



4 - Fouille de données François Poulet



Introduction à la fouille de données

François POULET
Université de Rennes 1 - IRISA


5 - Fouille de données François Poulet



Ressources - Bibliographie

- U.Fayyad, G.Piatetsky-Shapiro, P.Smith, R.Uthurusamy : *Advances in Knowledge Discovery and Data Mining*, AAAI Press, 1996
- T.Hastie, R.Tibshirani, J.Friedman : *The Elements of Statistical Learning : Data Mining, Inference and Prediction*, Springer Verlag, 2001
- L.Breiman, J.Friedman, R.Olshen, C.Stone : *Classification and Regression Trees*, Wadsworth, 1984
- D.Hand, H.Mannila, P.Smyth, *Principles of Data Mining*, MIT press, 2000


6 - Fouille de données François Poulet



Ressources - Bibliographie

- www.kdnuggets.com
- www.kdnet.org
- www.mlnet.org
- www.learningtheory.org
- www.kernel-machines.org
- www.acm.org/sigs/sigkdd
- www.classification-society.org (IFCS)
- archives :
 - UCI ML Repository : www.uci.edu/~mllearn/MLRepository.html
 - UCR : www.cs.ucr.edu/~eamonn/TSDMA/main.php
 - Kent Ridge Bio Medical Data Set Repository : sdmc.lit.org.sd/GEDatasets/Datasets.html
- www.infovis.net


7 - Fouille de données François Poulet



Ressources - Bibliographie

- Logiciels (libres) :
 - Weka (Université de Waikato)
 - Tanagra (Université Lyon2)
 - Mineset (SGI -> Purple Insight)
 - R
 - ... www.kdnuggets.com/software
- arbres de décision :
 - C4.5 (Quinlan)
 - OC1 (Murthy et al)
 - CART, Random Forest (Breiman)
 - (Mineset, Weka, Tanagra)


8 - Fouille de données François Poulet



Ressources - Bibliographie

- SVM :
 - libSVM (Chang, Lin)
 - SVMTorch (Collobert et al)
 - SVMLight (T.Joachims)
 - ... www.kernel-machines.org/software.html

9 - Fouille de données François Poulet



Introduction

- introduction fouille de données
- principaux algorithmes de fouille de données
 - classification supervisée
 - arbres de décision
 - SVM Séparateurs à Vaste Marge (Support Vector Machine)
 - Naive-Bayes
 - classification non supervisée
 - k-NN
 - k-means
- le traitement de grandes quantités de données?

7 - Fouille de données François Poulet

Introduction - fouille de données

- quelques chiffres:
 - puissance des machines double 18 mois
 - qté de données double 9 mois
 - + de données dans les 3 ans à venir que jamais
 - 2004 : ~5 exa octets (10^{18}) 1000000 Tera octets

8 - Fouille de données François Poulet

Introduction - fouille de données

- internautes > 2 milliards (1.2 dans pays en voie de développement: + 317% Afrique, 294% M-Orient, 143% Asie)
- 5.3 milliards abonnement mobile
- accessible pour 90% (80% en zone rurale)
- croissance mobile :
 - pays développés : constant (+1.6%)
 - en voie de développement : +18%

9 - Fouille de données François Poulet

Introduction - fouille de données

- en augmentation :
 - blogs, radio, TV... Twitter : 1734 twitts/sec, zettabyte sur IP 2016 (milliard de To)
 - images (Flickr > 5 milliards d'images, 3500 photos /sec), 90% tel. portables font appareil photo => 300 millions de clichés / jour, 100000 photos/utilisateur, nouveau capteur CMOS Sony : 18 Mp
 - vidéos (YouTube > 35h vidéo / mn, Youku, 11 jours visualisés / sec)
 - SMS : 6100 milliards in 2010, 200 000 / sec. (coût 0.07US\$ = 14000\$/sec = 10150€/ sec)
 - pub en ligne : 16.5 milliards \$ Google, 3.5 Yahoo!, 2.9 FaceBook...
 - quantité de données explose : comment y retrouver ce qui nous intéresse?

10 - Fouille de données François Poulet

Introduction - fouille de données

- un exemple : Yahoo! (source U.Fayyad)
 - 73% utilisateurs internet US utilisent Yahoo!
 - 500 millions utilisateurs
 - 25 To de données collectées / jour
 - entrepôt de données : 5000 To
- Texmex : 100 To (6 mois TF1, France 2, France Info...)

11 - Fouille de données François Poulet

Introduction - fouille de données

- quantité de données explose : comment y retrouver ce qui nous intéresse?
- fouille de données : à quoi ça sert ?
- exemple 1 : je cherche une image de jaguar

The screenshot shows a Google Images search results page for the keyword 'jaguar'. The browser's address bar shows the URL: <http://images.google.fr/images?gbv=1&oeq=jaguar&start=20&sa=M&oeq=20>. The search results are displayed in a grid of 12 thumbnails, each with a small caption and source. The captions include: 'Où la Jaguar XJ est sage', 'La Jaguar Advanced Lightweight Coupé', 'Jaguar', 'Jaguar', 'Jaguar Menace en diminution-chasse...', 'Par JAGUAR SOVEREIGN', 'Voici ma Jaguar (enfin exactement)', 'Jaguar XJ6', 'L'avenir de Jaguar repose sur la XF', 'comphotosdof-jaguar/f...', 'Jaguar prévoit 3 versions...', and 'Jaguar d'Amazona'. The sources listed are various automotive websites and forums.

13 - Fouille de données François Poullet

Introduction - fouille de données

- quasiment impossible faire le tri "manuellement" :
- résultats 1 - 20 sur un total d'environ 21 700 000 pour jaguar
- exemple 2 : Yahoo!, publicité contextuelle
- "traquer" les utilisateurs, détecter intentions
- réponse dans l'heure

14 - Fouille de données François Poullet

Introduction - fouille de données

- étude du cheminement utilisateur
- intention d'acheter une voiture
 - 70% ont acheté dans les 3 mois
 - 24% ont acheté dans le mois
- Yahoo! Autos
 - spécifications modèles
 - calculateur emprunt
 - comparateur de véhicules
 - configuration et prix

15 - Fouille de données François Poullet

Introduction - fouille de données

- Yahoo! search
 - fabricants automobiles
 - concessionnaires
 - guides d'achat
- Yahoo! local
 - recherche dans les concessions locales
- gains Yahoo! (2007) > 200 millions \$

16 - Fouille de données François Poullet

Introduction - fouille de données

- mais aussi ... exemple 3 :
- 7 sept. 2008 article Chicago Tribune, repris sur Florida Sun-Sentinel, dépôt de bilan de United Airlines (2^e compagnie mondiale)
 - réalité : article de 2002, suite au 11 sept.2001
- article indexé par robot de Google -> Google news
- newsletter de Bloomberg envoyée à des milliers de personnes
- résultat : -75% action UAL dans la journée

17 - Fouille de données François Poullet

Introduction - fouille de données

- fouille de données
- nécessité de traiter ces données
- MIT Technology Review 2001:
- KDD = 1 des 10 technologies émergentes XXIe

18 - Fouille de données François Poullet

Introduction - fouille de données

- coopération entre :
 - visualisation (d'informations)
 - statistiques
 - analyse de données
 - bases de données
 - IA : apprentissage automatique

19 - Fouille de données François Poullet

Introduction

Method	2007 votes		
Decision Trees/Rules	127 62.6%	+ grandes progressions	
Regression	104 51.2%		AG
Clustering	102 50.2%		Boosting
Statistics	94 46.3%		Visualization
Visualization	66 32.5%	Hybrid methods	
Association rules	53 26.1%	Bagging	
Sequence analysis	35 17.2%		
Neural networks	35 17.2%	déclin relatif	
SVM	32 15.8%	SVM	
BN / Naive Bayes	32 15.8%	Association rules	
Boosting	30 14.8%		
Nearest neighbor	42 19.7%		
Hybrid methods	24 11.8%		
Other	23 11.3%		
Genetic algorithms	23 11.3%		
Bagging	22 10.8%		

20 - Fouille de données François Poullet

Introduction

- CRM / consumer analytics : 38.3%
- Banking : 31.8%
- Fraud detection : 19.6%
- Finance : 16.8%
- Direct Marketing / Fundraising : 14%
- Investment / Stocks : 13.1%
- Credit scoring : 13.1%
- Retail : 12.1%
- Advertising : 12.1%
- Telecom / cable : 12.1%
- Biotech/genomics 11.2%
- Science : 10.3%
- Insurance : 10.3%

21 - Fouille de données François Poullet

Introduction

- définition :
 - [Fayyad, 1996]
 - The non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data.
 - process : ECD processus itératif, réitère étapes ua connaissances intéressantes
 - valid : généralisable dans le futur
 - novel : quelque chose d'inconnu auparavant
 - useful : dont on va pouvoir se servir
 - understandable : pour l'utilisateur final (boîte noire)

22 - Fouille de données François Poullet

Introduction

The diagram illustrates the ECD (Extract, Clean, Discover) process. It starts with 'données brutes' (raw data) represented by a cylinder. An arrow labeled 'sélection' leads to 'données sélectionnées' (selected data), represented by a smaller cylinder. Another arrow labeled 'pré-traitement' leads to 'données pré-traitées' (pre-processed data), represented by a document icon. A final arrow labeled 'modélisation' leads to 'données transformées' (transformed data), represented by a document icon. These four stages are grouped under the heading 'pré-traitement'. An arrow then points to a box labeled 'fouille de données' (data mining), which contains a tree diagram and is labeled 'formes (patterns)'. This box is part of a larger section labeled 'fouille de données'. An arrow from the 'fouille de données' box points to a box labeled 'post-traitement' (post-processing), which contains an icon of a person and is labeled 'interprétation & évaluation'. An arrow from the 'post-traitement' box points to 'connaissances' (knowledge). A feedback loop arrow returns from 'connaissances' to the 'pré-traitement' section.

- processus d'ECD
 - boucle
 - fouille de données : cœur du processus
 - arbres de décision, clustering, association, etc.

23 - Fouille de données François Poullet

Introduction

- processus :
 - pré-traitement = 80% temps
 - data-warehouse
- points critiques :
 - accessibilité données
 - consistance données
 - informations contenues

24 - Fouille de données François Poullet

Introduction

- tâches :
 - classification supervisée : prédire une classe
 - clustering (non supervisée) : trouver des groupes
 - associations : A et B apparaissent souvent
 - visualisation : données / modèle / fouille
 - agrégation : décrire un groupe
 - détection de déviation : concept drift (changement)
 - estimation : prédire une valeur
 - ...

25 - Fouille de données François Poullet

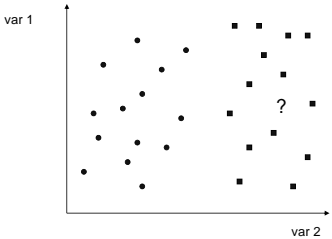
Introduction - Classification

- classification supervisée
- (clustering = non supervisée)
- grand nombre de méthodes
 - arbres de décision : C4.5, CART, OC1, ...
 - réseaux de neurones
 - statistiques / analyse de données
 - SVM (Séparateur à Vaste Marge - Support Vector Machine [Vapnik, 1995])

26 - Fouille de données François Poullet

Introduction - Classification

- but : prédire appartenance à une classe
- individus pré-étiquetés



27 - Fouille de données François Poullet

Introduction - Classification

- arbres de décision :
 - C4.5 [Quinlan, 93]
 - OC1 [Murthy et al, 94]
 - CART [Breiman et al., 84] Random Forest [Breiman et al, 2001]
- SVM
 - www.kernel-machines.org
 - www.support-vector.net

28 - Fouille de données François Poullet

Introduction - Classification

- modes de fonctionnement :
 - ensemble de données initial
 - partage en deux sous-ensembles :
 - apprentissage
 - test
 - f(taille) :
 - petits ensembles : validation croisée (simule ens. + gd)
 - n-fold cross validation : répète n fois
 - divise en n sous-ensembles : training : n-1, test : 1
 - moyenne
 - très petits ensembles : leave one out

29 - Fouille de données François Poullet

Introduction - Classification

- pb très grands ensembles :
 - ex : 1 milliard individus
 - statistique :
 - échantillonnage (sampling)
 - échantillon données représentatif à n% de l'ensemble global
- pb très grands ensembles (suite) :
 - gd nombre de dimensions, exemples :
 - fouille de texte (1 dimension / mot)
 - bio-informatique (bcp dim. et peu ind.)
 - ex : Kent Ridge Bio Medical Data Set Repository
 - Breast Cancer : 24000 dimensions, 100 individus

30 - Fouille de données François Poullet

Introduction - Classification

- critères de qualité des algorithmes :
 - temps d'exécution
 - taux de bonne classification (précision)
 - robustesse
 - scalabilité

31 - Feuille de données François Poulet

Introduction - Précision

- précision
 - taux individus dont label correct
 - mieux : matrice de confusion classifié comme...

l'étiquette est...

	Cat	Dog	Pig
Cat	100	0	0
Dog	9	90	1
Pig	45	45	10

32 - Feuille de données François Poulet

Introduction - Scalabilité

- scalabilité
 - temps construction modèle
 - doit être linéaire en fonction nb ind.
 - temps de test
 - cte
 - nb accès aux données
 - coût mémoire
 - data stream (flot de données)

33 - Feuille de données François Poulet

Algorithmes de classification supervisée

- données ont une étiquette connue a priori
- construire un modèle des données
- étiqueter nouvelles données
- algorithmes :
 - k-PPV (k-NN)
 - arbres de décision
 - NB
 - SVM

34 - Feuille de données François Poulet

Les plus proches voisins

- 1-NN
- attribut a classifier est étiqueté comme son plus proche voisin

35 - Feuille de données François Poulet

Les plus proches voisins

- pb identique diagramme Voronoï
- (= triangulation de Delaunay)
- surface de décision
- pb : sensible outliers
- solution : k-NN

36 - Feuille de données François Poulet

Les plus proches voisins

- généralisation aux k-ppv (k impair, vote)

K = 1

K = 3

37 - Feuille de données François Poulet

Les plus proches voisins

- pb : sensible aussi aux dimensions non pertinentes => mauvaise classification

38 - Feuille de données François Poulet

Les plus proches voisins

- sensible unités
- normalisation

39 - Feuille de données François Poulet

Les plus proches voisins

- distance euclidienne
- n'importe quelle distance ou similarité
 - distance de Manhattan
 - distance de Mahalanobis
- mesures de distances spécifiques domaine :
 - chaîne d'ADN
 - séries temporelles
 - images
 - empreintes digitales...

40 - Feuille de données François Poulet

Les plus proches voisins

- Avantages :
 - simple implémenter
 - traite dimensions corrélées
 - peut utiliser n'importe quelle distance
 - traite flot de données
- Inconvénients :
 - sensible aux dimensions non pertinentes
 - coût classification grands ensembles de données
 - convient mieux aux données à valeurs réelles

41 - Feuille de données François Poulet

Arbres de décision

- coupes successives pour séparer données
- ex : 2 classes

42 - Feuille de données François Poulet

Arbres de décision

- étape 1 :
 - si $x > v1$ alors classe = cercle
 - feuille de l'arbre de décision (données éliminées)
- étape 2 :
 - traite le reste
 - si $y > v2$ alors
 - carré
 - sinon
 - cercle

49 - Fodille de données

François Poulet

Arbres de décision

- critère de coupe : gain d'information
 - un choix possible (ID3, C4.5)
 - choisit var + haut gain
 - soit 2 classes P et N, ensemble S
 - n élts N et p élts de P, entropie de S est :

$$H(S) = -\frac{p}{p+n} \log_2 \left(\frac{p}{p+n} \right) - \frac{n}{p+n} \log_2 \left(\frac{n}{p+n} \right)$$
- gain de la coupe suivant attribut A est :

$$Gain(A) = H(Current\ set) - \sum_i H(child_sets)$$

50 - Fodille de données

François Poulet

Arbres de décision

- exemple simple :

nom	long. cheveux	poids	age	sexe
M1	0"	250	36	M
F1	9"	150	34	F
M2	2"	90	10	M
F2	6"	78	8	F
F3	4"	20	1	F
M3	1"	170	70	M
F4	8"	160	41	F
M4	10"	180	38	M
M5	7"	200	45	M

51 - Fodille de données

François Poulet

Arbres de décision

- entropy(4F,5M) = $-(4/9)\log_2(4/9) - (5/9)\log_2(5/9)=0.9911$
- élimine les noms !!!
- test coupe long.cheveux :
 - tri suivant valeurs croissantes

nom	long. cheveux	poids	age	sexe
M1	0	250	36	M
M3	1	170	70	M
M2	2	90	10	M
F3	4	20	1	F
F2	6	78	8	F
M5	7	200	45	M
F4	8	160	41	F
F1	9	150	34	F
M4	10	180	38	M

52 - Fodille de données

François Poulet

Arbres de décision

- coupes possibles aux changements de classe
 - coupes possibles : >2, >6, >7, >9
- chaque coupe possible : calcule entropie
- garde la coupe avec + faible entropie
- idem autres attributs
 - garde le min global = coupe sur attribut avec valeur
 - poids > 160
 - (Gain(poids>160)=0.9911– (5/9 * 0.7219 + 4/9 * 0)=0.5900)

53 - Fodille de données

François Poulet

Arbres de décision

nom	long. cheveux	poids	age	sexe
F3	4	20	1	F
F2	6	78	8	F
M2	2	90	10	M
F1	9	150	34	F
F4	8	160	41	F
M3	1	170	70	M
M4	10	180	38	M
M5	7	200	45	M
M1	0	250	36	M

feuille de l'arbre

54 - Fodille de données

François Poulet

Arbres de décision

- poids > 160 (Gain(poids>160)=0.9911– (5/9 * 0.7219 + 4/9 * 0)=0.5900)

poids > 160

oui

non

4M

1M,4F
- réitère sur sous arbres

55 - Feuille de données François Poulet

Arbres de décision

- ne considère que 5 individus restant
- coupe sur les cheveux :

```

    graph TD
      A[poids > 160] -- oui --> B[4M]
      A -- non --> C[cheveux > 2]
      C -- oui --> D[4F]
      C -- non --> E[1M]
  
```

56 - Feuille de données François Poulet

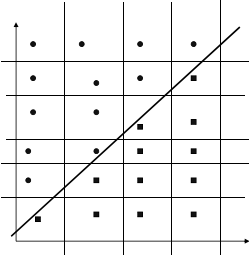
Arbres de décision

- autres possibilités que le gain
- critère de Gini (CART) :
 - soit 2 classes P et N, ensemble S
 - n élt N et p élt de P, Gini de S est :
- plus petite valeur

57 - Feuille de données François Poulet

Arbres de décision

- cas défavorable :
 - frontière oblique
 - approximation
 - marches d'escalier
 - arbre très profond
 - peu lisible
- augmente finesse
- diminue erreur



58 - Feuille de données François Poulet

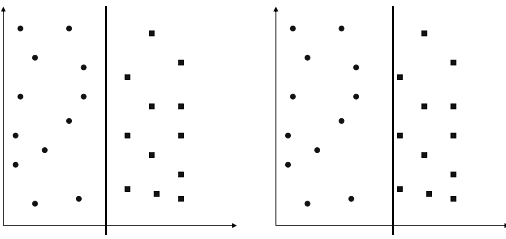
Arbres de décision

- sur-apprentissage (overfitting)
 - commun plusieurs méthodes d'apprentissage
- trop proche ensemble apprentissage
- petite variation test set => changement classe
- tx bonne classification élevé trn set (faible erreur apparente)
 - forte dégradation tst set (forte erreur réelle)
- élagage de l'arbre (pruning)

59 - Feuille de données François Poulet

Arbres de décision

- idée intuitive :



60 - Feuille de données François Poulet

Arbres de décision

- pré ou post élagage
- pré élagage
 - ne coupe que si critère > seuil
 - pb fixer le bon seuil et le bon critère
- post élagage
 - supprime branches
 - utilise test set pour tester efficacité
 - garde le meilleur
- encore beaucoup de recherches

61 - Feuille de données François Poulet

Arbres de décision

- avantages :
 - compréhension
 - génération de règles (si alors sinon...)
- inconvénients :
 - sur-apprentissage
 - très volumineux (nécessité élagage)
 - coupe univariée
 - pas de traitement de flot de données

62 - Feuille de données François Poulet

Arbres de décision

- coupe unaire (une seule coupe)
- coupe univariée (selon une seule variable)
- CART, C4.5, Random Forest, ...
- OC1 (Oblique Cut) : arbre oblique

63 - Feuille de données François Poulet

Arbres de décision

- OC1
- coupe hyperplan en dimension n (données dimension n)
- avantage : arbre bcp + pt
- inconvénient : compréhension

64 - Feuille de données François Poulet

Arbre de décision

- OC1 :
- recherche meilleure coupe unaire, univariée
- NP-complet
- complexité augmentée
 - coupe : n variables

65 - Feuille de données François Poulet

Arbres de décision

- OC1 :
- recherche meilleur hyperplan de séparation des données
- recherche exhaustive impossible
 - heuristique :
 - tirage aléatoire pour
 - coefficients hyperplan ("pente")
 - scalaire ("position")
 - très coûteux temps exécution

66 - Feuille de données François Poulet

Arbres de décision

- OC1 : exemple de résultat
- Australian : 14 dimensions, 2 classes, 690 individus
- Root Hyperplane: Left = [303,30], Right = [44,244], (87.83%) 85.51% accuracy
- $-187.354065 x[1] + -6.004185 x[2] + 0.110710 x[3] + -16.884249 x[4] + 2.762119 x[5] + 0.641304 x[6] + 50.861160 x[7] + 406.713806 x[8] + 104.010483 x[9] + 21.005371 x[10] + -38.602798 x[11] + -55.722027 x[12] + -0.002645 x[13] + 0.063269 x[14] + -0.312639 = 0$
- 1 coupe = hyperplan 14-D, 2 feuilles

67 - Fodille de données François Poulet

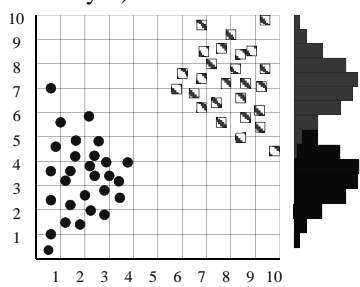
Arbres de décision

- CART: 85.5%, C4.5: 84.5% (85 feuilles)
- avantages :
 - arbre très compact
 - meilleure coupe que univariée
 - meilleure précision
- inconvénients :
 - coût calcul
 - compréhension du résultat
 - application cartes de crédit (tout modifié)

68 - Fodille de données François Poulet

Bayésien naïf

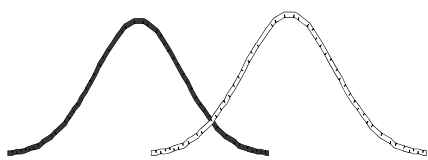
- NB (Naïve Bayes)



69 - Fodille de données François Poulet

Bayésien naïf

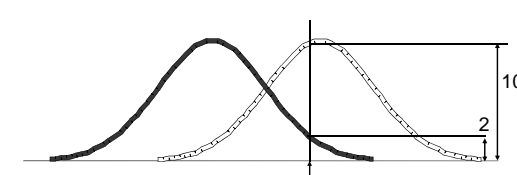
- remplace histo par distribution



70 - Fodille de données François Poulet

Bayésien naïf

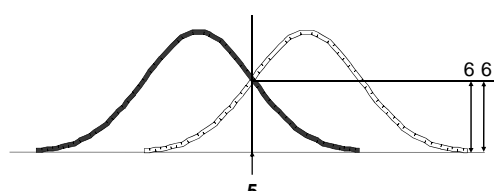
- $p(\text{rouge} | 3) = 2 / (10+2) = 0.166$
- $p(\text{blanc} | 3) = 10 / (10+2) = 0.833$



71 - Fodille de données François Poulet

Bayésien naïf

- $p(\text{rouge}|5) = 6 / (6+6) = 0.5$
- $p(\text{blanche}|5) = 6 / (6+6) = 0.5$



72 - Fodille de données François Poulet

Bayésien naïf

- Théorème de Bayes :
- $p(c_j | d) = \frac{p(d | c_j)p(c_j)}{p(d)}$
- $p(c_j | d)$ = proba individu $d \in$ classe c_j
- $p(d | c_j)$ = proba individu d connaissant c_j
- $p(c_j)$ = proba classe c_j
- $p(d)$ = proba de d (indépendant de la classe)

73 - Fodille de données François Poulet

Bayésien naïf

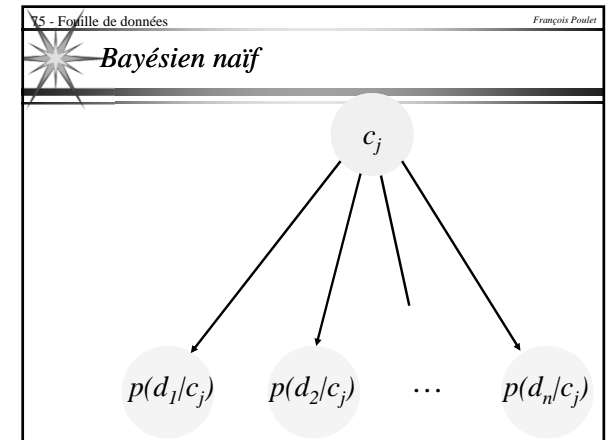
- exemple : 2 classes = {H,F}
- claud est H ou F?
- $p(H|claud) ? p(F|claud)$
- $p(H|claud) = \frac{p(claud|H)p(H)}{p(claud)}$
- $p(H|claud) = \frac{1/3 * 3/8}{3/8} = 1/3$
- $p(F|claud) = \frac{2/5 * 5/8}{3/8} = 2/3$

Name	Sex
Claude	H
Carine	F
Claude	F
Claude	F
Albert	H
Marie	F
Nina	F
Serge	H

74 - Fodille de données François Poulet

Bayésien naïf

- plusieurs attributs ?
- suppose distributions indépendantes
- produit des probabilités
- $p(d|c_j) = p(d_1|c_j) * p(d_2|c_j) * \dots * p(d_n|c_j)$
- représentation sous la forme d'un graphe



76 - Fodille de données François Poulet

Bayésien naïf

```

graph TD
    c_j((c_j)) --> p1((p(d_1/c_j)))
    c_j --> p2((p(d_2/c_j)))
    c_j --> dots[...]
    c_j --> pn((p(d_n/c_j)))
  
```

Animal	Mass > 10kg	
Cat	Yes	0.15
	No	0.85
Dog	Yes	0.91
	No	0.09
Pig	Yes	0.99
	No	0.01

Animal	Color	
Cat	Black	0.33
	White	0.23
	Brown	0.44
Dog	Black	0.097
	White	0.003
	Brown	0.90
Pig	Black	0.04
	White	0.01
	Brown	0.95

77 - Fodille de données François Poulet

Bayésien naïf

Spam detection software, running on the system "verdi.aaaa-ouest.fr", has identified this incoming email as possible spam. The original message has been attached to this so you can view it (if it isn't spam) or label similar future email. If you have any questions, see Centre Informatique <ccentrinfo@aaaa-ouest.fr> for details.

Content analysis details: (15.9 points, 5.0 required)

pts rule name	description
2.0 DATE_IN_FUTURE_96_XX	Date: is 96 hours or more after Received: date
0.1 RAZOR2_CF_RANGE_51_100	BODY: Razor2 gives confidence level above 50% [cf: 100]
3.5 BAYES_99	BODY: Bayesian spam probability is 99 to 100% [score: 1.0000]
1.5 RAZOR2_CHECK	Listed in Razor2 (http://razor.sf.net/)
1.0 URIBL_SBL	Contains an URL listed in the SBL blocklist [URLs: sewerio.com]
0.4 URIBL_AB_SURBL	Contains an URL listed in the AB SURBL blocklist [URLs: sewerio.com]
3.2 URIBL_OB_SURBL	Contains an URL listed in the OB SURBL blocklist [URLs: sewerio.com] ...

78 - Fodille de données François Poulet

Bayésien naïf

- avantages :
 - peu coûteux
 - insensible aux outliers
 - traite flot de données
- inconvenients :
 - indépendance des dimensions
 - (rarement le cas en réalité)

89 - Fouille de données

François Poullet

SVM

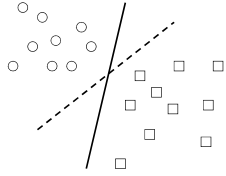
- Support Vector Machine
- Séparateurs à Vaste Marge
- utilisation récente en fouille données (V.Vapnik, 1995)
- premiers travaux plus anciens :
 - (Vapnik, 1965)
 - (Mangasarian, 1964)
- reconnaissance de formes

90 - Fouille de données

François Poullet

SVM

- SVM : mode le + simple : géométrique
- meilleur hyperplan séparation données

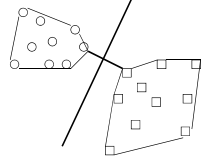


91 - Fouille de données

François Poullet

SVM

- méthode (données linéairement séparables)
 - enveloppe convexe 2 classes
 - recherche les 2 plus proches voisins
 - meilleur plan bissectrice

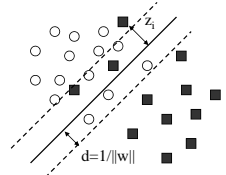


92 - Fouille de données

François Poullet

SVM

- principe :
 - ens. données en dimension d étiquetées ± 1
 - cherche meilleur hyperplan de séparation
 - meilleur = + éloigné des 2 classes, - d'erreurs



93 - Fouille de données

François Poullet

SVM

- séparatrice non linéaire

espace initial

x

y

Φ

↓

espace feature

x

y

x²

y²

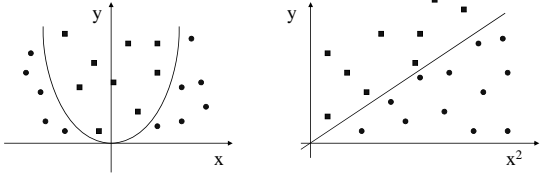
- hyperplan (espace feature)

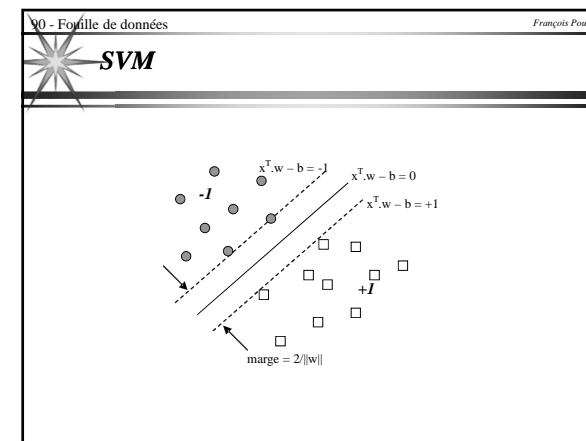
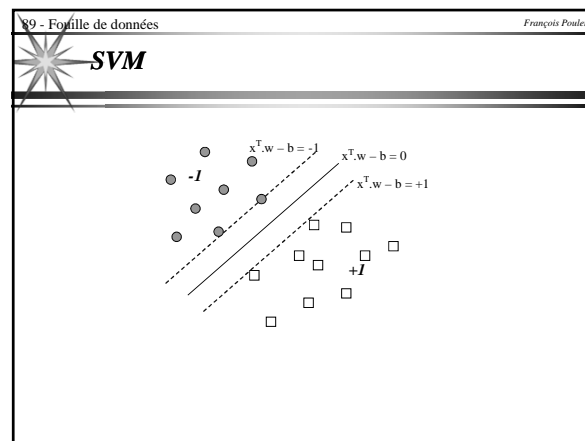
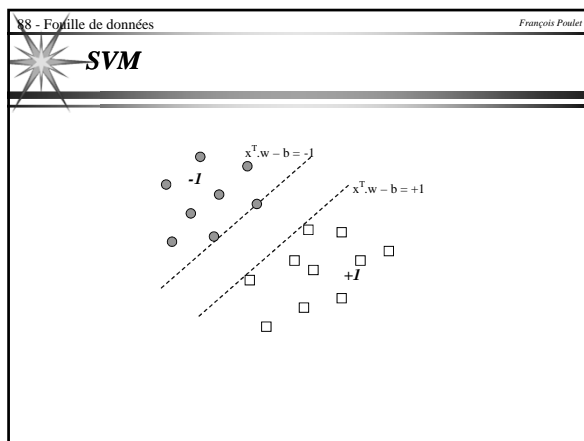
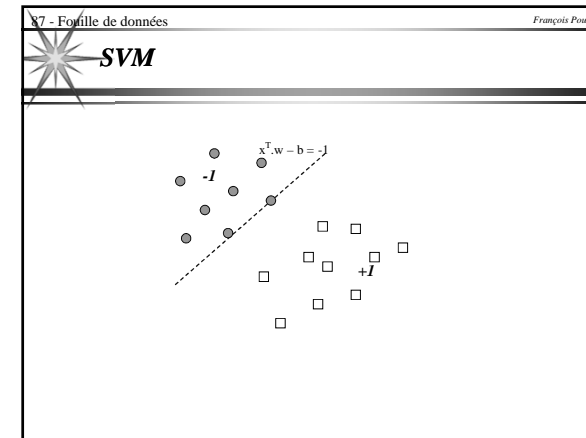
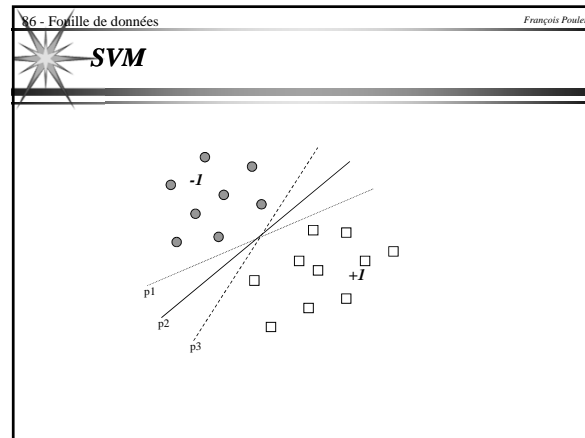
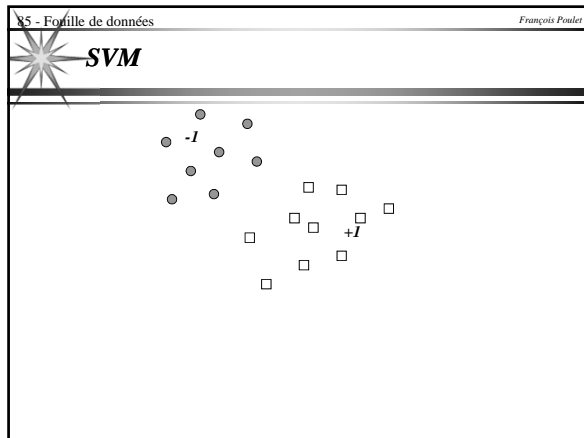
94 - Fouille de données

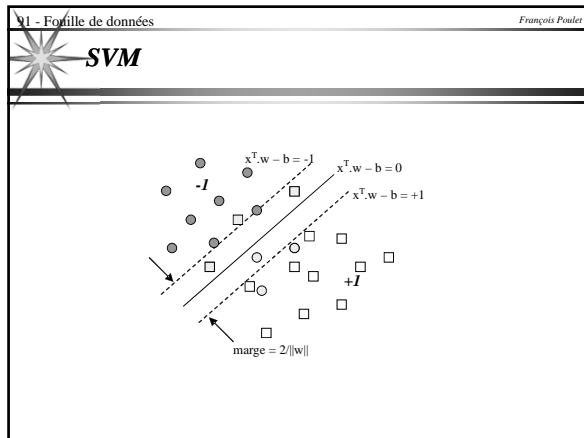
François Poullet

SVM

- exemple :







92 - Feuille de données

François Poullet

SVM

- maximisation de la marge + minimisation des erreurs

$$\min f(z, w, b) = (v/2) \|z\| + (1/2) \|w\|^2$$
 avec $D(Aw - eb) + z \geq e$ (1)
 où $z \geq 0$ variable de ressort et v constante positive
- résolution du programme quadratique (1) : w, b
- classification d'un nouvel individu x : $\text{sign}(x^T w - b)$

93 - Feuille de données

François Poullet

SVM

- avantages :
 - très performants (parmi les meilleurs)
 - aucune supposition sur données
 - pas de sur apprentissage
 - robustesse (mathématique VC-dimension)
- inconvénients :
 - coûteux (mémoire et temps calcul cf. scalabilité)
 - sensibles aux paramètres
 - que 2 classes

94 - Feuille de données

François Poullet

Conclusion classification supervisée

- 4 familles algorithmes :
 - k-NN, arbres de décision, NB, SVM
- autres :
 - réseaux de neurones
 - algorithmes génétiques
 - méthodes visuelles
- classification non supervisée...

95 - Feuille de données

François Poullet

Clustering

- appelé : classification, segmentation...
- but :
 - regrouper les individus en groupes (clusters)
 - grande similarité intra-cluster
 - grande dissimilarité inter-cluster
- définition similarité
- méthodes :
 - ascendante (on regroupe successivement)
 - descendante (on découpe successivement)

96 - Feuille de données

François Poullet

Clustering

- similarité basée distance
- distance :
 - symétrie : $d(x, y) = d(y, x)$
 - séparation : $d(x, y) = 0 \Leftrightarrow x = y$
 - inégalité triangulaire : $d(x, z) \leq d(x, y) + d(y, z)$

97 - Feuille de données François Poulet

Clustering

- propriétés intéressantes :
 - traitement de grands ensembles de données
 - traitement de différents types de données
 - résistant au bruit et aux outliers
 - non sensible à l'ordre de traitement des données
 - inclusion de contraintes utilisateur
 - interprétabilité et utilisabilité

98 - Feuille de données François Poulet

Clustering

- clustering hiérarchique
- dendrogramme :
 - similarité = hauteur du noeud interne partagé le plus bas

99 - Feuille de données François Poulet

Clustering

- plus grand saut = nb clusters (pas tjrs visible...)

100 - Feuille de données François Poulet

Clustering

- détection d'outlier

101 - Feuille de données François Poulet

Clustering

- calcul distance point-cluster ou cluster-cluster
 - distance du plus proche voisin : 2 points les + proches de chaque cluster
 - distance du voisin le plus éloigné : 2 points les + éloignés de chaque cluster
 - distance moyenne : moyenne des distances entre les paires de points
 - critère de Wards : minimise la variance entre 2 clusters

102 - Feuille de données François Poulet

Clustering

Average linkage

Single linkage

Wards linkage

103 - Feuille de données François Poulet

Clustering

- clustering hiérarchique
 - ascendant : part de chaque ind. cherche la meilleure paire à assembler, réitère ua 1 cluster
 - descendant : part d'un cluster, cherche la meilleure coupe en deux, réitère ua 1 individu / cluster

104 - Feuille de données François Poulet

Clustering

- conclusion méthodes hiérarchiques
 - pas besoin de fixer k (nb clusters)
 - facilement interprétable pour certaines applications
- passage a l'échelle difficile ($O(n^2)$)
- problème des optima locaux
- interprétation subjective
- pas la meilleure partition en k

105 - Feuille de données François Poulet

Clustering

- méthodes de partitionnement
 - non hiérarchique
 - utilisateur doit préciser le nb de clusters souhaité

106 - Feuille de données François Poulet

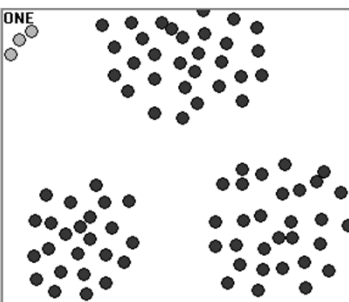
Clustering

- k-means
 - choix valeur de k
 - init : k centres de clusters
 - tant que pas fini
 - pour tous les points
 - assigne le point au cluster dont le centre est le plus proche
 - fpour
 - pour chaque cluster
 - calcul le nouveau centre
 - fpour
 - si aucun point n'a changé de cluster=> fini
 - ftantque

107 - Feuille de données François Poulet

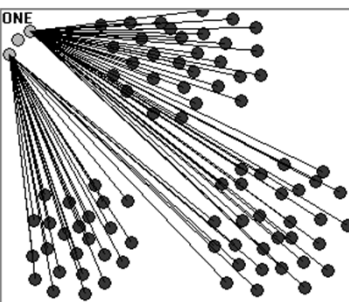
Clustering - k-means

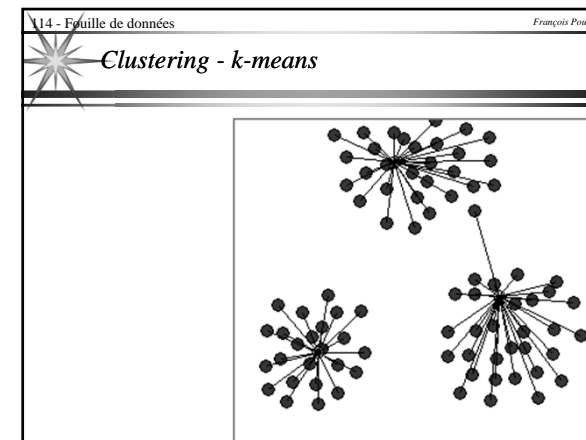
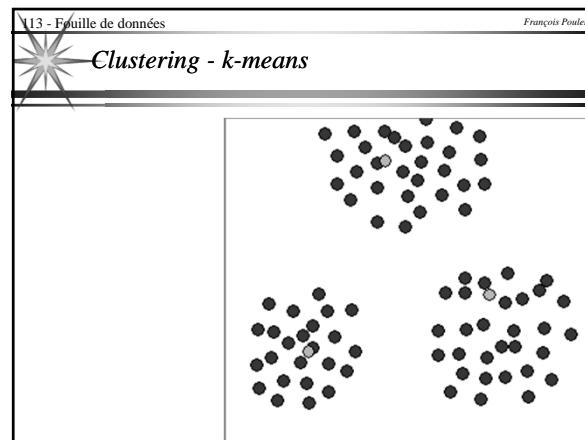
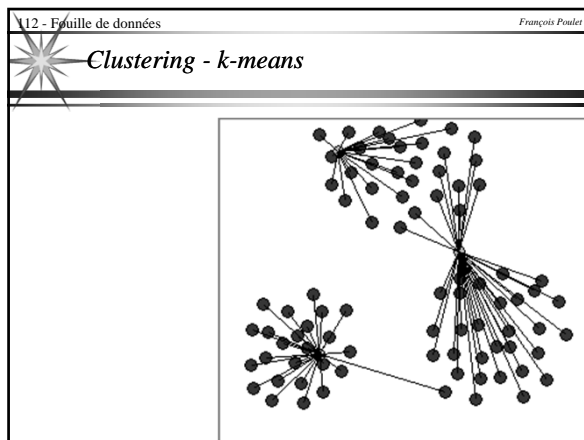
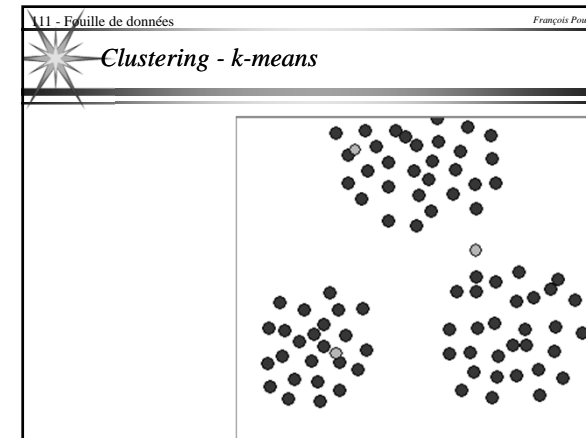
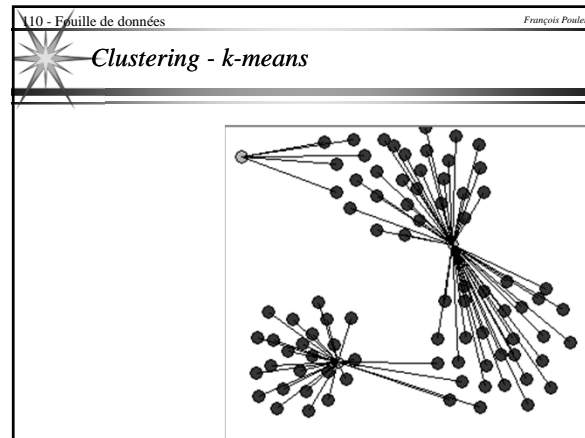
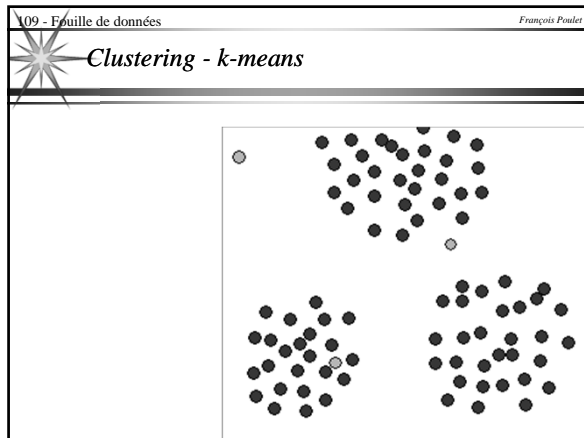
- exemple :
 - igs.wesleyan.edu/demos.htm

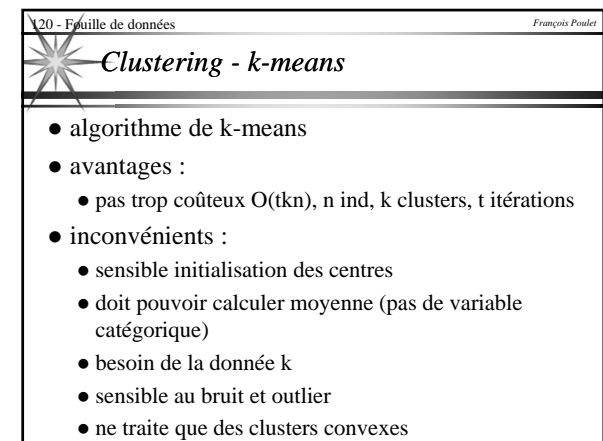
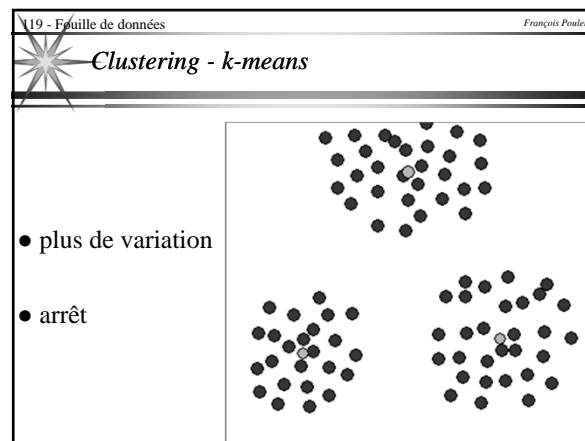
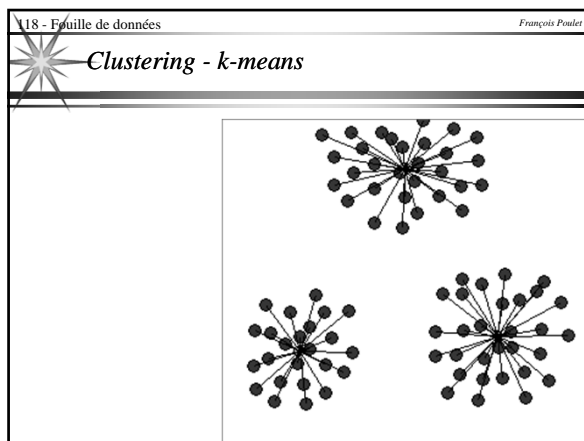
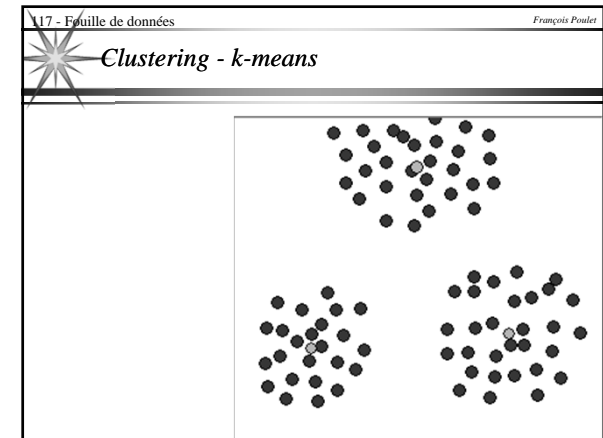
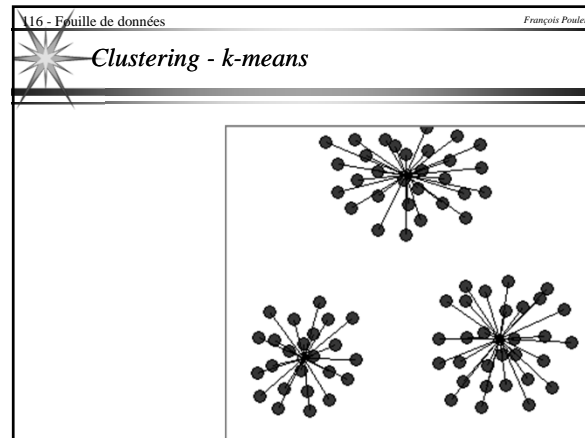
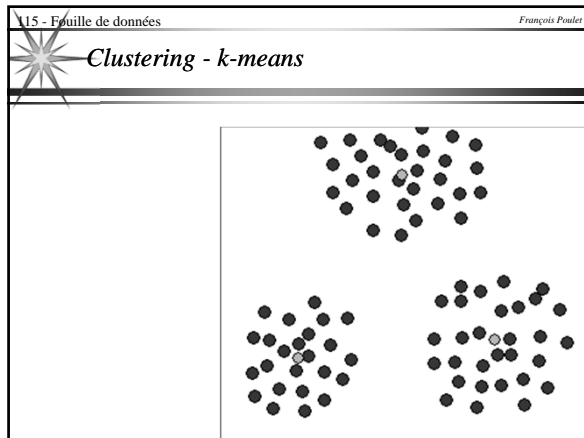


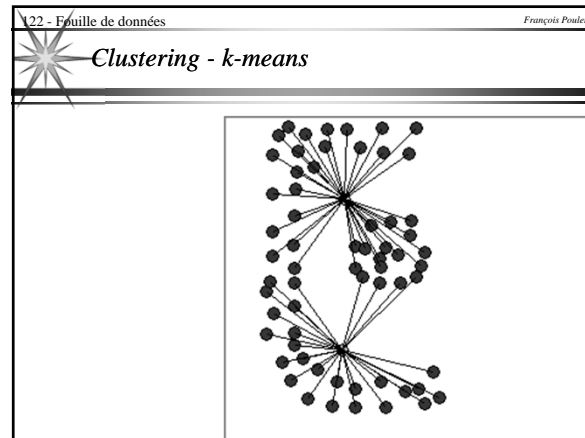
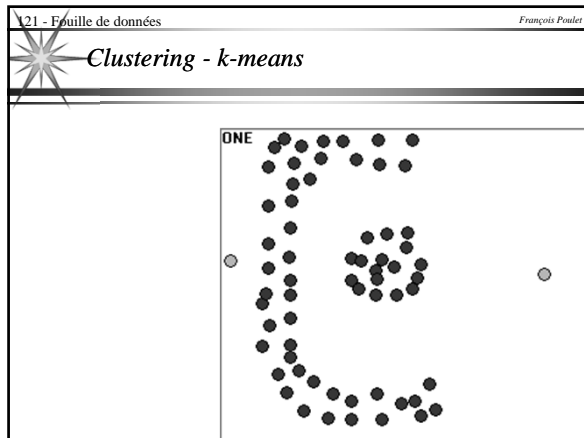
108 - Feuille de données François Poulet

Clustering - k-means









123 - Feuille de données François Poulet

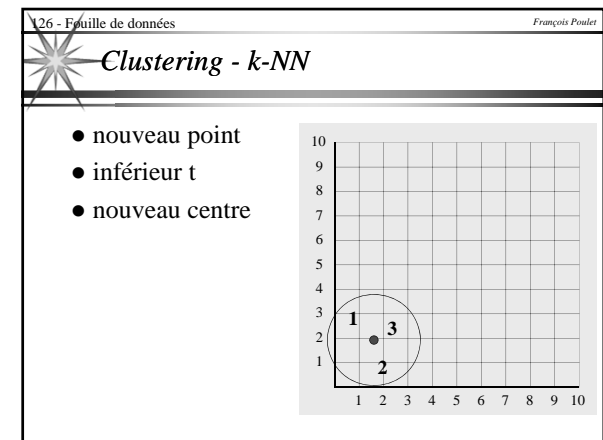
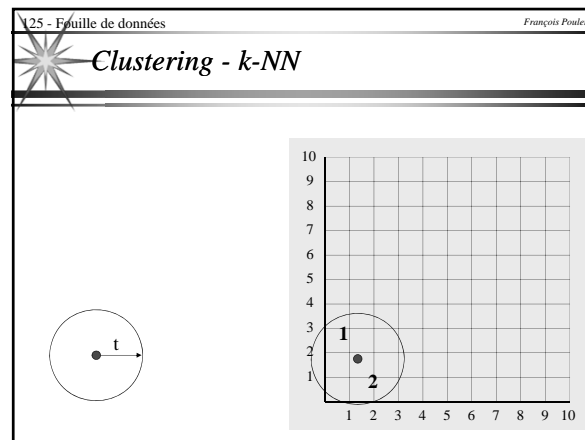
Clustering

- variante de k-means : k-medoids
- centres clusters = points réels

124 - Feuille de données François Poulet

Clustering

- plus proche voisin (<>classification sup.)
- points itérativement ajoutés
 - cluster existant si proche
 - nouveau cluster si éloigné
- fixer seuil t
- incrémental





Clustering - k -NN

- $> t$
- nouveau cluster
- inconvénients :
 - t ?
 - dépendant ordre

