« Court » de Data mining et Machine Learning *IAM 2014 - 2015*

14/10/2014

Frédéric Precioso
Laboratoire I3S – UMR UNS-CNRS 7271
Pôle GLC – Equipe MinD

Plan du Cours

1. **Introduction : classification et prédiction :** Ensembles d'apprentissage et de test, taux d'erreur, sur-apprentissage. Définition de distances, Problème des variables continues, Evaluation de la classification, Interprétation des classes obtenues.
2. **Description des méthodes par plus proches voisins :** Méthodes par partitionnement, exemple des K-Moyennes ou « *k-means* », des approches hiérarchiques et autres méthodes évoluées par densité.
3. **Techniques de classement par arbres de décision – Forêts aléatoires**
4. **Introduction aux Machines à noyaux :** Techniques de classification et de prédiction par Machine à Vecteurs de Supports. Fonctions de similarité noyaux..

De la Statistique …

- Quelques centaines d'individus,
- Quelques variables,
- Fortes hypothèses sur les lois statistiques suivies,
- Importance accordée au calcul,
- Échantillon aléatoire.

… au Data mining

- Des millions d'individus,
- Des centaines de variables,
- Données recueillies sans étude préalable,
- Nécessité de calculs rapides,
- Pas un échantillon aléatoire.

Systèmes d'information et data mining

- Les données forment le cœur des processus de base dans la plupart des entreprises.
- L'archivage des données crée la mémoire de l'entreprise.
- L'exploitation des données « data mining » crée l'intelligence de l'entreprise.

De OLAP à la Fouille de données

- **OLAP**
  - oblige l'utilisateur à formuler une question précise qui fait l'objet d'une requête ad hoc pour fournir un résultat factuel : « *Combien de chaussures de taille 42 ai-je vendu ces trois derniers mois ?* ».
  - ne fait que des comptages et pas de prévisions. Le résultat permet de valider une hypothèse ou apporte une information en vue d'une appréciation.
  - se concentre généralement sur des faits actuels, des utilisations prédéfinies de données agrégées, l'établissement de résultats factuels par des requêtes ad hoc finalisées sous la forme de rapports.

De OLAP à la Fouille de données

- **La fouille de données**
  - A partir d'un ensemble de données, des techniques d'exploration sont appliquées pour trouver des relations, souvent complexes, et des modèles inconnus qui ont du sens : « *Combien de chaussure d'été de taille 42 devrais-je commander pour la saison prochaine ?* ».
  - se concentre généralement sur des tendances, des estimations, des découvertes ou des prévisions, exige des données détaillées, met en œuvre des techniques statistiques, des algorithmes, et établit des modèles (explicites ou implicites, complets ou partiels).

L'exploitation des données est devenue une réalité industrielle

- Les techniques d'exploitation des données existent depuis des années.
- L'utilisation de ces techniques dans l'industrie est cependant beaucoup plus récente parce que:
  - Les données sont produites électroniquement,
  - Les données sont archivées,
  - La puissance de calcul nécessaire est abordable,
  - Le contexte est ultra-concurrentiel,
  - De nombreux algorithmes pour l'exploitation des données ont émergés.

Data mining ou …

- KDD (Knowledge Discovery in Databases)
- Fouille de données (terme français)
- Extraction automatique de connaissances à partir de données (ECD)
- Recherche d'Information (Information Retrieval)

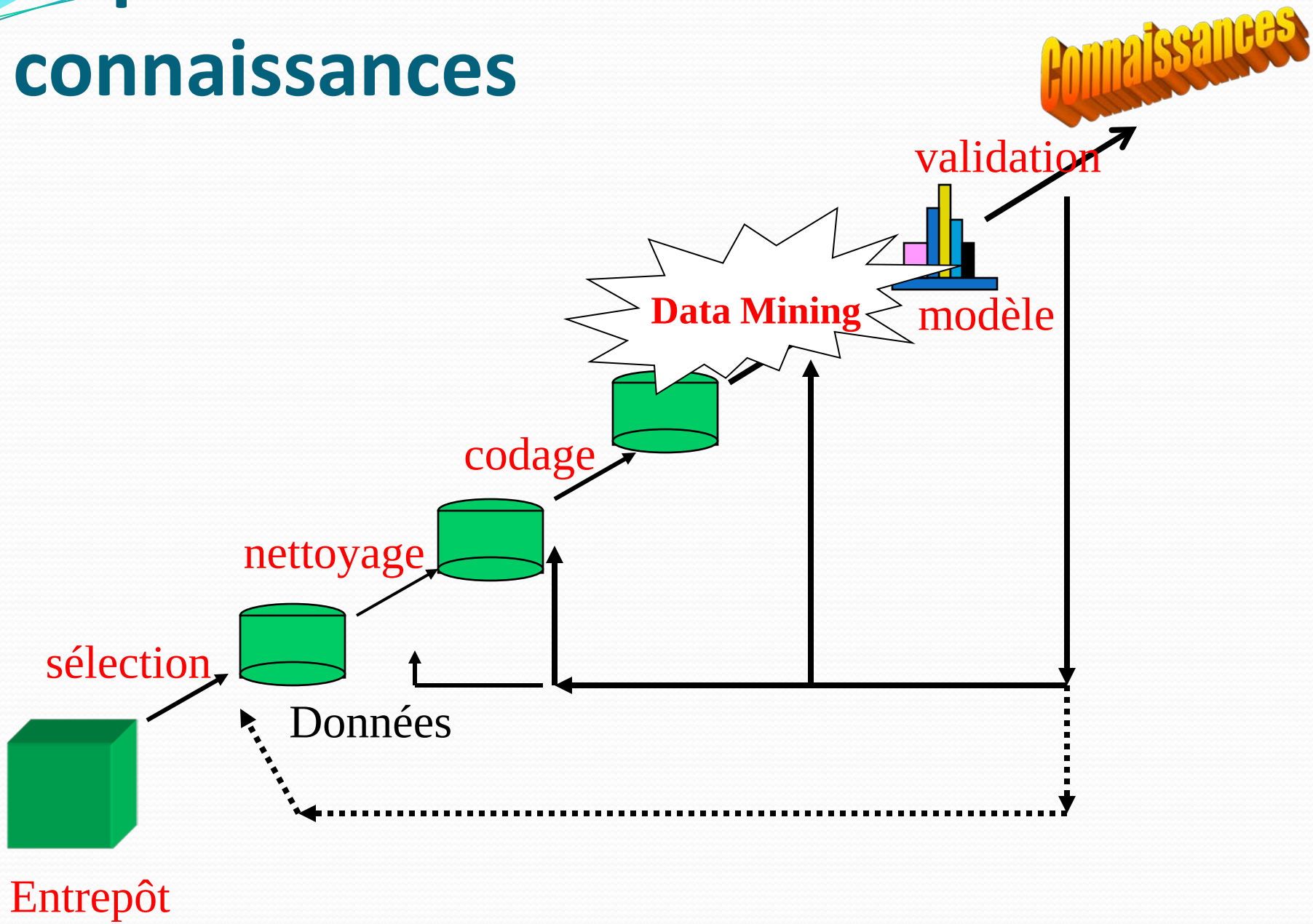
Extraction de connaissances à partir des données

- Dans de grands volumes de données:
  - «Extraction d'informations originales auparavant inconnues, potentiellement utiles»
  - «découverte de nouvelles corrélations, tendances et modèles»
  - «processus d'aide à la décision en cherchant des modèles d'interprétation des données»

Data mining : « La famille »

- Statistiques, analyse des données
- Apprentissage automatique (Machine Learning)
- Reconnaissance de formes
- Bases de données
- Entrepôt de données (Data Warehouse)
- Visualisation des données

Le processus d'extraction de connaissances



Le processus général de découverte de connaissances dans les données

1. Poser le problème
2. Recherche des données
3. Nettoyage des données
4. Codage des données, actions sur les variables
5. Recherche d'un modèle, de connaissances, d'information
   (Data mining)
6. Validation et interprétation du résultat, avec retour possible sur les étapes précédentes
7. Intégration des connaissances apprises

1. Poser le problème

- c'est comprendre le domaine d'application, la connaissance déjà existante et les buts de l'utilisateur final.
- Quel type de problème a-t-on à traiter ? on connaît les classes on veut identifier les facteurs d'affectation, ou on veut créer les classes facteurs de différentiation.
- Si on met en évidence de nombreux groupes de clients, dans une étude de marketing pourra-t-on revoir les processus marketing pour chaque groupe ?

1. Poser le problème: un exemple

- Un éditeur vend 5 sortes de magazines : sport, voiture, maison, musique et BD.
- Les questions qu'il se pose:

1. *Combien de personnes ont pris un abonnement à un magazine de sport cette année?*
2. *A-t-on vendu plus d'abonnements sport cette année que l'année dernière?*
3. *Est-ce que les acheteurs de magazines de BD sont aussi amateurs de sport?*
4. *Quelles sont les caractéristiques principales de mes lecteurs de magazines de voiture?*
5. *Peut-on prévoir les pertes de clients et prévoir des mesures pour les diminuer?*

1 et 2 sont de simples requêtes. Dans 2 on a la notion de temps donc les données doivent être historisées.

Pour 3, la réponse pourrait être une valeur estimant la proba que la règle soit vraie. 3 peut être généralisée : on peut chercher des associations fréquentes entre acheteurs de magazine.

4 est plus ouverte, 5 aussi c'est vraiment le domaine de la fouille de données

2. Recherche des données

- Données existantes ou à constituer
  - Entrepôt de données (Data Warehouse), magasin de données, Bases de données relationnelles, Bases de données temporelles, Web,…

- Échantillon ou travail sur toutes les données: dépend des données disponibles, de la puissance machine, de la fiabilité souhaitée. Très souvent le travail sur un échantillon est bien adapté au data mining qui est un processus itératif.

Entrepôt de données (Data Warehouse)

- « collection de données orientées pour un sujet, intégrées, non volatiles et historisées, organisées pour le support du processus d'aide à la décision ».

- base de données dans laquelle sont déposées après nettoyage et homogénéisation les informations en provenance des différents systèmes de production de l'entreprise.

  L'entrepôt facilite le data mining mais le data mining peut se faire aussi sur des données extraites pour l'occasion.

Types de bases de données

- Il existe plusieurs types de structure de bases de données :
  - “flat file”
    - Toute l’information du client est contenue dans un même ficher qui peut être de longueur variable
  - Relationnelle
    - L’information du client est contenu dans plusieurs fichiers unis par une « clef » commune, par exemple le numéro du client

3. Nettoyage des données

- Doublons, erreurs de saisie, pannes de capteurs…
- Valeurs aberrantes : rechercher les pics, les valeurs en dehors d'un espace déterminé par la moyenne et un certain nombre d'écarts-type, outils de visualisation : histogrammes, nuages de points, …
- Informations manquantes : exclure les enregistrements incomplets, remplacer les données manquantes (valeur moyenne, valeur par défaut), garder les manquants si la méthode de fouille sait les gérer

Valeurs manquantes

On peut :

- Ignorer l'observation
- Utiliser la valeur moyenne (la pire !!)
- Utiliser la valeur moyenne pour les exemples d'une même classe
- Utiliser la régression (plus précise mais plus complexe)

Valeurs aberrantes

- Il convient de définir une stratégie pour traiter les valeurs aberrantes (données hors norme) ou éventuellement de développer quand même un modèle reposant sur ces valeurs :
  - Par exemple, si l'objectif est de prévoir les taux de fréquentation et les revenus de rencontres sportives, il faut certainement éliminer les chiffres de fréquentations anormales dues à des événements particuliers, grève des transports, etc.…
  - Au contraire dans le cas de la détection de fraudes, il peut-être pertinent de se concentrer sur certaines valeurs aberrantes car elles sont peut-être la représentation de transactions frauduleuses.

4. Codage des données

- Agrégation (somme, moyenne)
- Discrétisation (réduire le nombre de valeurs d'une variable continue en divisant le domaine de valeurs en intervalles)
- Codage des attributs discrets
- Uniformisation d'échelle ou standardisation
- Construction de nouvelles variables

Cette étape de choix des bonnes variables peut être déterminante pour le succès du processus de fouille.

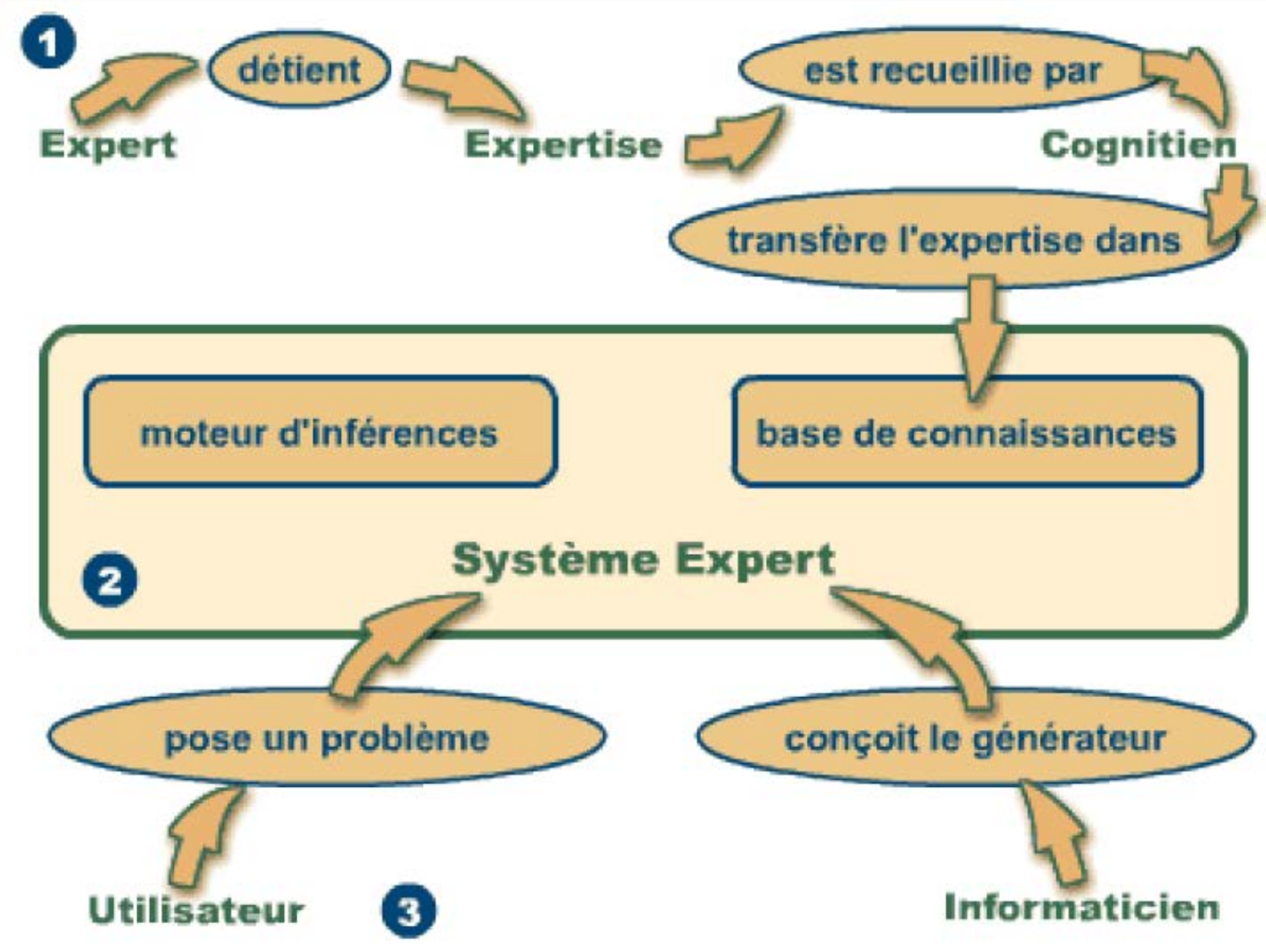
Exemples d'actions sur les variables

- Transformation d 'une variable :
  - Transformation des données géographiques : ville code postal en données géographiques (longitude, latitude) permettant de prendre en compte la proximité des lieux dans le raisonnement (géocodage utilisé en géomarketing)
  - Transformation des dates en durées : ancienneté d 'un client, durée entre l 'envoi d 'un catalogue et la 1ère commande

Exemples d'actions sur les variables

- Transformation multi-variables :
  - combiner plusieurs variables en une nouvelle variable agrégée, combinaison linéaire ou non-linéaire de plusieurs variables : revenu et nombre d'enfants combinés par Revenu/nbre d'enfants

Système Expert



Data mining vs Systèmes Experts

- ***Déduction : base des systèmes experts***
  - schéma logique permettant de déduire un théorème à partir d'axiomes
  - le résultat est sûr, mais la méthode nécessite la connaissance de règles

- ***Induction : base du data mining***
  - méthode permettant de tirer des conclusions à partir d'une série de faits
  - généralisation « un peu abusive »
  - indicateurs de confiance permettant la pondération

5. Data mining

Des algorithmes d'inspirations …

- Mathématiques : statistiques et Analyse de Données
- **Calculatoires**
  - « *Clustering* »
  - Arbres de décision
  - Règles d'association
  - Programmation dynamique
  - Machines à Vecteurs de Support (SVM)
- **Biologiques**
  - Réseaux de neurones
  - Algorithmes génétiques

5. Data mining

Des algorithmes :

- **Non Supervisés - Apprentissage *a priori en mode Découverte***
  - « *Clustering* »
  - Algorithmes génétiques
  - Règles d'association
- **Supervisés - Apprentissage *a posteriori* en mode Reconnaissance/Prédiction**
  - Réseaux de neurones / Machines à Vecteurs de Support
  - Arbres de décision
  - Programmation dynamique

5. Data mining

- Classification
- Estimation
- Recherche d'associations
- Clustering

Data mining : Classification

- Affecter un objet à une classe en fonction de ses caractéristiques A1,…An

- Exemple
  - Déterminer si un message est un mail de SPAM ou non (2 classes)
  - Affecter une page web dans une des catégories thématiques de l'annuaire Yahoo (multi-classes)
  - Diagnostic : risque d'accident cérébral ou non (2 classes)

Classification

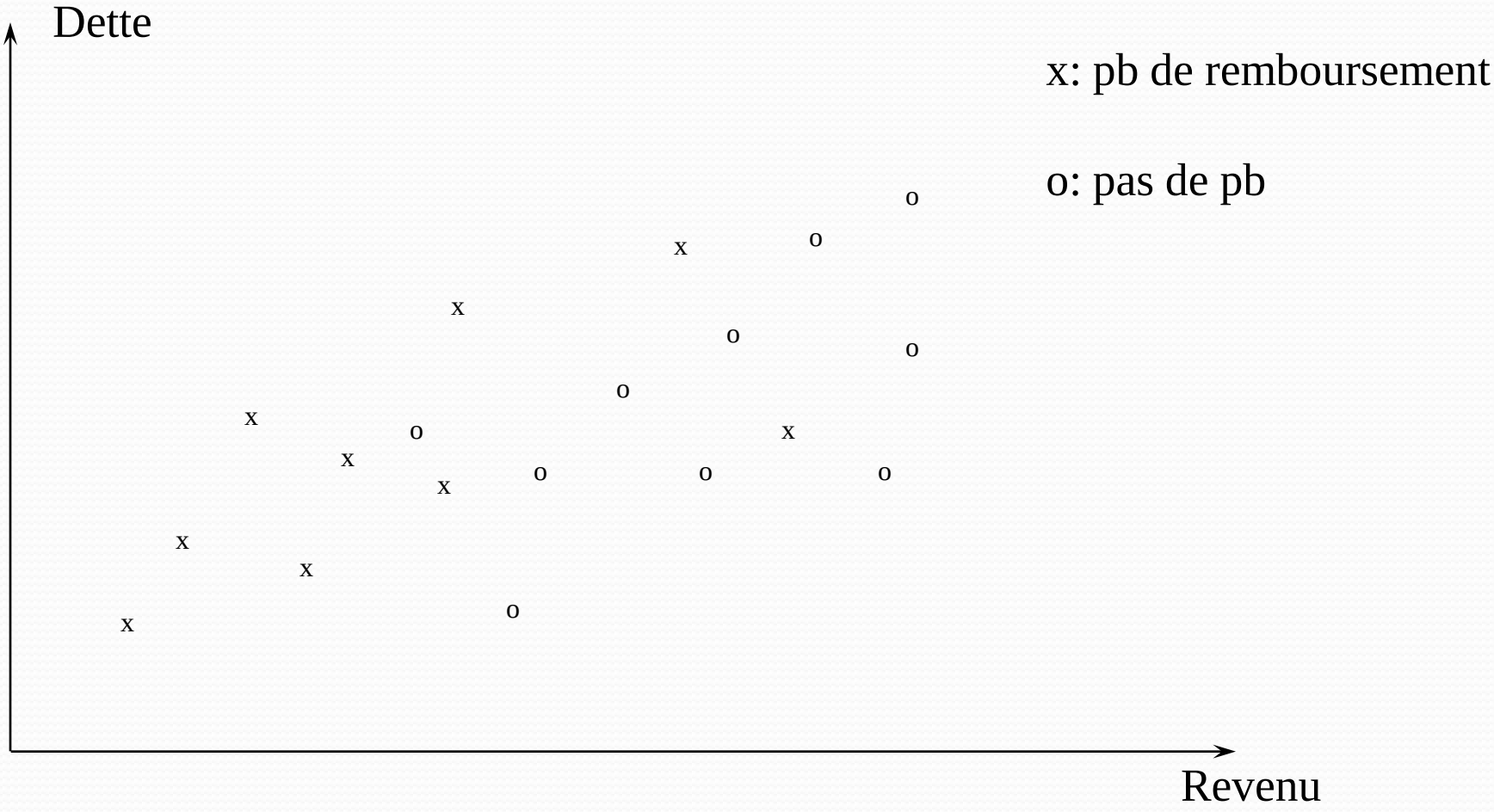
- Si pas de connaissance *a priori* pour définir la classe en fonction de A1, …, An alors on étudie un ensemble d'exemples pour lesquels on connaît A1,…, An et la classe associée et on construit un modèle

$$\text{Classe} = f(A1,\ldots,An)$$

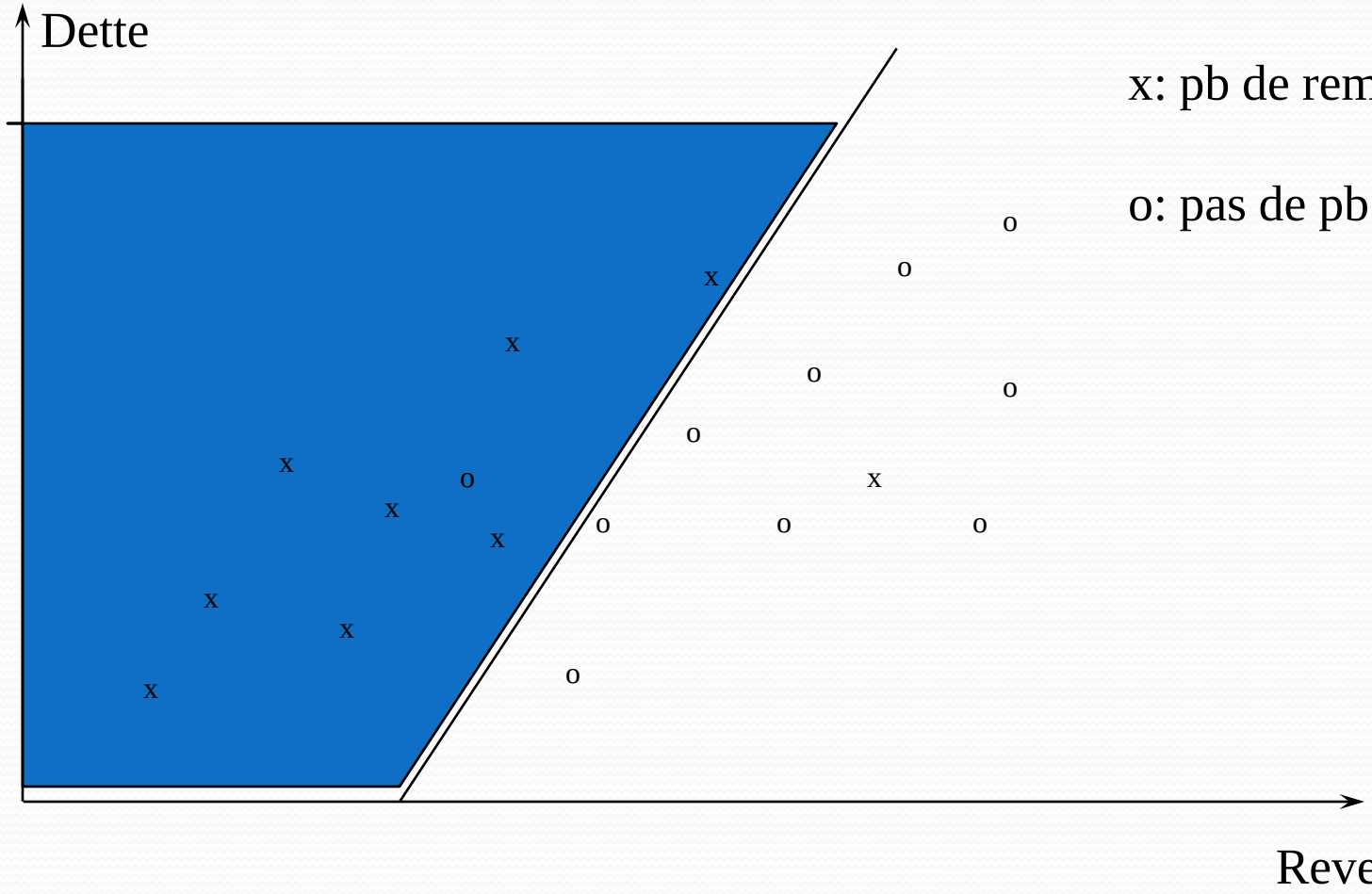
  - Analyse discriminante
  - Arbres de classification
  - Machines à noyaux

Un exemple schématique avec 2 classes

La classification c'est apprendre une fonction qui permet d'affecter un nouvel individu dans une classe ou une autre.



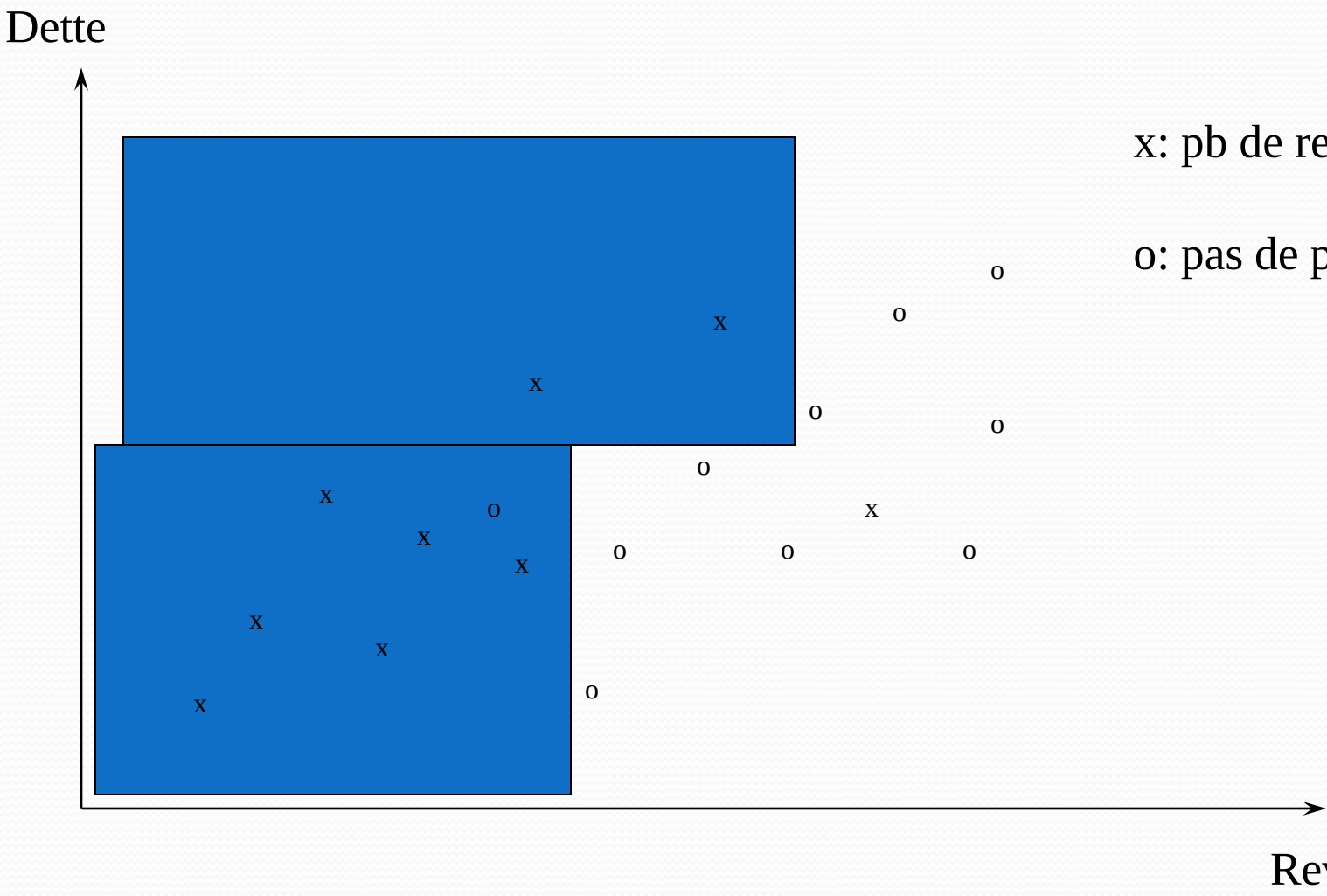
Classification par analyse discriminante



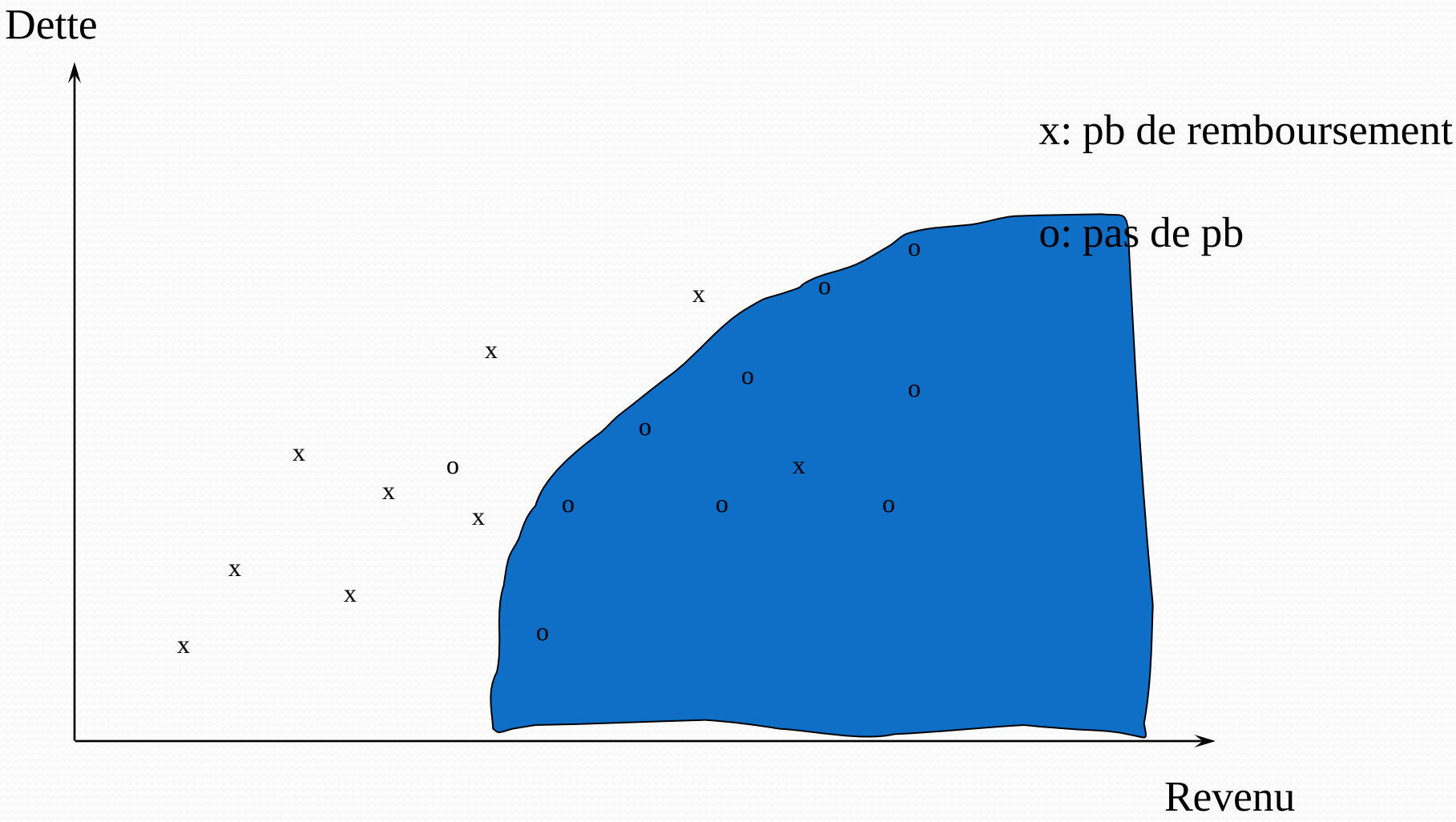
x: pb de remboursement

o: pas de pb

Classification par arbre de décision



Classification par machine à noyaux

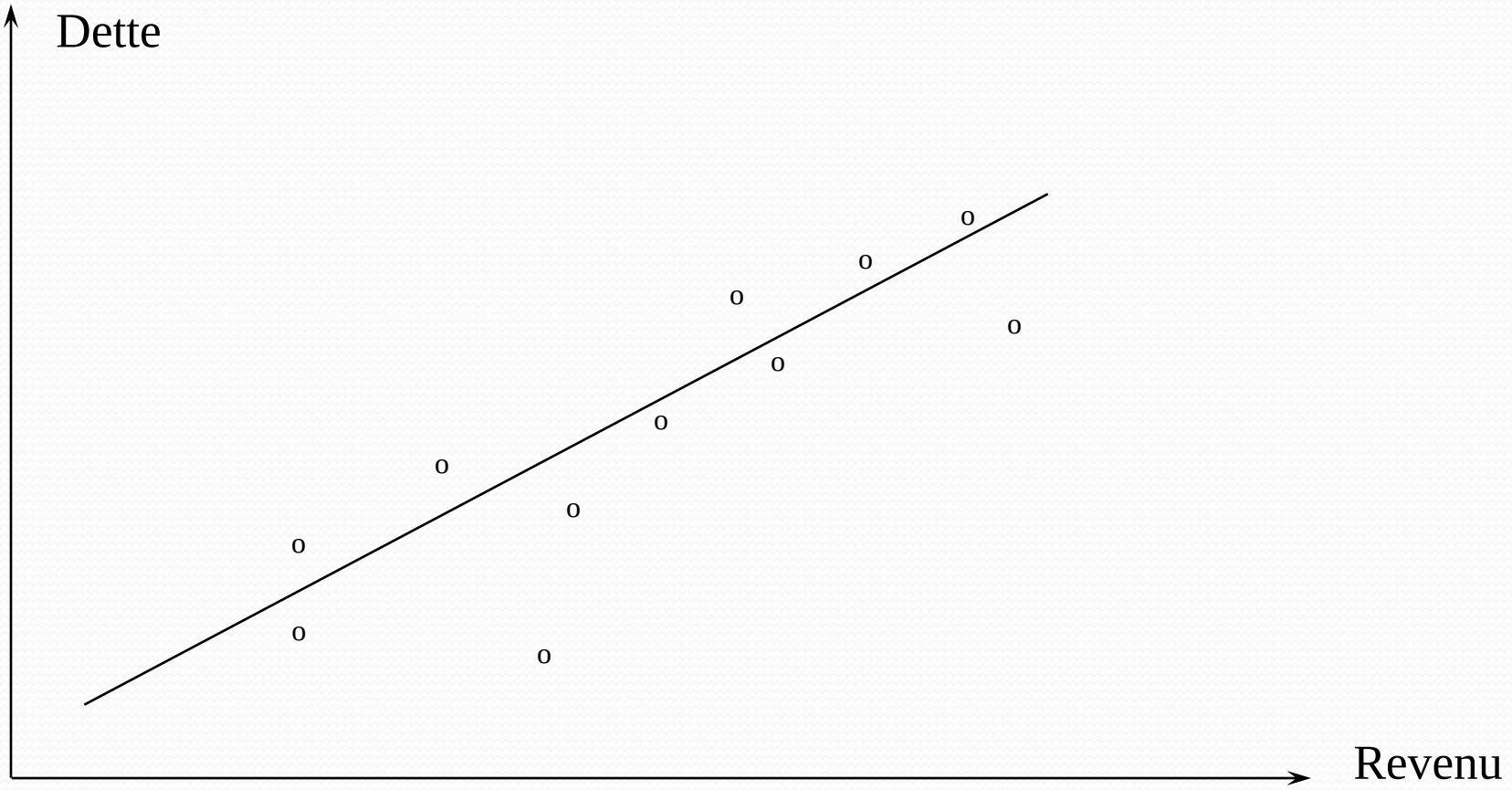


Data mining : Estimation

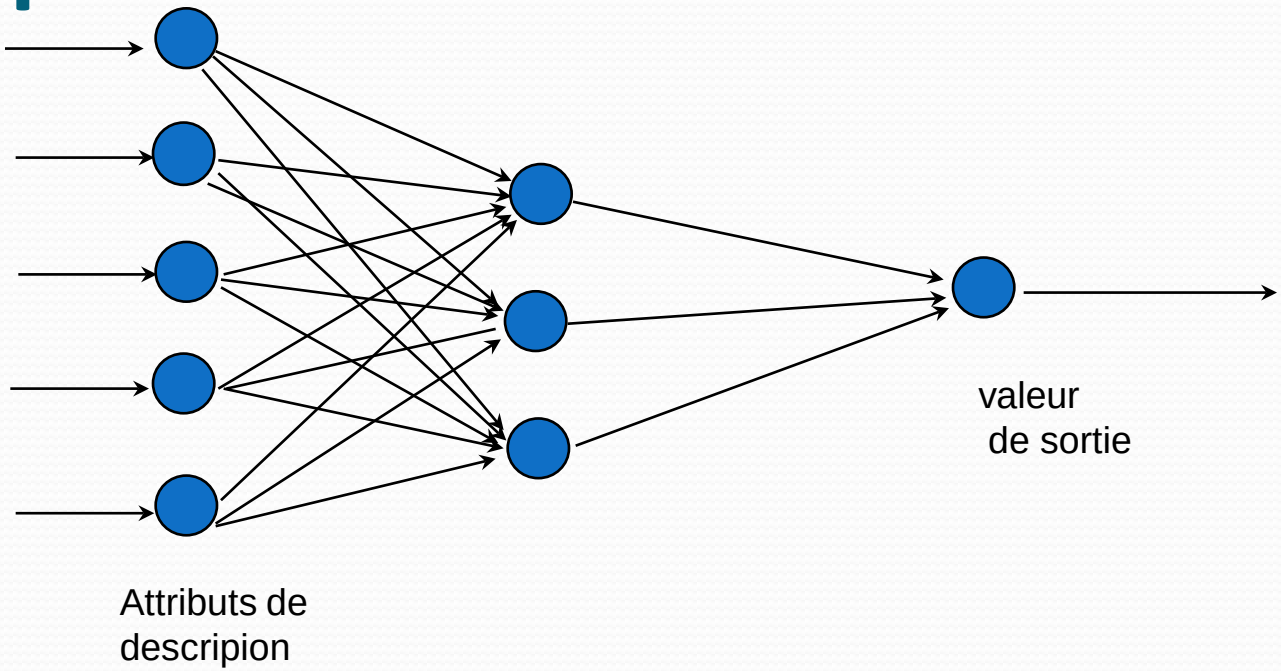
- Estimer (prédire) la valeur d'une variable à valeurs continues à partir des valeurs d'autres attributs
  - Régression
  - Machines à noyaux

Régression linéaire simple

La régression explique les variations d'une variable par une fonction des autres variables : ici la dette est représentée comme une fonction du revenu, le résultat est médiocre car il y a peu de corrélation.



Machines à noyaux : Réseau de neurones & SVM



- La couche d'entrées correspond aux entrées, la couche de sortie(s) au résultat
- Système non-linéaire
- L'apprentissage va ajuster les poids des connexions mais l'architecture et le nombre de neurones dans la couche cachée est un choix arbitraire.

Data mining : Recherche de règles

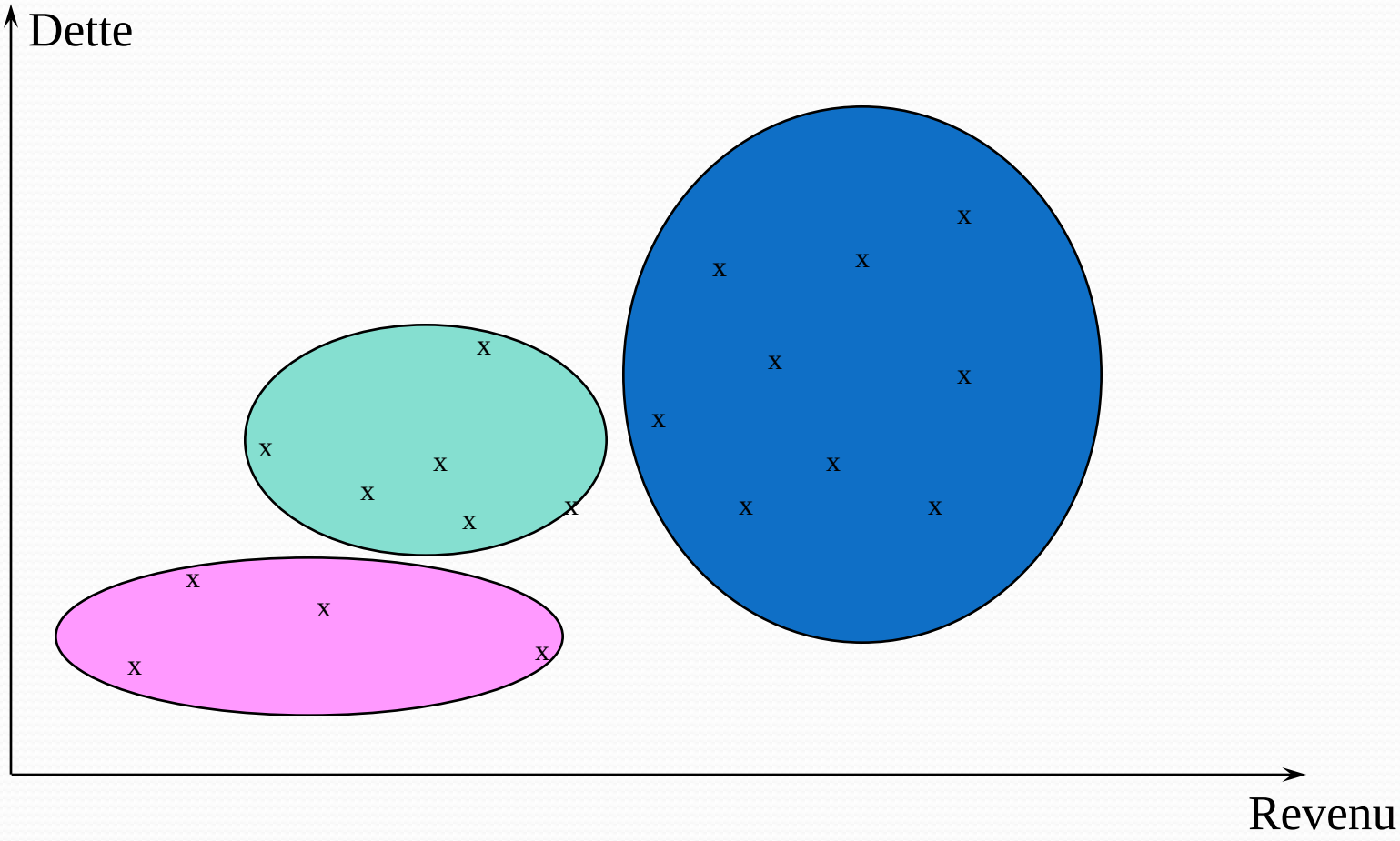
- Règles d'associations : analyse du panier de la ménagère
  - « le jeudi, les clients achètent souvent en même temps des packs de bière et des couches. »
  - Y-a-t-il des liens de causalité entre l'achat d'un produit P et d'un autre produit P' ?

Data mining : Segmentation / partitionnement (clustering)

- Apprentissage non supervisé : les données ne sont pas classées, on isole des sous-groupes d'enregistrements similaires les uns aux autres (nuées dynamiques ou agrégation)
- Un fois les clusters détectés, on pourra appliquer des techniques de modélisation sur chaque cluster

Clustering

Pas d'affectation à une classe connue au départ : on regroupe les individus par leur proximité en classes qui peuvent se recouper.

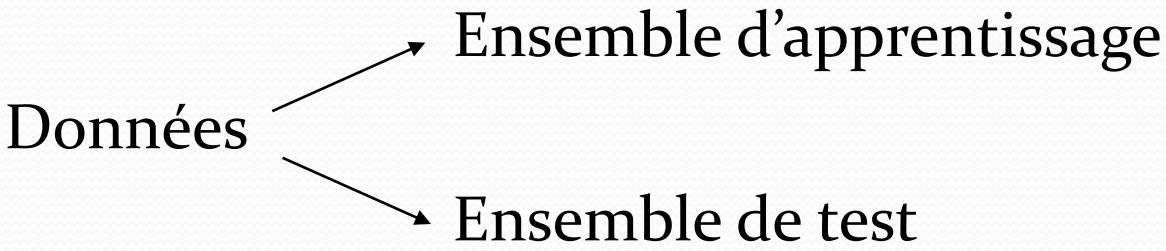


Supervisé vs NON supervisé

- La classification, la régression sont des tâches supervisées
  - Data mining prédictif (on dispose d'une variable dépendante à prédire ou à estimer notée généralement par Y).
- Le clustering, la recherche de règles d'associations sont des tâches non supervisées
  - Data mining explicatif (on cherche plus à expliquer les relations entre les variables sans disposer d'une variable dépendante).

6. Validation dans le cas supervisé

- Validation par le test



Construction d'un modèle sur l'ensemble d'apprentissage et test du modèle sur le jeu de test pour lequel les résultats sont connus

- Évaluation quantitative (ne pas oublier les intervalles de confiance)

7. Intégration de la connaissance

- Prise de décision grâce aux connaissances extraites

- **Les experts métiers sont essentiels pour donner du sens aux informations extraites !**

Cinq Mythes sur le Data Mining

1. Le data mining pourrait instantanément prévenir l'avenir, à la manière d'une boule de cristal.
2. Le data mining ne serait pas encore viable pour des applications professionnelles.
3. Le data mining exigerait une base de données distincte et dédiée.
4. Il faudrait être polytechnicien pour faire du data mining.
5. Le data mining serait réservé aux grandes entreprises disposant d'un large volume de données client.

Références

- Data Mining : techniques appliquées au marketing, à la vente et aux services clients, Berry & Linoff. InterEditions (1997).

- Data Mining et statistique décisionnelle, Stéphane Tufféry (2005)

Sur le Web…

- http://www.lsp.ups-tlse.fr/Besse/enseignement.html
- http://www.web-datamining.net/
- http://www.kdnuggets.com
- http://www.data-miners.com
- http://www.cs.bham.ac.uk/~anp/TheDataMine.html

Regroupement (Clustering) et similitude

Objectifs

- Comprendre la notion d'apprentissage **NON supervisé**
- Le lier à la notion de **Découverte de Structures**
- Connaître des algorithmes de regroupement
  - Hiérarchiques et leur représentation graphique : **dendrogramme**
  - par Optimisation type **K-Means, ISODATA**
- Comprendre que la notion de Similitude liée à la vaste notion mathématique de Distance est **subjective mais centrale** dans cette problématique
- Savoir **construire un espace** de mesure multi-dimensionnelle et définir **une mesure de similarité** dans cette espace
- Savoir **choisir l'algorithme** à utiliser en fonction des données en entrée

Clustering

- Principes
  - Contexte non supervisé
  - « Révéler » l'organisation de motifs en groupes cohérents
  - Processus Subjectif

- Disciplines : Biologie, Zoologie, Psychiatrie, Sociologie, Géologie, Géographie…

- Synonymes : Apprentissage non supervisé, Taxonomie, Typologie, Partition

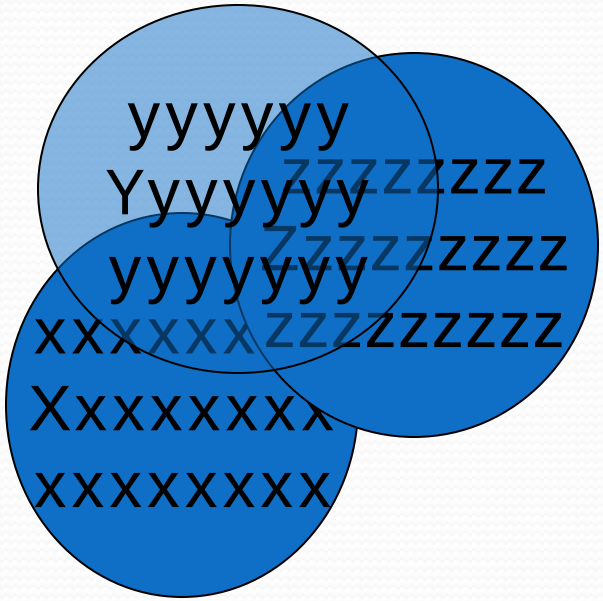
Qu’est-ce que le clustering ?

- Groupe ou “Cluster”: un ensemble d’objets ou d’individus
  - Semblables entre eux à l’intérieur d’un groupe
  - Différents d’un groupe à l’autre
- Segmentation ou “Cluster analysis”
  - Classement des individus ou objets dans différents groupes ou segments
- Le clustering est une technique **non dirigée**

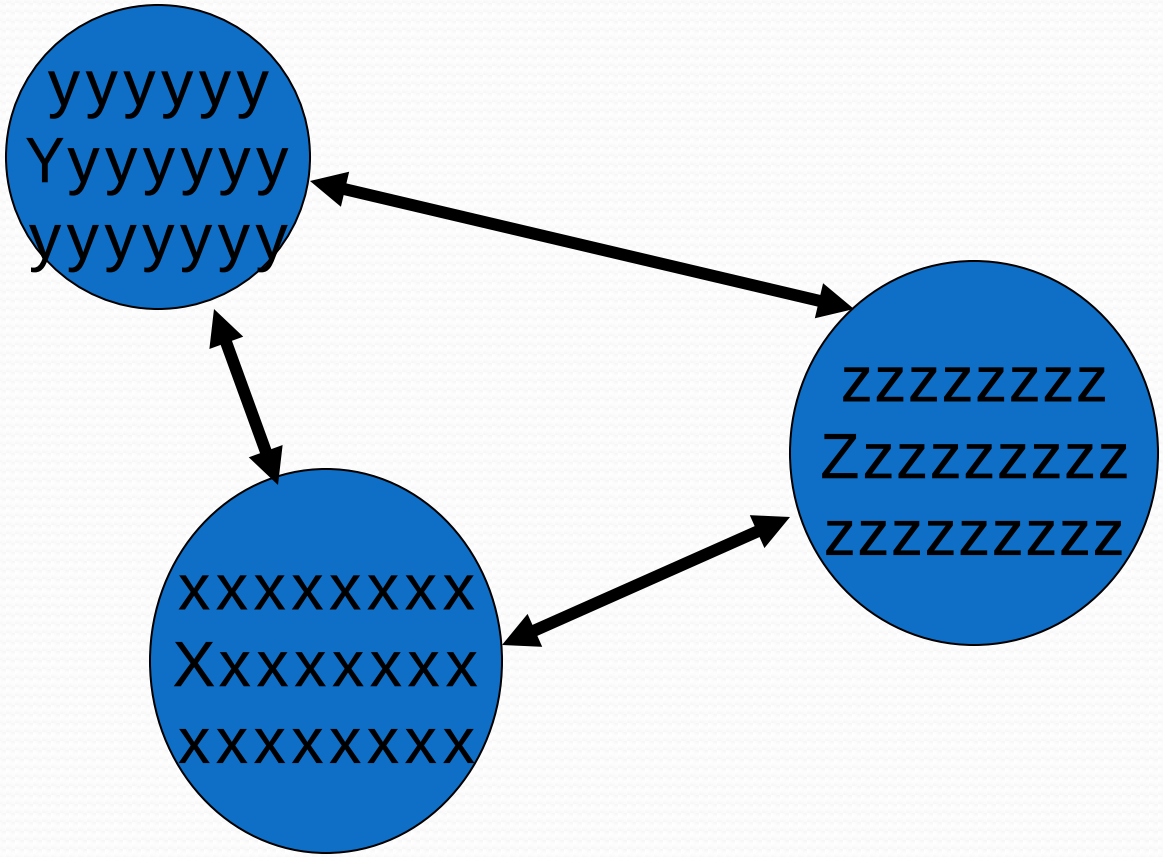
Comment déterminer une bonne segmentation ?

- Un bon algorithme de classification fera en sorte qu'il y aura une :
  - petite variabilité intra-classe (c-à-d petite distance entre les individus d'un même groupe)
  - grande variabilité inter-classe (c-à-d grande distance entre les individus de groupes différents)
- La qualité des résultats de la classification dépendra de la mesure de distance utilisée et de l'algorithme choisi pour l'implanter.

Comment déterminer une bonne segmentation ?



Comment déterminer une bonne segmentation ?



Algorithmes de Clustering

Méthodes de clustering : caractéristiques

- Extensibilité
- Abilité à traiter différents types de données
- Découverte de clusters de différents formes
- Connaissances requises (paramètres de l'algorithme)
- Abilité à traiter les données bruitées et isolées.

Approches de Clustering

- Algorithmes de Partitionnement: Construire plusieurs partitions puis les évaluer selon certains critères

- Algorithmes hiérarchiques: Créer une décomposition hiérarchique des objets selon certains critères

- Algorithmes basés sur la densité: basés sur des notions de connectivité et de densité

Algorithmes à partionnement

- Construire une partition à **k** clusters d'une base **D** de **n** objets
- Les k clusters doivent optimiser le critère choisi
  - Global optimal: Considérer toutes les k-partitions
  - Heuristic methods: Algorithmes k-means et k-medoids
  - k-means (MacQueen'67) : Chaque cluster est représenté par son centre
  - k-medoids or PAM (Partition Around Medoids) (Kaufman & Rousseeuw'87) : Chaque cluster est représenté par un de ses objets

La méthode des k-moyennes (K-Means)

- L'algorithme k-means est en 4 étapes :

1. Choisir k objets formant ainsi k clusters
2. (Ré)affecter chaque objet O au cluster C_i de centre M_i tel que dist(O,M_i) est minimal
3. Recalculer M_i de chaque cluster (le barycentre)
4. Aller à l'étape 2 si on vient de faire une affectation

Illustration (1)

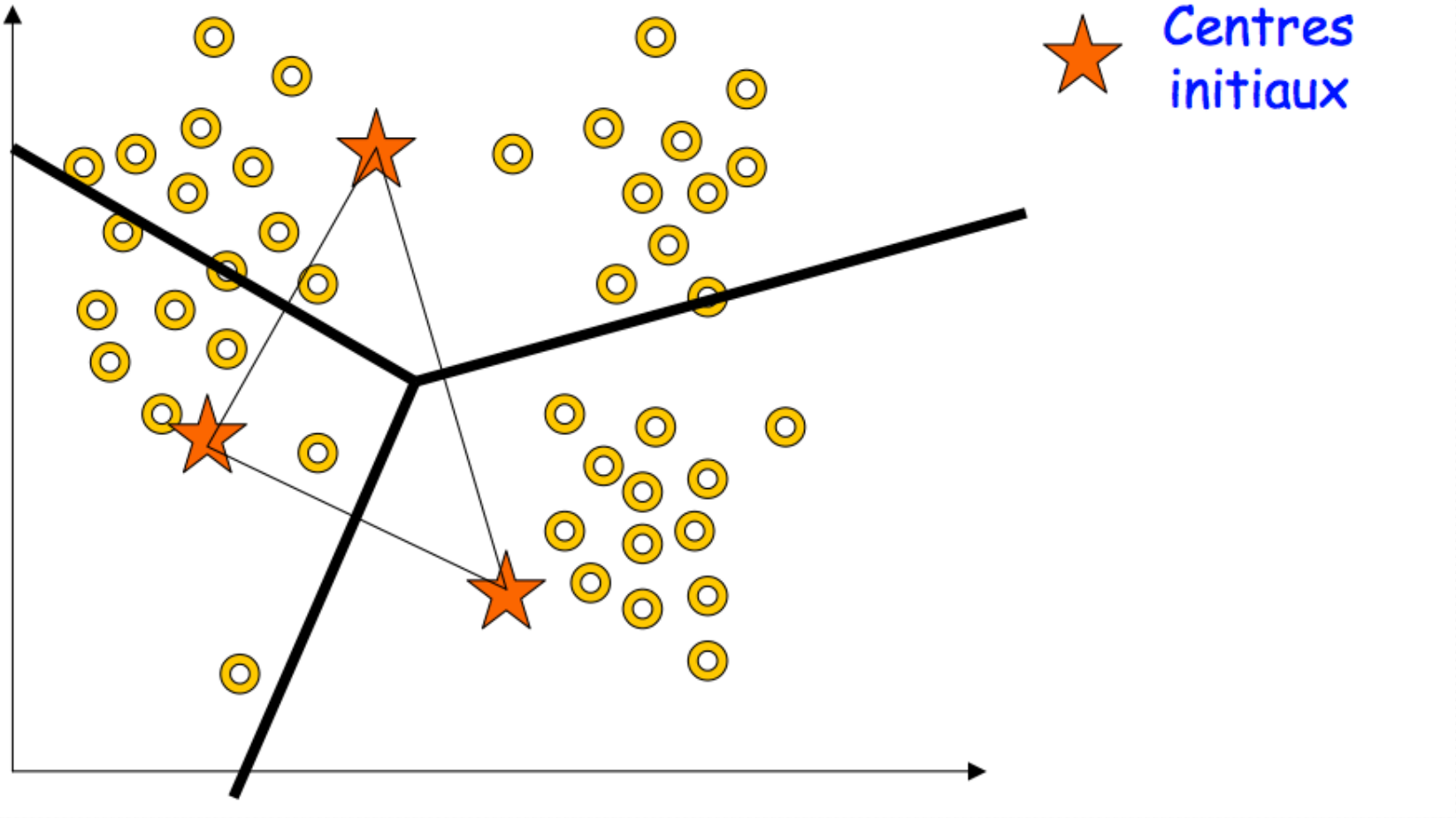


Illustration (2)

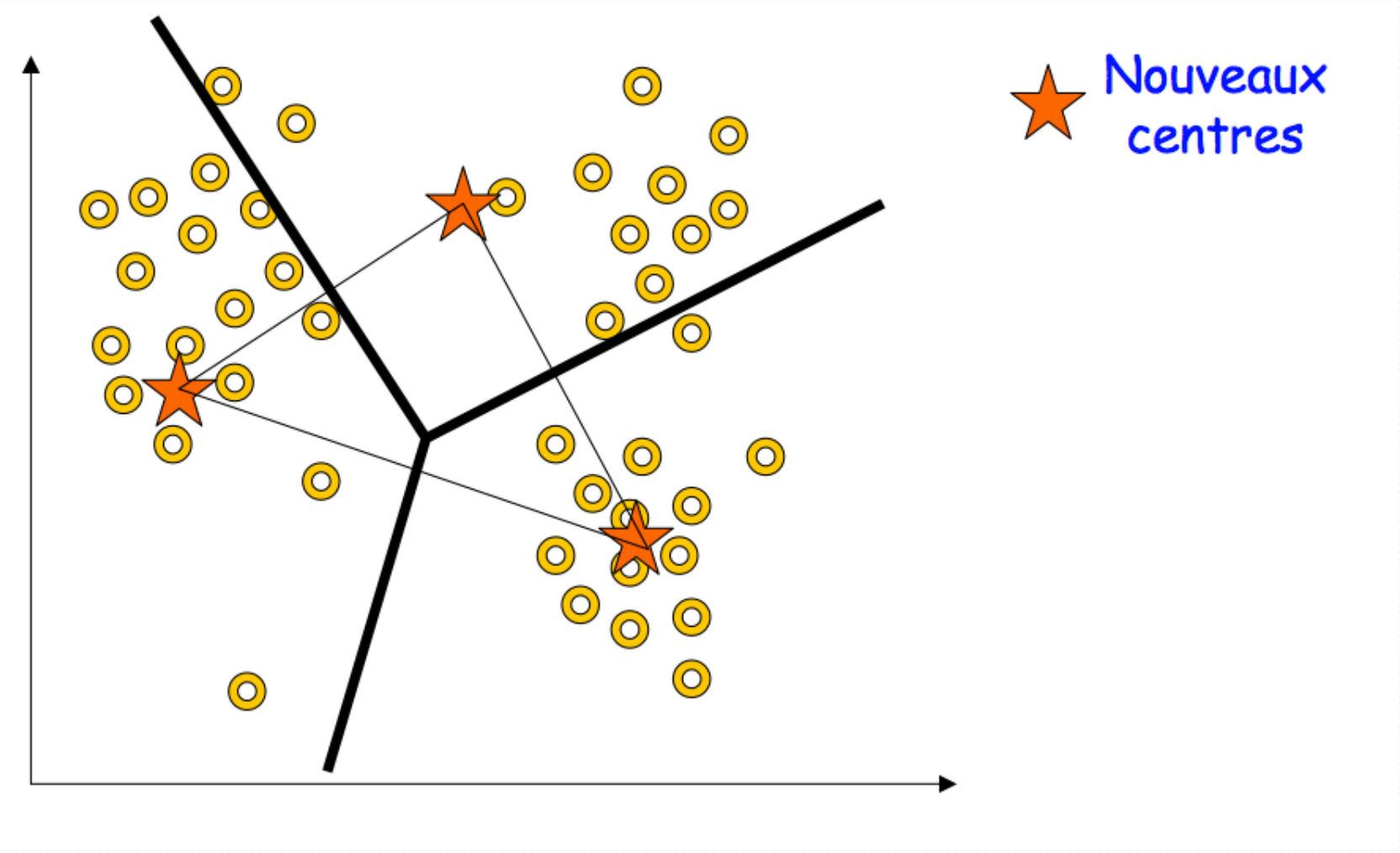
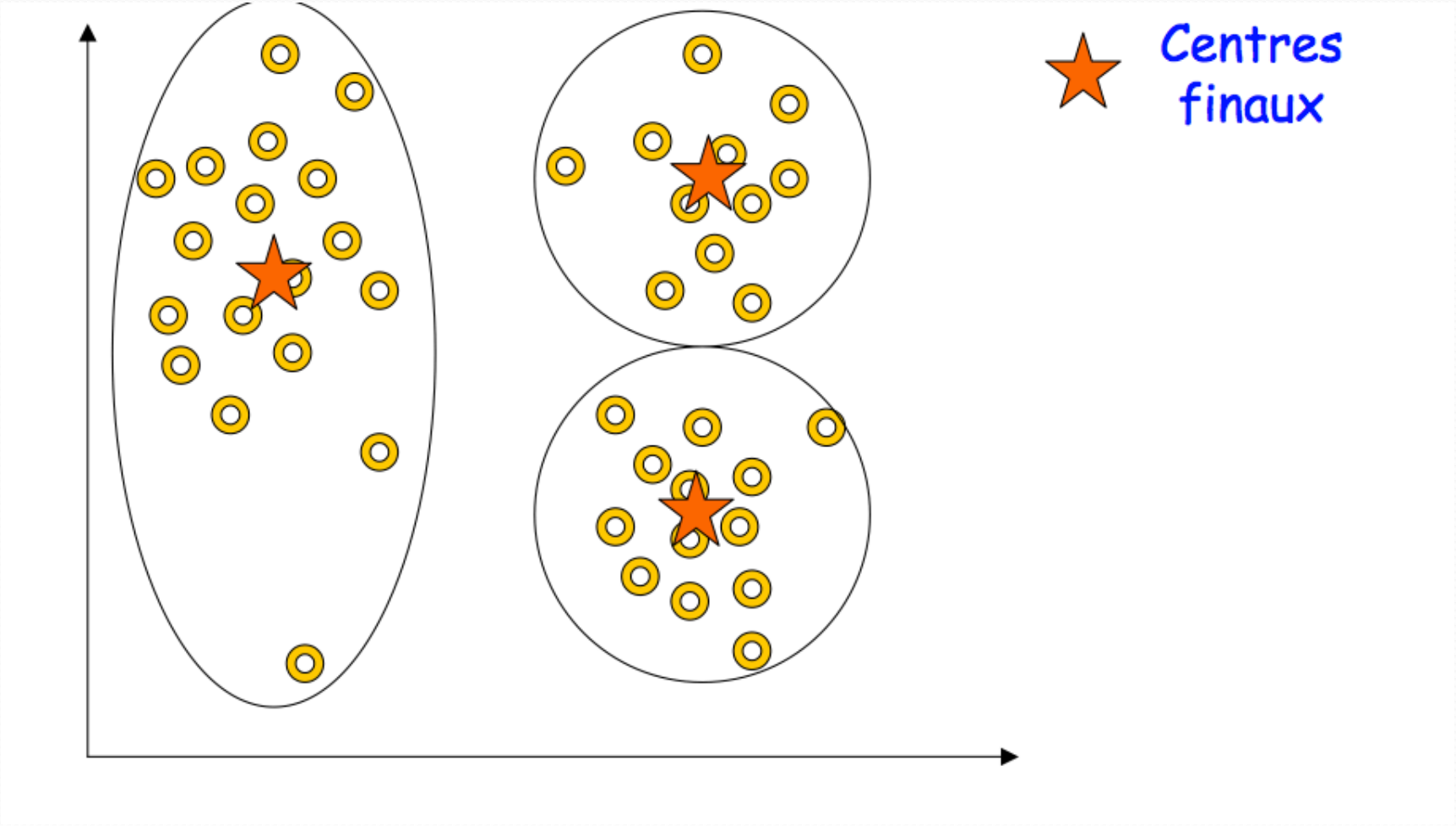


Illustration (3)



Commentaires sur la méthode des K-Means

- Force
  - Relativement extensible dans le traitement d'ensembles de taille importante
  - Relativement efficace : O(t.k.n), où n représente # objets, k # clusters, et t # iterations. Normalement, k, t << n.
- Faiblesses
  - N'est pas applicable en présence d'attributs où la moyenne n'est pas définie
  - On doit spécifier k (nombre de clusters)
  - Incapable de traiter des données bruitées
  - Les clusters sont construits par rapports à des objets inexistants (les milieux)
  - Ne peut pas découvrir les groupes non-convexes
  - Les outliers sont mal gérés.

Variantes des K-means

- Sélection des centres initiaux
- Calcul des similarités
- Calcul des centres (K-medoids : [Kaufman & Rousseeuw'87])
- GMM : Variantes de K-moyennes basées sur les probabilités
- K-modes : données catégorielles [Huang'98]
- K-prototype : données mixtes (numériques et catégorielles)

Approches de Clustering

- Algorithmes de Partitionnement: Construire plusieurs partitions puis les évaluer selon certains critères

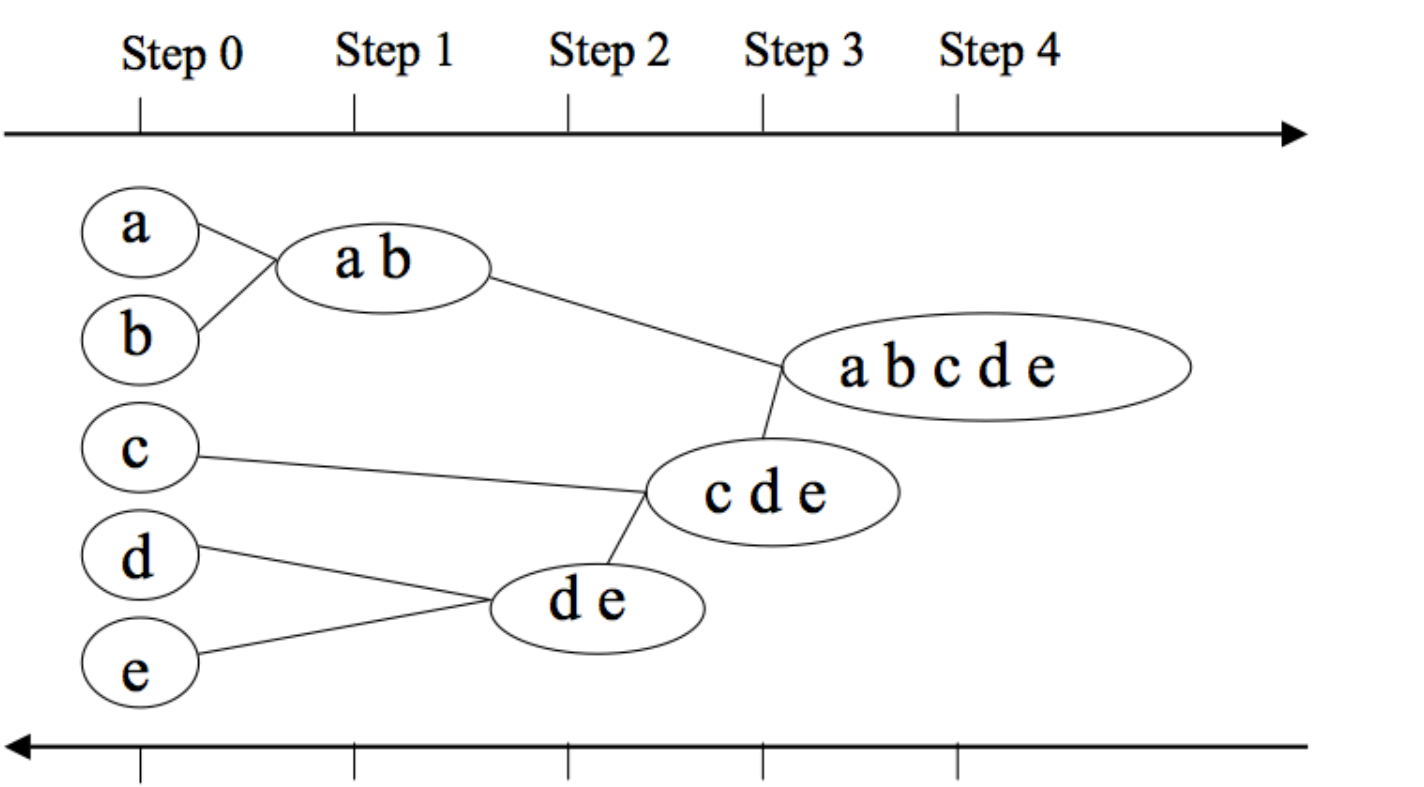
- Algorithmes hiérarchiques: Créer une décomposition hiérarchique des objets selon certains critères

- Algorithmes basés sur la densité: basés sur des notions de connectivité et de densité

Méthodes hiérarchiques

- **Une méthode hiérarchique :** construit une hiérarchie de clusters, non seulement une partition unique des objets.
- Le nombre de clusters **k** n'est pas exigé comme donnée
- Utilise une matrice de distances comme critère de clustering
- Une **condition de terminaison** peut être utilisée (ex. Nombre de clusters)

Arbre de clusters



Dendogramme

- Résultat : Graphe hiérarchique qui peut être coupé à un niveau de **dissimilarité** pour former une **partition**.
- La hiérarchie de clusters est représentée comme un arbre de clusters, appelé **dendrogramme**
- Les **feuilles** de l'arbre représentent les **objets**
- Les noeuds intermédiaires de l'arbre représentent les **clusters**

Distances entre clusters

- Distance entre les centres des clusters (Centroid Method)

- Distance minimale entre toutes les paires de données des 2 clusters (***Single Link Method***)

$$d(i, j) = \min_{x \in Ci, y \in Cj} \{d(x, y)\}$$

- Distance maximale entre toutes les paires de données des 2 clusters (***Complete Link Method***)

$$d(i, j) = \max_{x \in Ci, y \in Cj} \{d(x, y)\}$$

- Distance moyenne entre toutes la paires d'enregistrements (Average Linkage)

$$d(i, j) = avg_{x \in Ci, y \in Cj} \{d(x, y)\}$$

+ et -

- Avantages :
  - Conceptuellement simple
  - Propriétés théoriques sont bien connues
  - Quand les clusters sont groupés, la décision est définitive => le nombre d'alternatives différentes à examiner est réduit
- Inconvénients :
  - Groupement de clusters est définitif => décisions erronnées sont impossibles à modifier ultérieurement
  - Méthodes non extensibles pour des ensembles de données de grandes tailles

Approches de Clustering

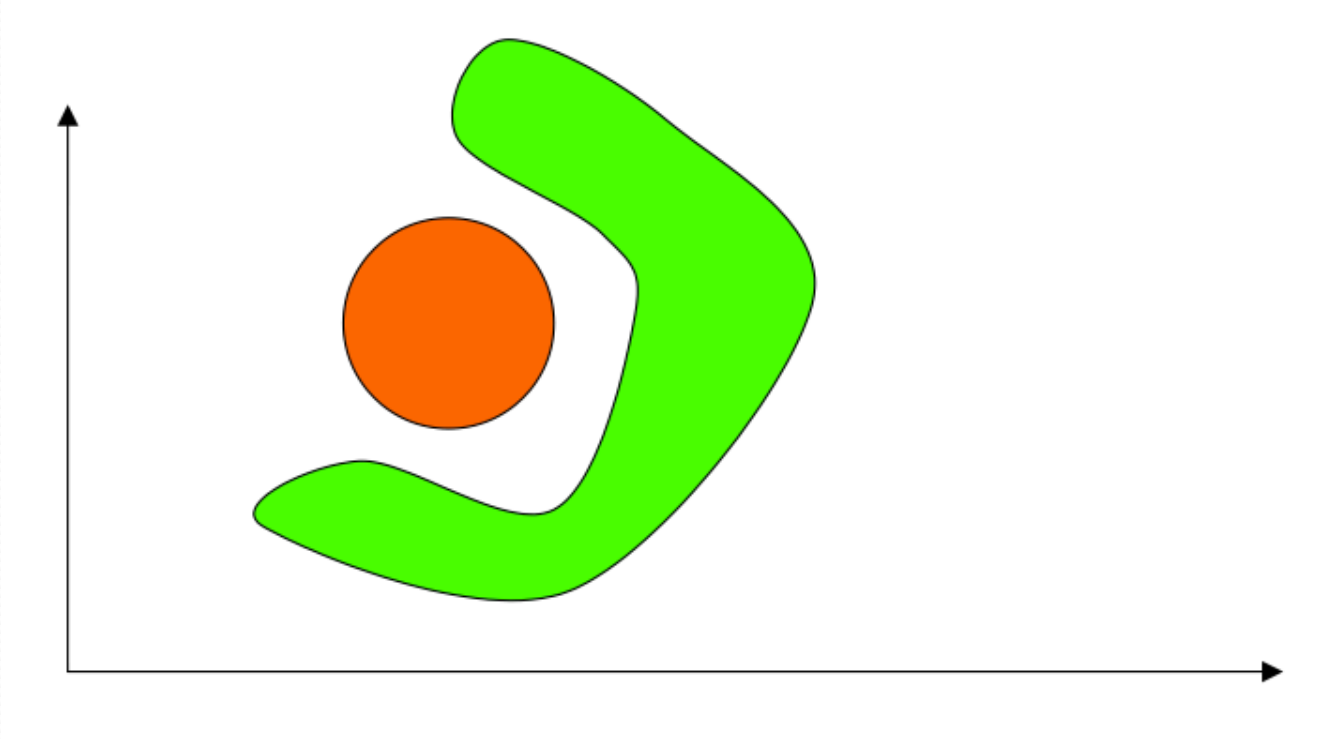
- Algorithmes de Partitionnement: Construire plusieurs partitions puis les évaluer selon certains critères

- Algorithmes hiérarchiques: Créer une décomposition hiérarchique des objets selon certains critères

- Algorithmes basés sur la densité: basés sur des notions de connectivité et de densité

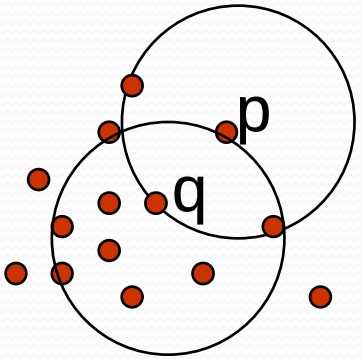
Méthode basée sur la densité

- Pour ce types de problèmes, l'utilisation de mesures de similarité (distance) est moins efficace que l'utilisation de **densité de voisinage.**



Clustering basé sur la densité

- Voit les clusters comme des régions denses séparées par des régions qui le sont moins (bruit)
- Deux paramètres:
  - Eps: Rayon maximum du voisinage
  - MinPts: Nombre minimum de points dans le voisinage-Eps d'un point
- Voisinage : $V_{Eps}(p)$: $\{q \in D \mid dist(p,q) <= Eps\}$
- Un point p est directement densité-accessible à partir de q resp. à Eps, MinPts si
  - $p \in V_{Eps}(q)$
  - $|V_{Eps}(q)| >= MinPts$

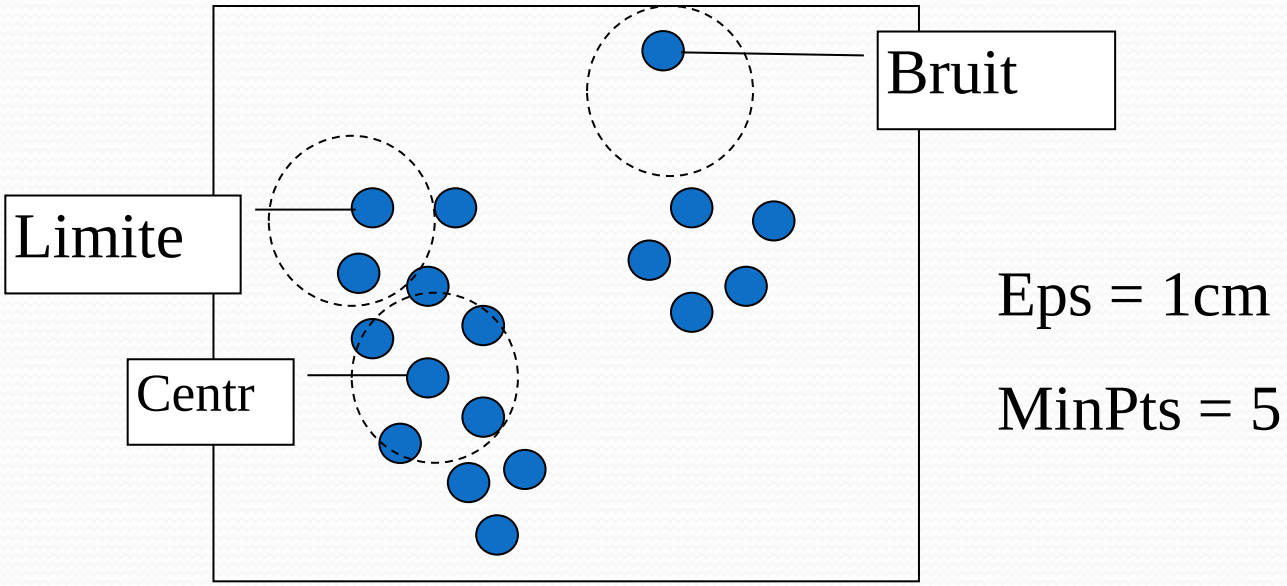


MinPts = 5

Eps = 1 cm

DBSCAN: Density Based Spatial Clustering of Applications with Noise

- Un cluster est l'ensemble maximal de points connectés
- Découvre des clusters non nécessairement convexes



DBSCAN: l'algorithme

- Choisir p
- Récupérer tous les points accessibles à partir de p resp. à Eps et MinPts.
- Si p est un centre, un cluster est formé.
- si p est une limite, alors il n'y a pas de points accessibles de p : passer à un autre point
- Répéter le processus jusqu'à épuiser tous les points.

Résumé

- Le clustering groupe des objets en se basant sur leurs **similarités**.
- Le clustering possède plusieurs applications.
- La mesure de similarité peut être calculée pour **différents types** de données.
- La sélection de la **mesure de similarité** dépend des données utilisées et le type de similarité recherchée.

Notions de distance et de similarité

Notions de distance

- Définition :

Soit x et y deux vecteurs, d est une distance si et seulement si $d(x,y)$ vérifie les propriétés suivantes

- $d(x,y) \geq 0$
- $d(x,y) = 0 \Leftrightarrow x=y$
- $d(x,y) = d(y,x)$
- $d(x,y) \leq d(x,z) + d(z,y)$

Mesure de la qualité de la segmentation

- Mesure de Dissemblance/Ressemblance : la ressemblance est exprimée par une fonction de distance $d(x, y)$
- La définition des fonctions de distance diffère selon le type de variables (intervalle, binaire, nominale, ordinale)
- Il est difficile de définir “assez ressemblant” ou “bonne ressemblance” pour inclure deux individus dans le même groupe :
  - Il y a typiquement une grande part de subjectivité dans la décision.

Dissemblance et ressemblance entre objets ou individus à valeurs réelles

- Une fonction de distance est normalement utilisée pour mesurer la ressemblance ou dissemblance entre deux individus.
- Une fonction de distance parmi les plus populaires pour des variables de type intervalle: *Minkowski distance* :

$$d(x,y)=\sqrt[q]{(|x_1-y_1|^q+|x_2-y_2|^q+...+|x_p-y_p|^q)}$$

où $x = (x_1, x_2,..., x_p)^T$ et $y = (y_1, y_2,..., y_p)^T$ sont deux vecteurs de dimension p représentant deux objets et q est un entier positif.

Dissemblance et ressemblance entre objets ou individus à valeurs réelles (suite)

- Si $q = 1$, $d(x,y)$ est la distance de Manhattan:

$$d(x,y) = |x_1 - y_1| + |x_2 - y_2| + \ldots + |x_p - y_p|$$

- Si $q = 2$, $d(x,y)$ est la distance Euclidienne

$$d(x,y) = \sqrt{(|x_1 - y_1|^2 + |x_2 - y_2|^2 + \ldots + |x_p - y_p|^2)}$$

- Si $q = \infty$, $d(x,y)$ n'est plus une distance mais correspond à une mesure très utilie

$$d(x,y) = \max(x,y)$$

Dissemblance et ressemblance entre objets ou individus à valeurs réelles (suite)

- Nous pouvons également utiliser une fonction de distance pondérée :

$$d(x,y)=\sqrt[q]{(w_1|x_1-y_1|^q+w_2|x_2-y_2|^q+...+w_p|x_p-y_p|^q)}$$

Par exemple, la distance de Mahalanobis d'une série de valeurs de moyenne $\mu=\left(\mu_1,\mu_2,\mu_3,...,\mu_p\right)^T$ et possédant une matrice de covariance Σ pour un vecteur à plusieurs variables $x=\left(x_1,x_2,x_3,...,x_p\right)^T$ est :

$$d(x)=\sqrt{(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

Dissemblance et ressemblance entre objets ou individus à valeurs réelles (suite)

- La distance de Mahalanobis peut aussi être définie comme étant la mesure de dissimilarité entre deux vecteurs aléatoires et de **même distribution** avec une matrice de covariance Σ :

$$d(x,y)=\sqrt{(x-y)^T \Sigma^{-1}(x-y)}$$

- Si la matrice de covariance est la matrice identitaire, cette distance est alors la même que la distance euclidienne.
- Si la matrice de covariance est diagonale, elle est appelée distance euclidienne normalisée :

$$d(x,y)=\sqrt{\sum_{i=1}^{p}\frac{(x_i-y_i)^2}{\sigma_i^2}}$$

- A la différence de la distance euclidienne où toutes les composantes des vecteurs sont traitées de la même façon, la distance de Mahalanobis accorde un poids moins important aux composantes les plus bruitées (en supposant que chaque composante soit une variable aléatoire de type gaussien).

Dissemblance et ressemblance entre objets ou individus à valeurs réelles (suite)

- D'autres similarités et distances
  - Similarité de corrélation : $S_{inner}(x, y) = x^T y$
  - Similarité cosinus :

$$S_{cosinus}(x, y) = \arccos \frac{x^T y}{\|x\| \cdot \|y\|}$$

comme l'angle $S_{cosinus}$ est comprise dans l'intervalle $[0,\pi]$, la valeur π indiquera des vecteurs résolument opposés, $\pi/2$ des vecteurs indépendants (orthogonaux) et 0 des vecteurs colinéaires. Les valeurs intermédiaires permettent d'évaluer le degré de similarité (*cette métrique est fréquemment utilisée en fouille de textes*).

Exemple de problème de normalisation des données : cas des intervalles

Il faut standardiser les données en calculant une mesure normalisée par la moyenne et l'écart type du feature *f* : ce qu'on appelle le z-score.

Soit la matrice de n données suivantes :

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

On définit alors la matrice standardisée des z-scores.

Intervalle (discrètes)

- On définit alors la matrice standardisée des z-scores par :
  - Calculer l'écart absolu moyen:

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|)$$

où

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \ldots + x_{nf}).$$

  - Calculer la mesure standardisée (*z-score*)

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

Exemple: distance de Manhattan

	Age	Salaire
Personne1	50	11000
Personne2	70	11100
Personne3	60	11122
Personne4	60	11074

→ d(p1,p2)=120

d(p1,p3)=132

Conclusion: p1 ressemble plus à p2 qu'à p3!!! ☹

z-scoring

	Age	Salaire
Personne1	-2	-0,5
Personne2	2	0,18
Personne3	0	0,32
Personne4	0	0

→ d(p1,p2)=4,675

d(p1,p3)=2,324

Conclusion: p1 ressemble plus à p3 qu'à p2 !!! ☺

$m_{age}=60$, $s_{age}=5$

$m_{salaire}=11074$, $s_{salaire}=48$

Dissemblance et ressemblance entre objets ou individus à valeurs discrètes

- Les coordonnées des vecteurs appartiennent à un ensemble fini $F = \{0, 1,\ldots, k\text{-}1\}$, $k \geq 0$.
- *Si* $x, y \in F^p$, on définit la matrice de contingence $A(x,y)_{kxk} = [a_{ij}]$ par a_{ij} = le nombre de places où le premier vecteur a le symbole i et l'élément correspondant du second vecteur a le symbole j.

$$d_{\text{Hamming}}(x, y) = \sum_{i=0}^{k-1} \sum_{j=0,\, j \neq i}^{k-1} a_{ij}$$

Variables binaires

- Une table de contingence pour données binaires

		y		
		1	0	*sum*
x	1	a	b	$a+b$
	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p

a = nombre de positions où x à *1* et y à *1*

- Exemple $x=(1,1,0,1,0)$ et $y=(1,0,0,0,1)$

  $a=1$, $b=2$, $c=1$, $d=2$

Mesures de distances

- Coefficient d'appariement (matching) simple (invariant pour variables symétriques): $$S_{\text{appariement}}(x, y) = \frac{b + c}{a + b + c + d}$$
- Exemple *x=(1,1,0,1,0) et y=(1,0,0,0,1)*

  $S_{appariement}(x, y)=3/5$
- Coefficient de Jaccard : $$S_{\text{Jaccard}}(x, y) = \frac{b + c}{a + b + c}$$

  $S_{Jaccard}(x, y)=3/4$

Distances et similarités

- L'indice de Tanimoto reprend l'idée de la similarité cosinus dans le cas des attributs discrets binaires :

$$S_{\text{Tanimoto}}(x, y) = \frac{x^T y}{\|x\|^2 + \|y\|^2 - x^T y}$$

- La Earth Mover's Distance (EMD, inventée par Rubin et al. en 1998!!) a été définie comme "*the* minimal cost that must be paid to transform one histogram (P) into the other (Q)":

$$EMD^D(P,Q) = (\min_{\{F_{ij}\}} \sum_{i,j} F_{ij} D_{ij}) / (\sum_{i,j} F_{ij}) \quad s.t. \quad F_{ij} \geq 0 \qquad (1)$$

$$\sum_j F_{ij} \leq P_i, \quad \sum_i F_{ij} \leq Q_j, \quad \sum_{i,j} F_{ij} = min(\sum_i P_i, \sum_j Q_j) \qquad (2)$$

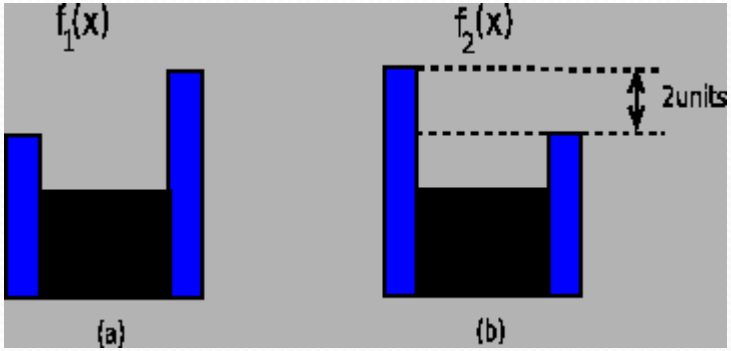
where $\{F_{ij}\}$ denotes the flows. Each F_{ij} represents the amount transported from the i^{th} supply to the j^{th} demand.

Variables binaires (I)

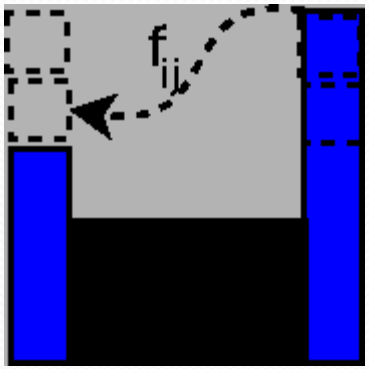
- Variable symétrique: Ex. le sexe d'une personne, i.e coder masculin par 1 et féminin par 0 c'est pareil que le codage inverse

- Variable asymétrique: Ex. Test HIV. Le test peut être positif ou négatif (0 ou 1) mais il y a une valeur qui sera plus présente que l'autre. Généralement, on code par 1 la modalité la moins fréquente
  - 2 personnes ayant la valeur 1 pour le test sont plus similaires que 2 personnes ayant 0 pour le test

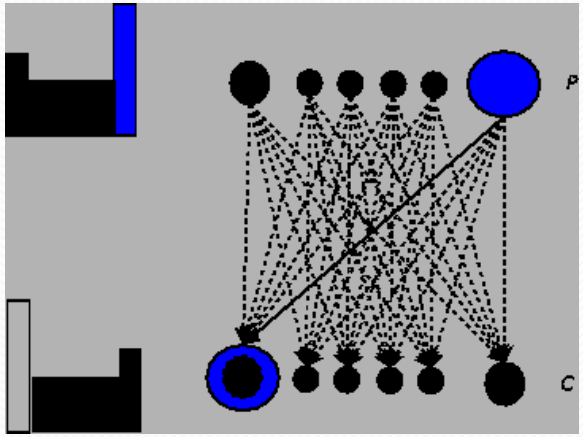
Distances et similarités



- *The Earth Mover's Distance is the minimum amount of work needed to transform one distribution to another one. We assume that both distributions have the same mass, in our case:*



Distances et similarités



In 2008, Pele andWerman proposed a modified version of EMD extending standard EMD not normalized histograms.

- In many histogram comparison situations, the difference between large bins is less important than the difference between small bins and should then be reduced. The **Chi-Squared distance** is a histogram distance that takes this into account:

$$\chi^2(P,Q) = \frac{1}{2}\sum_i \frac{(P_i - Q_i)^2}{(P_i + Q_i)}$$

Variables binaires(II)

- Exemple

Nom	Sexe	Fièvre	Toux	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Sexe est un attribut symétrique
- Les autres attributs sont asymétriques
- Y et P ≡ 1, N ≡ 0, la distance n'est mesurée que sur les asymétriques

$$d(jack, mary) = \frac{0 + 1}{2 + 0 + 1} = 0.33$$

$$d(jack, jim) = \frac{1 + 1}{1 + 1 + 1} = 0.67$$

$$d(jim, mary) = \frac{1 + 2}{1 + 1 + 2} = 0.75$$

Les plus similaires sont Jack et Mary⇒atteints du même mal

Variables Nominales

- Une généralisation des variables binaires, ex: rouge, vert et bleu
- Méthode 1: Matching simple
  - m: # d'appariements, p: # total de variables

$$S(x, y) = \frac{p - m}{p}$$

- Méthode 2: utiliser un grand nombre de variables binaires

  Créer une variable binaire pour chaque modalité (ex: variable rouge qui prend les valeurs vrai ou faux)

Variables Ordinales

- Une variable ordinale peut être discrète ou continue
- L'ordre peut être important, ex: classement
- Peuvent être traitées comme les variables intervalles
  - remplacer x_{if} par son rang $r_{if} \in \{1,\ldots,M_f\}$
  - Remplacer le rang de chaque variable par une valeur dans [0, 1] en remplaçant la variable f dans l'objet I par

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

  - Utiliser une distance pour calculer la similarité

En Présence de Variables de différents Types

- Pour chaque type de variables utiliser une mesure adéquate.
- Problèmes : les clusters obtenus peuvent être différents

- On utilise une formule pondérée pour faire la combinaison

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

  - f est binaire ou nominale :
    $d_{ij}^{(f)} = 0$ si $x_{if} = x_{jf}$, sinon $d_{ij}^{(f)} = 1$
  - f est de type intervalle : utiliser une distance normalisée
  - f est ordinale
    - Calculer les rangs r_{if}
    - Puis traiter z_{if} comme une variable de type intervalle

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$