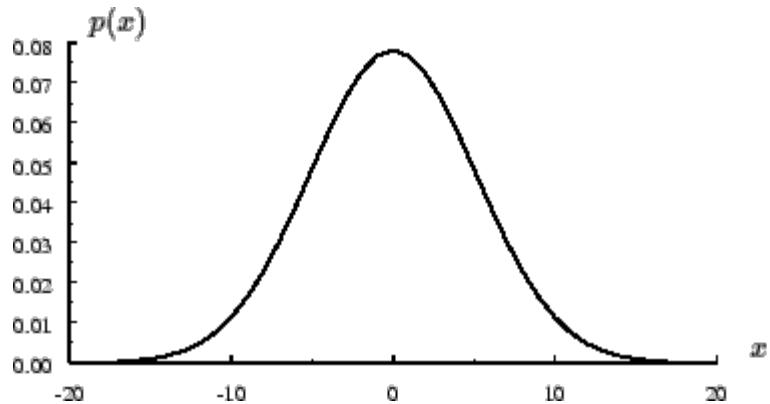
Et les maths dans tout ça ?

Et les maths dans tout ça ?

- Chaque variable est modélisée à l'aide d'une loi de probabilité
Exemple : La loi associée à un pile ou face est une loi de Bernoulli avec : probabilité d'avoir pile = p et probabilité d'avoir face = $1-p$

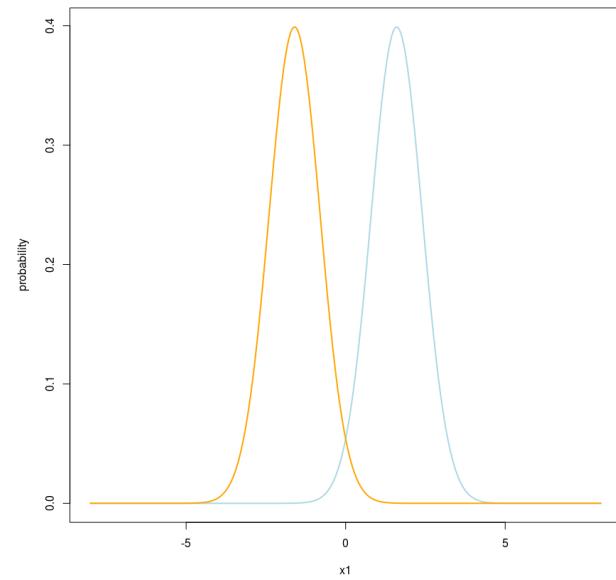
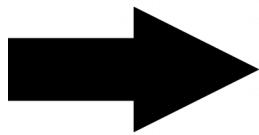
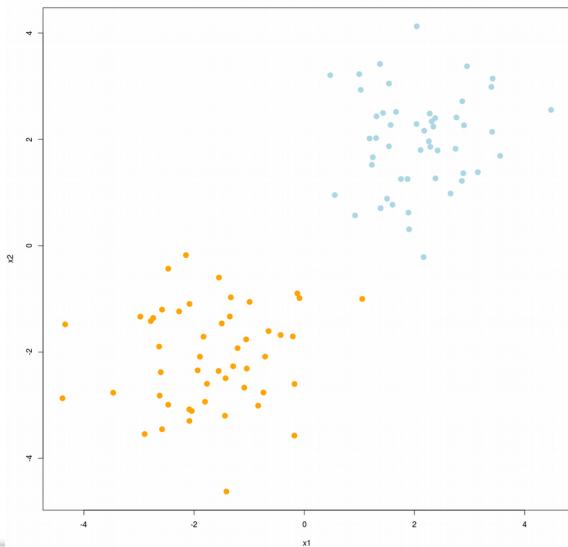
Et les maths dans tout ça ?

- Chaque variable est modélisée à l'aide d'une loi de probabilité
 - On peut mettre une loi sur tout type de variables
- Exemple : Une loi normale (ou gaussienne) pour les variables réelles. Cette loi se caractérise par une moyenne et un écart-type : plus on s'éloigne de la moyenne moins une valeur est probable.



Et les maths dans tout ça ?

- Chaque variable est modélisée à l'aide d'une loi de probabilité
 - On peut mettre une loi sur tout type de variables
 - Les paramètres de cette loi changent en fonction de la classe
- Exemple : une pièce non truquée à un paramètre $p=0,5$. Une pièce truquée aura sa probabilité de tomber sur pile plus grande ou plus petite ($p=0,8$ par exemple).



Et les maths dans tout ça ?

- Chaque variable est modélisée à l'aide d'une loi de probabilité
 - Les paramètres de cette loi changent en fonction de la classe
 - On peut mettre une loi sur tout type de variables
 - Ces lois servent à caractériser la notion de « ressemblance » entre individus
- Deux individus « proches » sont issus de la même loi (= dans la même classe)

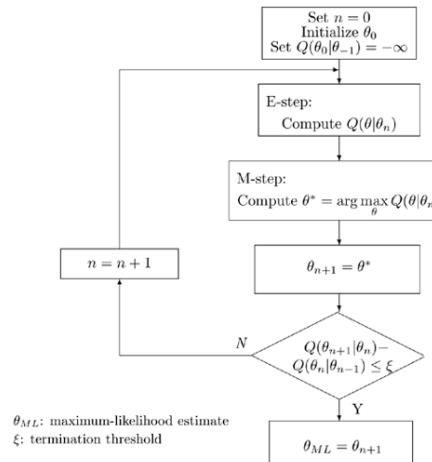
Et les maths dans tout ça ?

- Chaque variable est modélisée à l'aide d'une loi de probabilité
- Les paramètres de cette loi changent en fonction de la classe
- On peut mettre une loi sur tout type de variables
- Ces lois servent à caractériser la notion de « ressemblance » entre individus
- La classification se fait à l'aide d'un algorithme estimant les paramètres des différentes lois

Expectation Maximization (EM) Algorithm
 $\log \text{ expectation of } P(x|z)$

Goal: $\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \log \left(\sum_z p(\mathbf{x}, \mathbf{z} | \theta) \right) \quad f(\mathbf{E}[X]) \geq E[f(X)]$

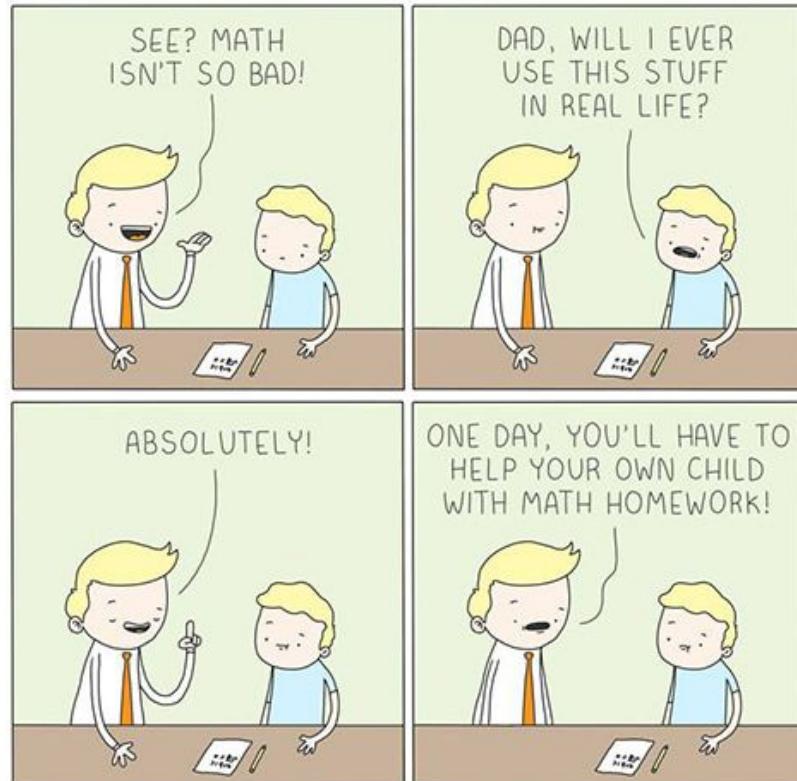
1. E-step: compute expectation of log of $P(x|z)$
 $E_{z|x, \theta^{(t)}} [\log(p(\mathbf{x}, \mathbf{z} | \theta))] = \sum_z \log(p(\mathbf{x}, \mathbf{z} | \theta)) p(\mathbf{z} | \mathbf{x}, \theta^{(t)})$
2. M-step: solve
 $\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}} \sum_z \log(p(\mathbf{x}, \mathbf{z} | \theta)) p(\mathbf{z} | \mathbf{x}, \theta^{(t)})$



Les maths, c'est utile dans la vraie vie ?

Les maths, c'est utile dans la vraie vie ?

- Aider vos enfants avec leurs devoirs de maths



Les maths, c'est utile dans la vraie vie ?

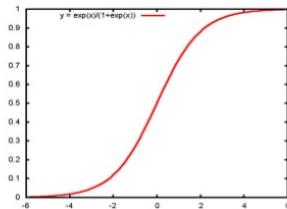
- Banque, crédit : un client va-t-il rembourser son prêt ? (*scoring*)

Consumer Credit Scoring using Logistic Regression and Random Forest

This is the *logistic regression* model. As $x \rightarrow \infty$, $\pi(x) \downarrow 0$ when $\beta < 0$ and $\pi(x) \uparrow 1$ when $\beta > 0$.

The odds are $\frac{\pi(x)}{1-\pi(x)} = \exp(\alpha + \beta x)$. The log odds called the *logit* has the linear relationship:

$$\text{logit}(\pi(x)) = \log \frac{\pi(x)}{1-\pi(x)} = \alpha + \beta x.$$



The curve in the above is defined by the equation $\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$. We can see that it is *S-shaped*.

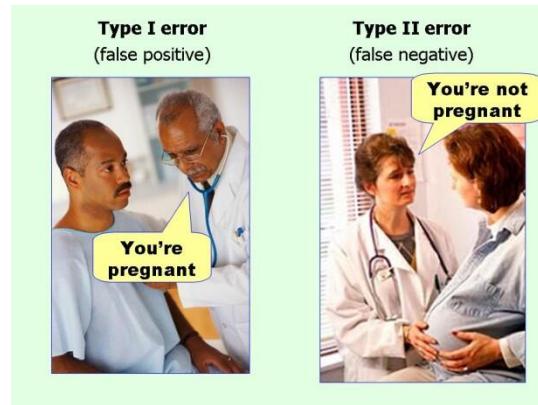
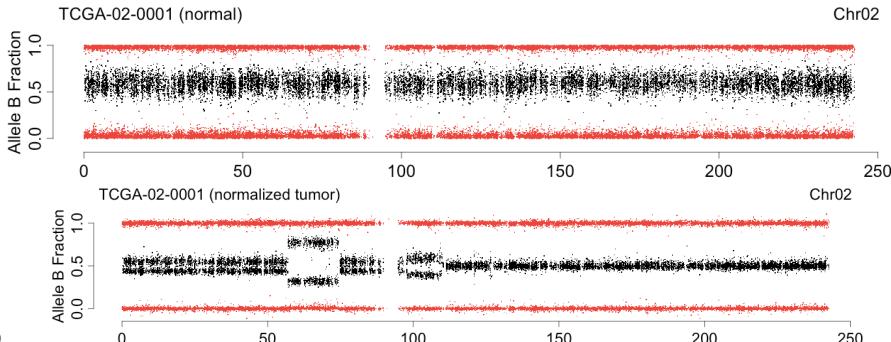
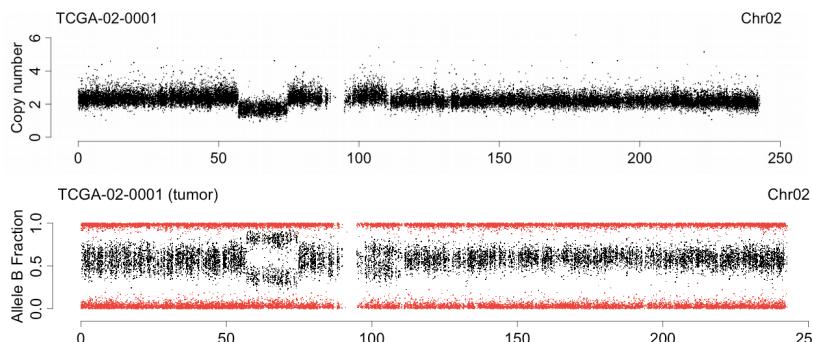
4.4 Fitting the logistic regression model:

Suppose we have a sample of n independent observations of the pair (x_i, y_i) , $i = 1, 2, \dots, n$, where y_i denotes the value of a dichotomous outcome variable and x_i is the value of the independent variable for the i th subject. Furthermore, assume that the outcome variable has been coded as 0 or 1, representing the absence or the presence of the characteristic, respectively. This coding for a dichotomous outcome is used throughout the text. Fitting the logistic regression model in equation to a set of data requires that we estimate the values of β_0, β_1 , the unknown parameters.

To fit a logistic regression model $\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$ to a set of data requires that the value of β_0, β_1 to be estimated. Now with some models, like the logistic curve, there is no mathematical solution that will produce explicit expressions for least square estimates of

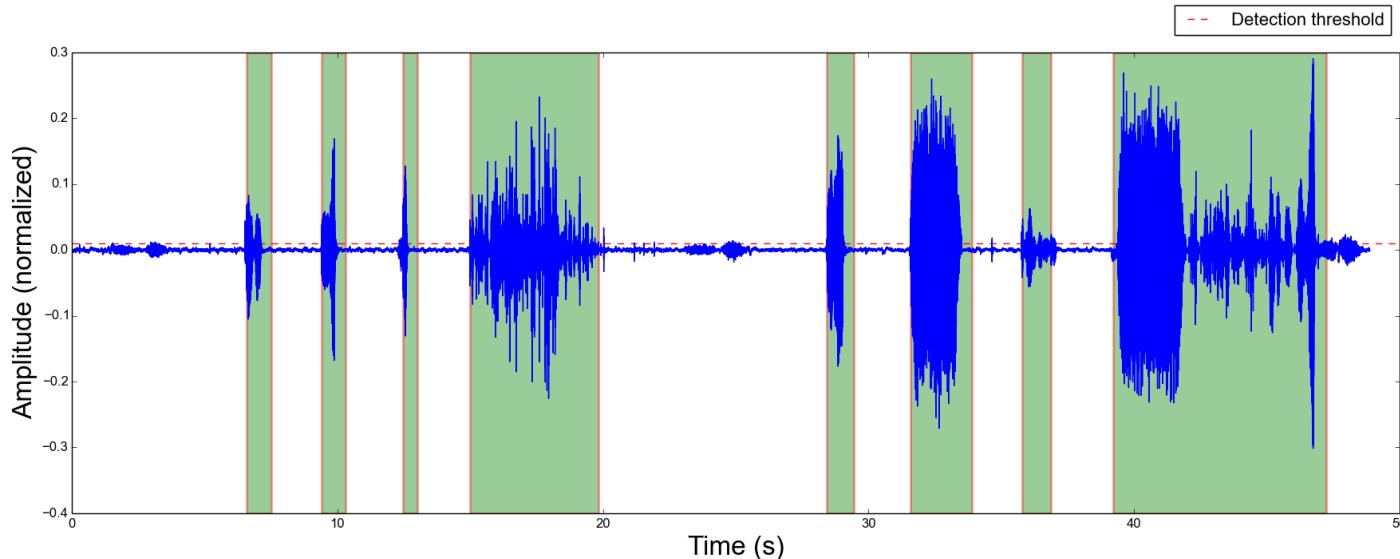
Les maths, c'est utile dans la vraie vie ?

- Banque, crédit : un client va-t-il rembourser son prêt ? (*scoring*)
- Biologie/médecine : segmentation de signaux, fiabilité d'un diagnostic, prédition de la survie d'un patient, test de l'effet d'un médicament, ...



Les maths, c'est utile dans la vraie vie ?

- Banque, crédit : un client va-t-il rembourser son prêt ? (*scoring*)
- Biologie/médecine : segmentation de signaux, fiabilité d'un diagnostic, prédition de la survie d'un patient, test de l'effet d'un médicament, ...
- Audio : segmentation de signaux pour détecter les différents sons



Les maths, c'est utile dans la vraie vie ?

- Banque, crédit : un client va-t-il rembourser son prêt ? (*scoring*)
- Biologie/médecine : segmentation de signaux, fiabilité d'un diagnostic, prédition de la survie d'un patient, test de l'effet d'un médicament, ...
- Audio : segmentation de signaux pour détecter les différents sons
- Marketing : segmentation client (classification), recommandations internet (netflix, amazon, ...)

The screenshot shows the official Netflix Prize website. At the top, a large yellow banner reads "Netflix Prize" and "COMPLETED". Below it, a message says "Congratulations!" and describes the purpose of the prize. It mentions the awarding of the \$1 million Grand Prize to BellKor's Pragmatic Chaos team. The page also links to the Leaderboard, Forum, and Rules.

The screenshot shows an Amazon product page for "Méthodes et 800 exercices corrigés de mathématiques : Pour réussir son Bac S" by Alain Gastineau. The page includes the book cover, price (EUR 19,00), and a list of related books under "Livres Scolaires".

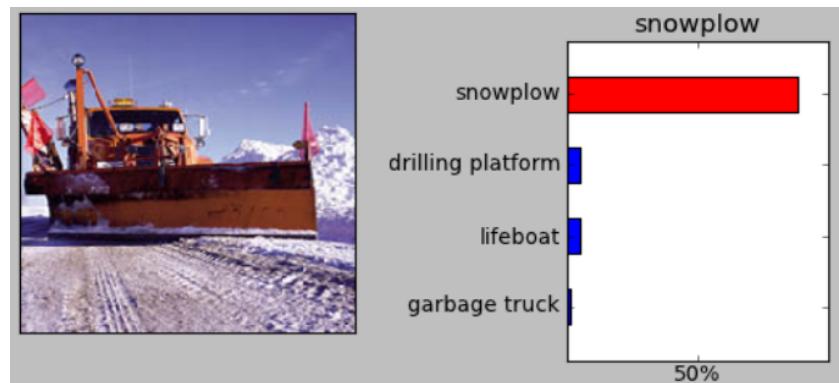
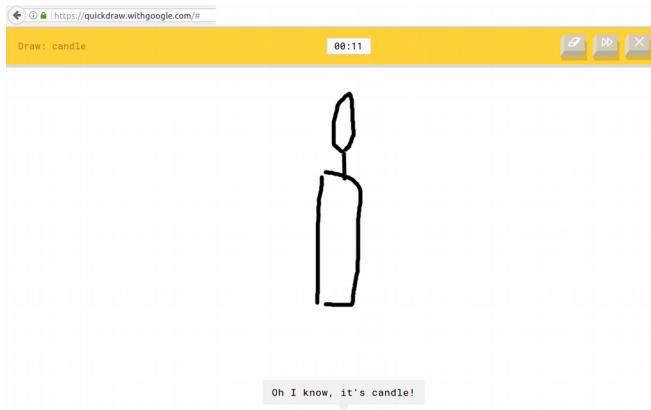
Méthodes et 800 exercices corrigés de mathématiques : Pour réussir son Bac S Relié – 28 août 2012
de Alain Gastineau • (Auteur)
Rentrée scolaire 2012 : livres, fournitures, cartables, ordinateurs, vêtements ...
★★★★★ • 10 commentaires client
Voir les formats et éditions
Référence : EUR 19,00
12 d'occasion à partir de EUR 9,00
5 neufs à partir de EUR 19,00
Vous avez envie de faire l'école le samedi 12 août ? Commandez-le dans les 7 h et 5 mins et choisissez la livraison en 1 jour ouvré au cours de votre commande. En savoir plus.
Note: Cet article est éligible à la livraison en points de collecte. Détails
Livrables scolaires

Les clients ayant acheté cet article ont également acheté

Produit	Prix	Avis
METHOD'S Physique-Chimie Terminale S	EUR 19,00 prime	★★★★★ 15
METHOD'S Les Exos de mathématiques Terminale S	EUR 19,00 prime	★★★★★ 3
500 Formules et propriétés à connaître pour le Bac S	EUR 2,99 prime	★★★★★ 111
Pour Réussir Sa Première S	EUR 19,00 prime	★★★★★ 20
METHOD'S Mathématiques Terminale S Conforme au Programme 2012	EUR 18,00 prime	★★★★★ 1
METHOD'S Méthodes et 300 Exercices Corrigés de Maths Terminale S	EUR 19,00 prime	★★★★★ 1
MATHS Spécialité Terminale S	EUR 12,22 prime	★★★★★ 2
Pour Réussir Sa Seconde : Méthodes et 300 Exercices Corrigés de Maths Terminale S	EUR 19,00 prime	★★★★★ 1

Les maths, c'est utile dans la vraie vie ?

- Banque, crédit : un client va-t-il rembourser son prêt ? (*scoring*)
- Biologie/médecine : segmentation de signaux, fiabilité d'un diagnostic, prédition de la survie d'un patient, test de l'effet d'un médicament, ...
- Audio : segmentation de signaux pour détecter les différents sons
- Marketing : segmentation client (classification), recommandations internet (netflix, amazon, ...)
- Reconnaissance de formes



Les maths, c'est utile dans la vraie vie ?

- Banque, crédit : un client va-t-il rembourser son prêt ? (*scoring*)
- Biologie/médecine : segmentation de signaux, fiabilité d'un diagnostic, prédition de la survie d'un patient, test de l'effet d'un médicament, ...
- Audio : segmentation de signaux pour détecter les différents sons
- Marketing : segmentation client (classification), recommandations internet (netflix, amazon, ...)
- Reconnaissance de formes
- Construction, architecture



GOLDEN RATIO 1.618

Based on the Fibonacci sequence, the Golden Ratio describes the relationship between two consecutive Fibonacci numbers. These follow a 1:1.618 ratio - this is what we refer to as the Golden Ratio.

Throughout history, the ratio for length to width of rectangles of 1:1.618 has been considered the most pleasing to the eye.

GOLDEN SPIRAL **GOLDEN TRIANGLE**

In architecture, the Parthenon is a famous example of the golden ratio being extensively used in the exterior dimensions of a building.

Les maths, c'est utile dans la vraie vie ?

- Banque, crédit : un client va-t-il rembourser son prêt ? (*scoring*)
- Biologie/médecine : segmentation de signaux, fiabilité d'un diagnostic, prédition de la survie d'un patient, test de l'effet d'un médicament, ...
- Audio : segmentation de signaux pour détecter les différents sons
- Marketing : segmentation client (classification), recommandations internet (netflix, amazon, ...)
- Reconnaissance de formes
- Construction, architecture
- Distribution : optimisation des stocks, prédictions des ventes
- ...

Les maths, c'est utile dans la vraie vie ?



a-volute
3D Sound Expert



ArcelorMittal



Merci



Inria Lille – Nord Europe
www.inria.fr