UBO

lab-sticc.univ-brest.fr/~babau/

Analyse de données

Jean-Philippe Babau

Département Informatique, UFR Sciences, UBO
Laboratoire Lab-STICC

jean-philippe.babau@univ-brest.fr

UBO

Objectifs du cours

- Appréhender le domaine de l'analyse de données
- Appliquer sur des exemples de traitement de données
  - Environnement Knime

jean-philippe.babau@univ-brest.fr

UBO

Pourquoi le Data Mining ?

- De plus en en plus de *capteurs* (web, objets connectés, senseurs, …)
- Des capacités de stockage importantes

- De plus en plus de données
- du tera (10^{12}) octets au peta (10^{15}) octets
- 2018 ; chaque seconde, 29.000 Gigaoctets (Go) d'informations sont publiées dans le monde, soit 2,5 trillions d'octets de données chaque jour et 88% de ces données disponibles ne sont pas analysées

- Besoin fort en analyse automatique des données

- Transformer des données (des faits) en informations

jean-philippe.babau@univ-brest.fr

UBO

Des exemples

- Domaine bancaire

si (salaire annuel élevé (>= 30 000) et propriétaire = vrai)
alors (risque de défaut de paiement = faux)

si (salaire mensuel élevé et achats Web nombreux)
alors (risque de découvert important)

- Vente

« les gens qui achètent des couches culottes le week-end, achètent souvent de la bière »

jean-philippe.babau@univ-brest.fr

UBO

Un peu d'histoire

- 1960 : création des bases de données
- 1970 : les bases de données relationnelles
- 1980 : les bases de données objet, spatiales, …
- 1990 : les entrepôts de données, le data mining
- 2000 : la globalisation, les applications

jean-philippe.babau@univ-brest.fr

UBO

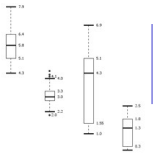
Les objectifs du data mining

- Aider l'expert métier à analyser un flot de données important
- Extraire une connaissance
  - non triviale et implicite
  - utile et pertinente

jean-philippe.babau@univ-brest.fr

UBO

Des données à la connaissance



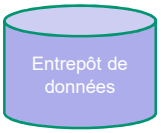
statistiques

arbre de décision

motifs

clusters

extraction,
transformations

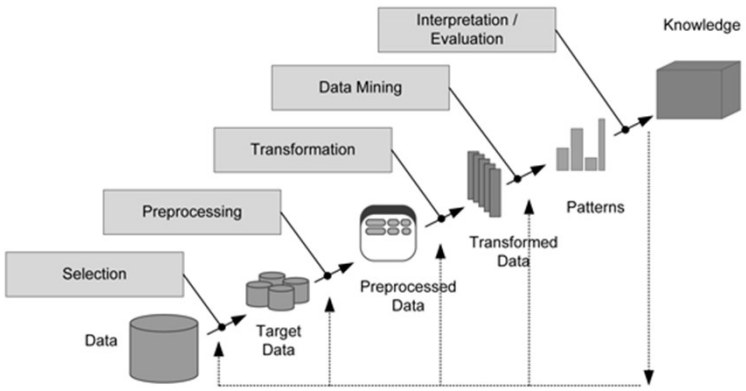


jean-philippe.babau@univ-brest.fr

UBO

Le processus KDD

- Knowledge Discovery in Databases



jean-philippe.babau@univ-brest.fr

UBO

Les étapes d'un processus d'analyse de données

- Extraction de multiples sources
- Nettoyage des données
- Transformation des données
- Intégration de données dans une base commune
- Exploration des données
- Sélection des données utiles

- *Analyse des données*

- Présentation des résultats
- Interprétation des résultats et itérations

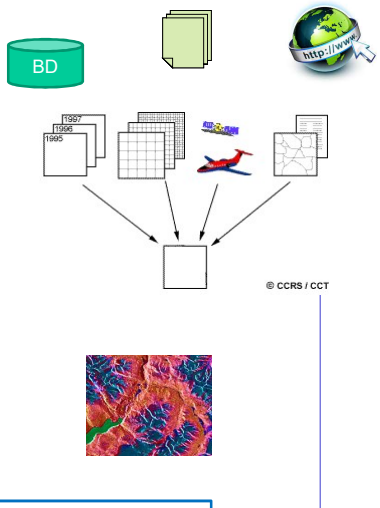
jean-philippe.babau@univ-brest.fr

UBO

Extraction de multiples sources

- Sources hétérogènes
  - fichiers
  - Web
  - BD relationnelles, objet, géographiques, …
  - données capteur (séries temporelles, …)

  - données structurées
  - images
  - textes

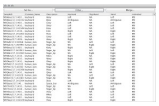


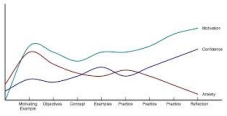
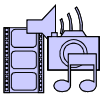
jean-philippe.babau@univ-brest.fr

UBO

Les types de données manipulées

- Tableau, matrices
  - à n dimensions
  - données géo-référencées : longitude, latitude, altitude, temps
  - représentation sous forme de tables
- Graphes
  - réseaux sociaux
- Textes
  - documents, rapports, …
- Séquence
  - série temporelle, trajectoire, …
- Multimédia
  - image, vidéo,




jean-philippe.babau@univ-brest.fr

UBO

Attributs des données

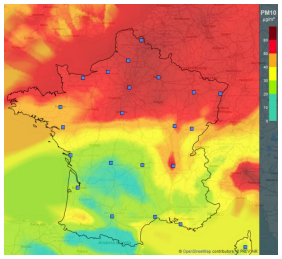
- Une donnée est caractérisée par des attributs
- Types d'attributs
  - nominal (noms)
  - date
  - booléen, énuméré, ordinal
  - valeur numérique (entier ou réel)
  - intervalles de valeurs numériques (entier ou réel)
- Types d'attributs
  - discret
  - continu (interpolation possible)

jean-philippe.babau@univ-brest.fr

UBO

Qualité de la donnée

- Vérification
  - valeurs crédibles
  - valeurs corrélées
  - valeurs validées
  - analyses spécifiques pour la vérification



- Corrections
  - manuelles : processus couteux
  - automatiques : attention aux modifications

jean-philippe.babau@univ-brest.fr

UBO

Nettoyage de la donnée

- Problèmes
  - valeur manquante
    - problème de saisie, de transmission, effacement
  - valeur de mauvaise qualité
    - Incorrecte, inconsistante, aberrante ou non crédible
    - QC :
  - erreur de mesure
    - qualité du capteur (biais, dérive, …)
  - duplication de données
- Correction
  - ajout d'une valeur par défaut
    - constante
    - valeur moyenne, valeur estimée
  - interpolation, lissage
  - correction de format
  - saisie manuelle
  - élimination
  - ajout d'un code qualité

jean-philippe.babau@univ-brest.fr

UBO

Exemple : SeaDataNet

- Data Quality Procedure
  - Normes
  - Tests automatiques et manuels
  - Obligations de commentaires
- MetaData : méta-données obligatoires
  - Lieu et date de la mesure
  - Type d'instrument et de mesure
  - Organisme de collecte
  - Opérations effectuées sur la donnée
  - Commentaires sur la mesure
- Vocabulaire et unités imposées
- Tests automatiques
  - Format des dates, latitudes (-90/90) et longitudes (-18//180), position (mer)
  - Intervalles (min/max)
  - Spécifique : la pression augmente avec la profondeur, la variation est limitée …
- Tests manuels
  - Spécifiques à la donnée

jean-philippe.babau@univ-brest.fr

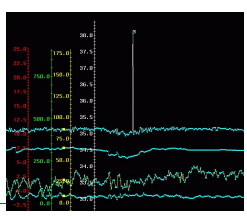
UBO

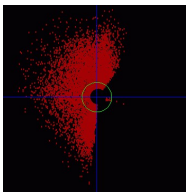
Exemple : SeaDataNet

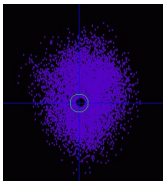
Test 9 : gradient
This test is failed when the difference between adjacent measurements is too steep.
Test value = | V2 - (V3 + V1)/2 |
where V2 is the measurement being tested as a spike, and V1 and V3 are the previous and next values.
• Temperature: The V2 value is flagged when the test value exceeds 9.0 degree C.
• Salinity: The V2 value is flagged when the test value exceeds 1.5 PSU

Values that fail the test (i.e. value V2) should be flagged as wrong.

Test 11 : instrument comparison
If two different sensors measure a same parameter, the difference between 2 measurements should not be greater than a fixed limit.
Example : on research vessels the difference between the temperature of the tank of the TSG and the measurement of the hull monted temperature sensor should be less than 1° Celsius. If the test fails, the measurements of both sensors are flagged as wrong.







jean-philippe.babau@univ-brest.fr

UBO

Exemple : SeaDataNet

Key	Entry Term	Abbreviated term	Term definition
0	no quality control	none	No quality control procedures have been applied to the data value. This is the initial status for all data values entering the working archive.
1	Good value	good	Good quality data value that is not part of any identified malfunction and has been verified as consistent with real phenomena during the quality control process.
2	probably good value	probably_good	Data value that is probably consistent with real phenomena but this is unconfirmed or data value forming part of a malfunction that is considered too small to affect the overall quality of the data object of which it is a part.
3	probably bad value	probably bad	Data value recognised as unusual during quality control that forms part of a feature that is probably inconsistent with real phenomena.
4	bad value	bad	An obviously erroneous data value.
5	changed value	changed	Data value adjusted during quality control. Best practice strongly recommends that the value before the change be preserved in the data or its accompanying metadata

UBO

Exemple : SeaDataNet

Key	Entry Term	Abbreviated term	Term definition
6	value below detection	BD	The level of the measured phenomenon was too small to be quantified by the technique employed to measure it. The accompanying value is the detection limit for the technique or zero if that value is unknown.
7	value in excess	excess	The level of the measured phenomenon was too large to be quantified by the technique employed to measure it. The accompanying value is the measurement limit for the technique.
8	interpolated value	interpolated	This value has been derived by interpolation from other values in the data object.
9	missing value	missing	The data value is missing. Any accompanying value will be a magic number representing absent data.
A	value phenomenon uncertain	ID_uncertain	There is uncertainty in the description of the measured phenomenon associated with the value such as chemical species or biological entity.

jean-philippe.babau@univ-brest.fr

UBO

Transformation des données

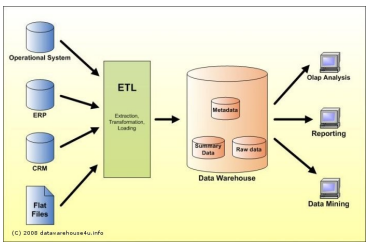
- Agrégation de données représentant la même information
  - #super #genial #génial
- Normalisation
  - normes (long, longitude) et unités (°C <->°F)
  - format, types, précision
  - standardisation : harmonisation d'évaluations laxistes et sévères (moyenne/écartType)
- Numérisation
  - catégorie -> valeur : génial : 5, super : 4, ok : 3, moyen : 2, faible : 1, nul : 0
- Filtrage
  - élimination du bruit
- Abstraction
  - les données deviennent des concepts
  - Un vent entre 0° et 15° est considéré comme un « vent de face »
- Attention à conserver la correction des données

jean-philippe.babau@univ-brest.fr

UBO

Intégration des données

- Entrepôts de données
  - un lieu centralisé pour stocker des données en provenance de sources multiples de manière homogène, multidimensionnelle pour des extractions et analyses efficaces
  - structure en cube
  - structure adaptée au requêtes



jean-philippe.babau@univ-brest.fr

UBO

La sélection des données

- Données pertinentes pour l'étude
- Filtrage
  - Conserver uniquement les données utiles
    « le nom n'est pas important pour des études sur le poids des personnes »
  - Conserver uniquement les valeurs utiles
    « on sélectionne une population spécifique pour l'étude »

jean-philippe.babau@univ-brest.fr

UBO

Exploration des données

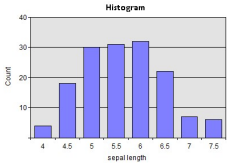
- Comprendre la donnée
  - Analyse statistique
  - Analyse mathématique
  - Analyse graphique
- Première analyse de données

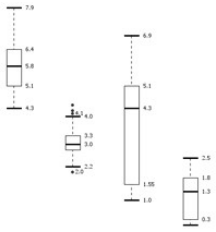
jean-philippe.babau@univ-brest.fr

UBO

Analyse de données numériques

- Analyse statistique
  - minimum, quartile (25%), moyenne, quartile (75%), maximum
  - médiane, variance, écart-type
  - distribution
  - données dépendantes Vs indépendantes
- Analyse graphique
  - box plot
  - histogrammes



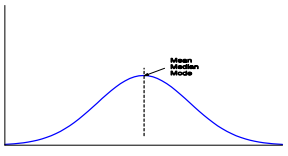


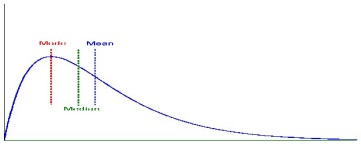
jean-philippe.babau@univ-brest.fr

UBO

Distribution de données numériques

- Distribution Gaussienne
  - 95 % à moyenne ± 2 σ
  - 99,7% à moyenne ± 3 σ
- Inégalité de Bienaymé-Tchebychev
  - 89 % à moyenne ± 3 σ
- La médiane n'est pas la moyenne





- Estimateurs de distribution
  - Évaluation de la distribution sur une population représentative

jean-philippe.babau@univ-brest.fr

UBO

Corrélation de données

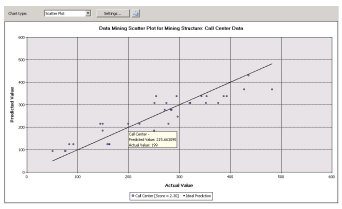
- Evaluation de l'indépendance / dépendances de données
- Evaluation d'une loi de distribution
- Test du χ^2
  - Evaluation d'une distribution vis à vis d'une hypothèse statistique
  - Test de la non-corrélation entre deux variables
  - Résultat significatif si $p < 0,05$

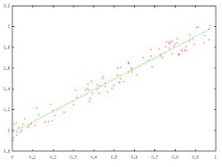
jean-philippe.babau@univ-brest.fr

UBO

Analyse mathématique

- Régression linéaire : un ensemble de valeurs caractérisé par une droite
- Comparaison données réelles / prévues
  - scatter plot





jean-philippe.babau@univ-brest.fr

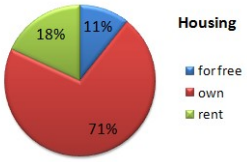
UBO

Analyse de données énumérées

- Analyse en nombre et en fréquence

Frequency Table

Housing	Count	Count%
for free	96	10.67%
own	641	71.22%
rent	163	18.11%



jean-philippe.babau@univ-brest.fr

UBO

Discrétisation

- Réduction de l'espace d'entrée
  - préparation de l'analyse
- Regroupement d'informations
  - ensemble de valeurs -> une valeur représentative
- Basée sur l'exploration des données
  - analyse statistique, analyse graphique
  - intégration de considérations métier
  - guidé par les attributs
- Techniques
  - binning
    - découpage régulier en taille, en nombre d'éléments
    - découpage en domaine (jeune, adulte, vieux)
  - hiérarchisation et perte de précision (jour, mois, année)
  - clustering

jean-philippe.babau@univ-brest.fr

UBO

Des données aux connaissances

- Analyse descriptive : trouver les caractéristiques des données fouillées
- Analyse prédictive : consiste à faire de l'inférence à partir des données actuelles pour prédire des évolutions futures

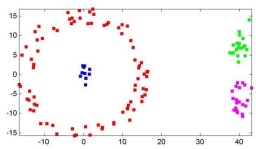
« Ce que les humains apprennent provient
de la reconnaissance de formes
(estimé à environ 300 millions de motifs) »
Ray Kurzweil

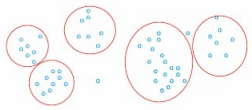
jean-philippe.babau@univ-brest.fr

UBO

Clustering (classification non supervisée)

- Regrouper les données en classe de valeurs
- Regroupement
  - partition
  - hiérarchie
- Techniques
  - étude basée sur une distance
    - méthode k-means
  - étude des densités
  - étude de voisinage





jean-philippe.babau@univ-brest.fr

UBO

Algorithme k-means

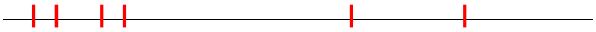
- Étant donné un ensemble d'éléments E=(e_1, e_2, …, e_n), on cherche à partitionner les *n* éléments en *k* ensembles S = {S_1, S_2, …, S_k} (*1* ≤ *k* ≤ *n*)
- Objectif : minimiser la distance entre les points à l'intérieur de chaque ensemble
  - calcul du *barycentre*
  - distance moyenne au barycentre
- k, algorithme de calcul de distance : entrées du problème
- Il existe une heuristique
- Problème de prise en compte des éléments isolés et des répartitions non denses

jean-philippe.babau@univ-brest.fr

UBO

Algorithme k-means : exemples

- E={1, 2, 4, 5,15, 20}

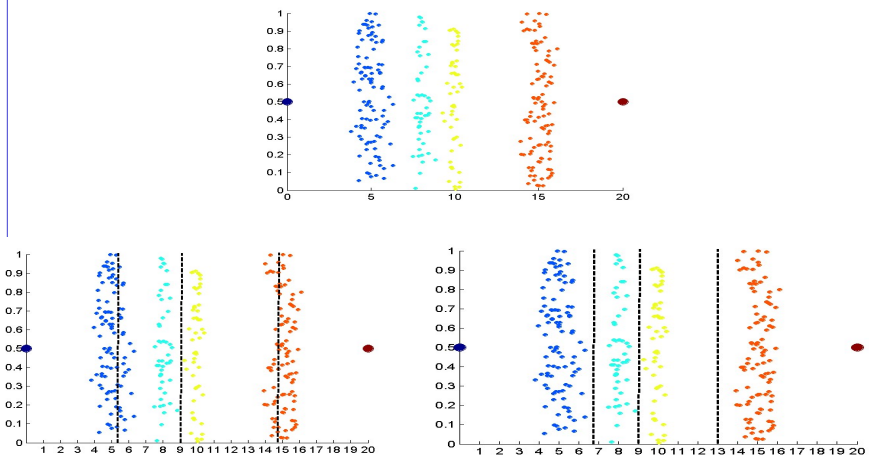


- {v1,v2,… } (barycentre; *distance moyenne au barycentre*)
- Moyenne des distances moyennes (distances minimales entre ensembles)
- k= 1 -> {1, 2, 4, 5,15, 20} (7,83; *6,44*) 6,44 (*0*)
- k= 2 -> {1,2,4,5} *(3;1,5*) {15,20} *(17,5;2,5*) 2 (10)
- k= 3 -> {1, 2} (1,5;*0,5*) {4, 5,} (4,5;*0,5*) {15, 20} (17,5;*2,5*) 1,16 (2,10,13)
- k= 3 -> {1, 2, 4, 5} *(3;1,5)* {15 } *(15;0*) { 20} *(20;0*) 0,5 (10, 5 15)
- k= 4 -> {1, 2} (1,5;*0,5*) {4, 5,} (4,5;*0,5*) {15 } *(15;0*) { 20} *(20;0*)
  0,25 (2,13,18,10,15,5)
- k= 6 -> {1} {2} {4} {5} {15} {20} 0 (1,3,4,14,19, 2,3,13,18,1,11,16,10,15,5)

jean-philippe.babau@univ-brest.fr

UBO

Binning / Clustering



jean-philippe.babau@univ-brest.fr

UBO

Qualité du clustering

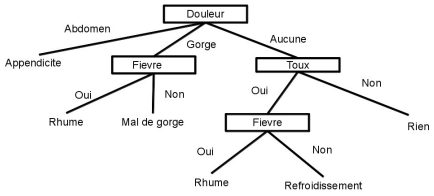
- Mise en place de principes de similarité
  - Basée sur la notion de distance
  - Distance simple d(x,y) = |x-y| ou |x-y| / dmax
  - Distance combinée : Manhattan (somme des distances) Euclidienne (racine carrée de la somme des distances au carré), Minkowski, …
  - Distance entre énumérés, images, textes, …
- Similarités inter classes faible
- Similarités intra classes importante
- Capacité à découvrir des patterns

jean-philippe.babau@univ-brest.fr

UBO

Arbre de décision

- Une donnée -> un ensemble d'attributs
- Un arbre : outil de classification
  - nœud : test d'un attribut
  - branche : valeur de l'attribut
  - feuille : classe



Fievre	Douleur	Toux	Maladie
oui	Abdomen	non	Appendicite
non	Abdomen	oui	Appendicite
oui	gorge	non	rhume
oui	gorge	oui	rhume
non	gorge	oui	mal de gorge
oui	non	non	aucune
oui	non	oui	rhume
non	non refroidissement	oui	
non	non	non	aucune

jean-philippe.babau@univ-brest.fr

UBO

Recherche de motifs (patterns)

- Règles d'association ou *frequent itemset*
  - *Item* : un fait (donnée validée, normalisée, sélectionnée et discrétisée)
  - *ItemSet* ou transaction : ensemble de faits
  - corrélation et causalité : X1 → X2 [support, confiance]
  - support : probabilité d'avoir X1 et X2
  - confiance : probabilité d'avoir X2 lorsqu'on a X1

- Exemples

  achat couches culottes → achat bière [0.5%, 75%]

  achat bière → achat couches culottes [0.5%, 2%]

  age([20,29]) & revenu([20,30]→ achète(PC) [2%, 60%]

  PC contient("Linux") → PC contient("OpenOffice") [30%, 75%]

jean-philippe.babau@univ-brest.fr

UBO

Classification

- Classification et prédiction
  - modèle par apprentissage
  - on prend un échantillon représentatif (jeu d'essai) dans lequel chaque donnée est associé à une classe (définie a priori)
  - on cherche ensuite les caractéristiques communes de la classe
  - classification automatique sur les nouvelles données
- Technique
  - régression linéaire ou polynomiale (série temporelle)
  - réseaux Bayésien (jeux d'attributs)
    - basée sur une notion de distance

jean-philippe.babau@univ-brest.fr

UBO

Prédiction

- Préparation des données
  - nettoyage, discrétisation, normalisation, renommage, fusion, …
- Partition
  - Données pour la construction du modèle /données pour le test du modèle
- *Learner* (produit le modèle) et *Predictor (interprète le modèle)*
  - Selon la technique choisie
- Prédiction sur les séries temporelles
  - À l'instant t : dépend des données à l'instant (t-i), (t-2i), … (t-nxi)
  - i est appelé lag
  - découpage de la time series selon la saisonnalité (cycles)
- Evaluation de l'erreur

jean-philippe.babau@univ-brest.fr

UBO

Recherche d'exceptions

- Entité hors classification
- Comportement anormal
- Comportement déviant
- Détection de fautes
- Sécurité

jean-philippe.babau@univ-brest.fr

UBO

Interprétation

- Présentation adaptée à la décision
  - graphiques, tableaux, cartes
  - basée sur l'exemple : trouver des cas particuliers représentatifs
- Interprétation des résultats avec l'expert du domaine
  - séparer les résultats évidents et résultats non évidents
  - extraire des résultats crédibles
- Un système de data mining peut générer des milliers de motifs,
  … pas tous intéressants.

jean-philippe.babau@univ-brest.fr

UBO

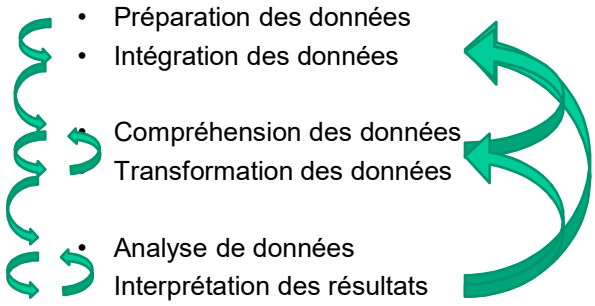
Interprétation

- Sélection de motif avec l'expert du domaine
  - résultats intéressants
- C'est quoi un motif intéressant ?
- **Mesure d'intérêt** : un motif est intéressant s'il est facilement compréhensible (data miner), possède un degré de certitude (support, confiance), nouveau (expert), peut servir à valider ou invalider une hypothèse utilisateur (expert)

jean-philippe.babau@univ-brest.fr

UBO

Processus itératif



jean-philippe.babau@univ-brest.fr

Conclusion

- Comprendre le domaine métier
  - processus en lien fort avec l'expert
  - la qualité de la donnée est primordiale
- Préparation des données
  - environ 60% de l'effort
- Trouver la bonne technique d'analyse
  - interprétation avec l'expert
- Mettre en place un processus complet
  - itératif

Références

- http://hanj.cs.illinois.edu/bk3/bk3_slidesindex.htm
- https://wiki.cites.illinois.edu/wiki/display/cs512/Lectures
- http://www.lifl.fr/~talbi/Cours-Data-Mining.pdf