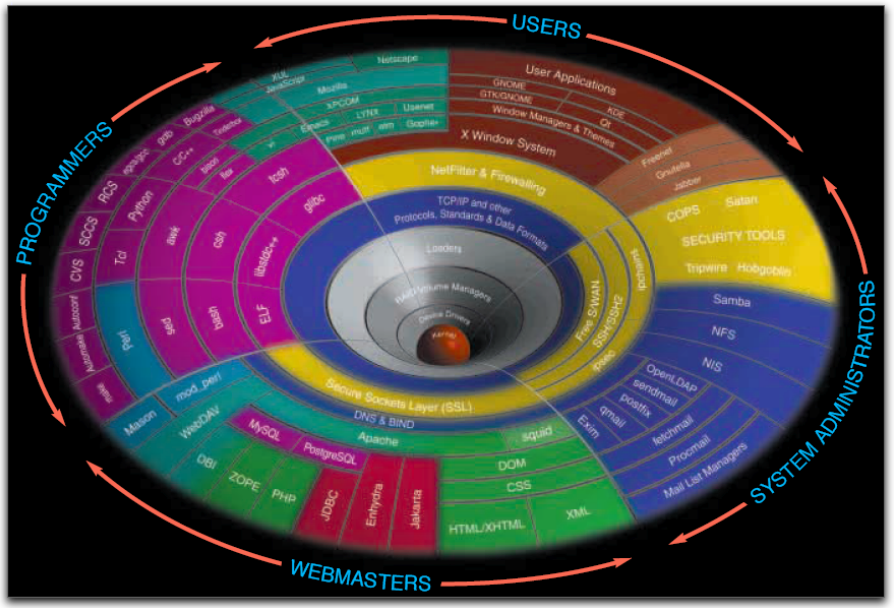
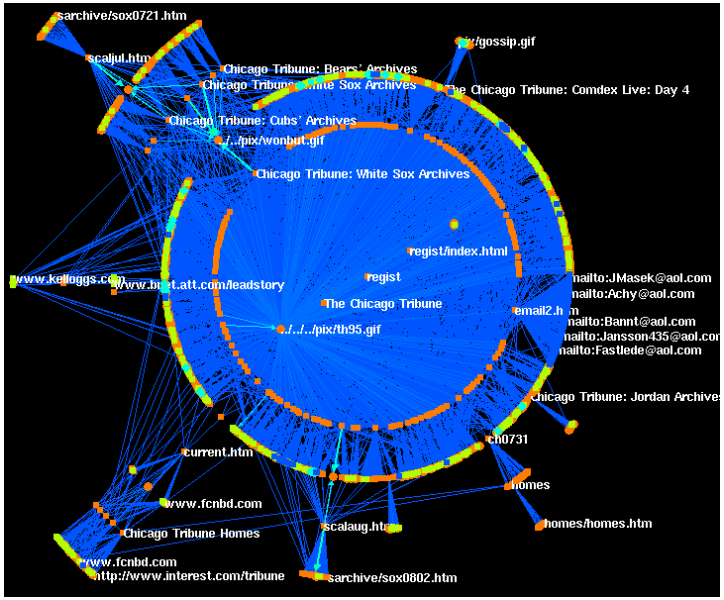
Modèle de classification de données

P-F. M., 2020-2021

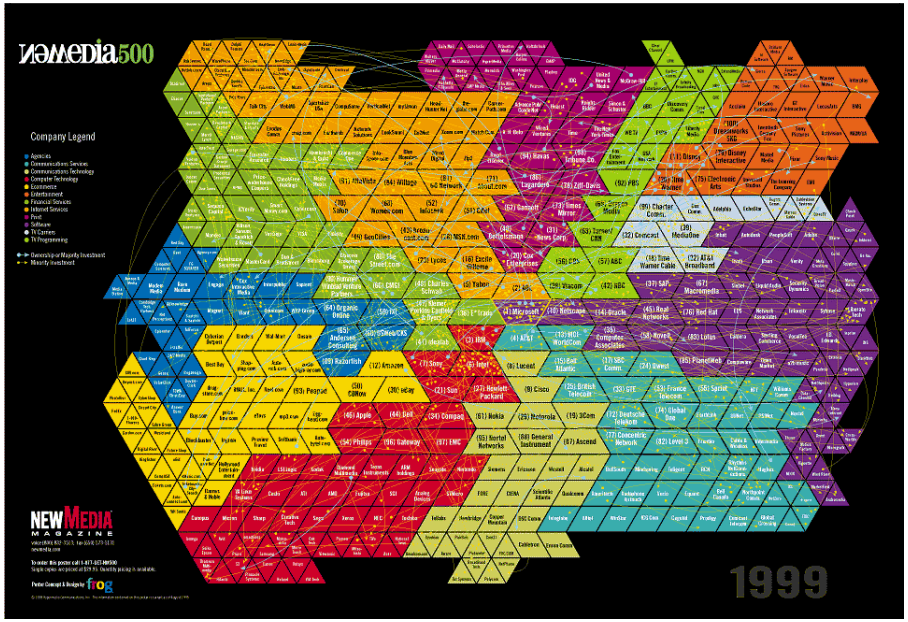
Pourquoi classer ?



Cartes de sites Web

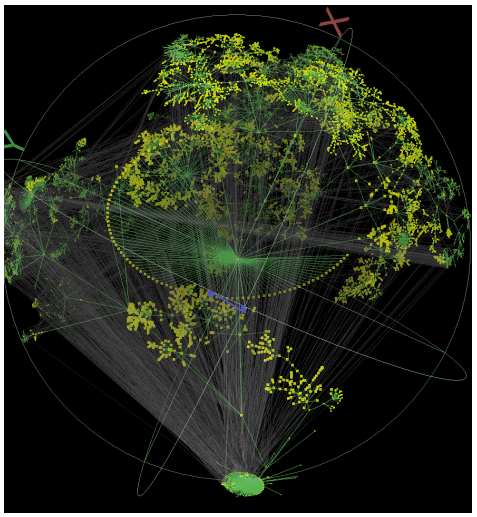


Cartes de sites Web : Inter-relations financières entre 50 000 sociétés

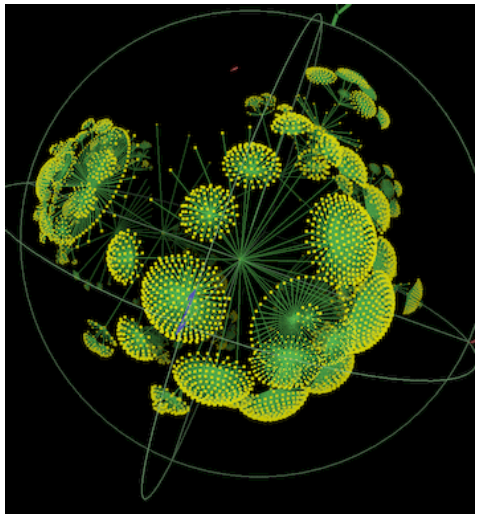


Visualisation de graphes (Walrus)

http://www.caida.org/tools/visualization/walrus/



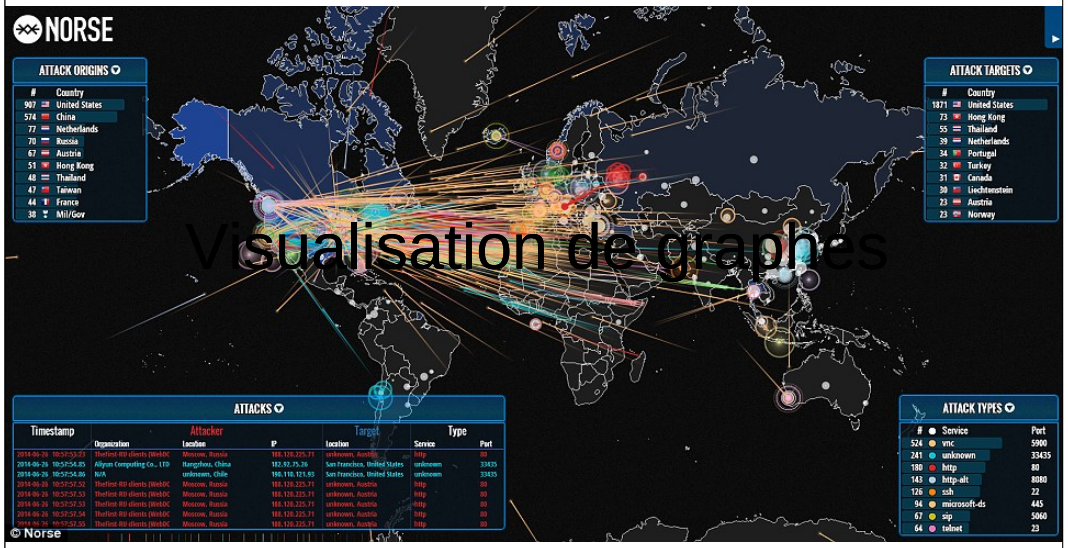
54 893 nœuds, 79 409 liens



535 102 nœuds, 601 678 liens

Géolocalisation de cyber-attaques

http://map.norsecorp.com/



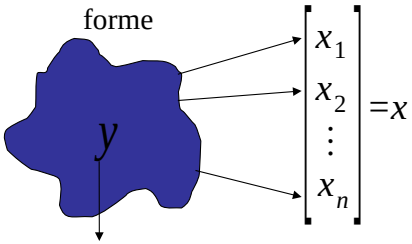
Classification supervisée des documents

- La classification est une tâche qui consiste à répartir des objets dans des catégories :
  - Lorsque les catégories sont définies a priori, on parle de classification supervisée
  - A *contrario*, lorsque les catégories sont découvertes a *posteriori*, on parle de classification non supervisée

2 Types de Classification

- Classification supervisée
  - Avec des exemples d'apprentissage pour lesquels la classe d'appartenance est connue a priori
- Classification non supervisée (Clustering)
  - Sans exemples d'apprentissage, on part d'un ensemble d'individus sans classe d'appartenance

Concepts de base de la classification



Vecteur de caractéristiques $x \in X$

- Définit à partir d'un vecteur d'observations, de mesures.

- **x** est un point dans l'espace des caractéristiques **X**, appelé aussi espace de représentation.

Etat caché $y \in Y$

- Ne peut pas être directement mesuré.

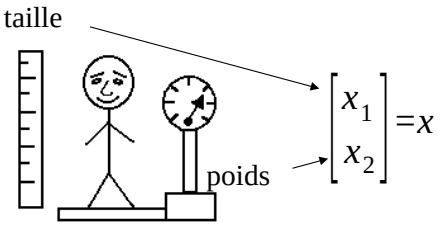
- Les formes ayant le même état caché appartiennent à la même classe.

Tâche :

- Concevoir un classifieur (règle de décision) $q: X \rightarrow Y$

Qui permet de décider de l'état caché d'une forme sur la base des observations effectuées sur cette forme.

Exemple



Tâche : classification jockey OU basketteur.

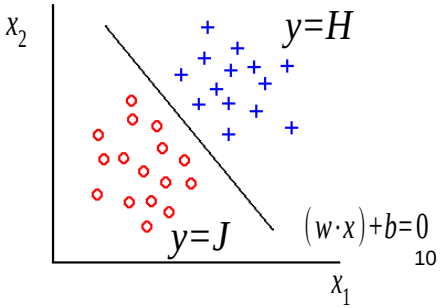
L'ensemble des états cachés est $Y = \{H, J\}$

L'espace des caractéristiques est: $X = \Re^2$

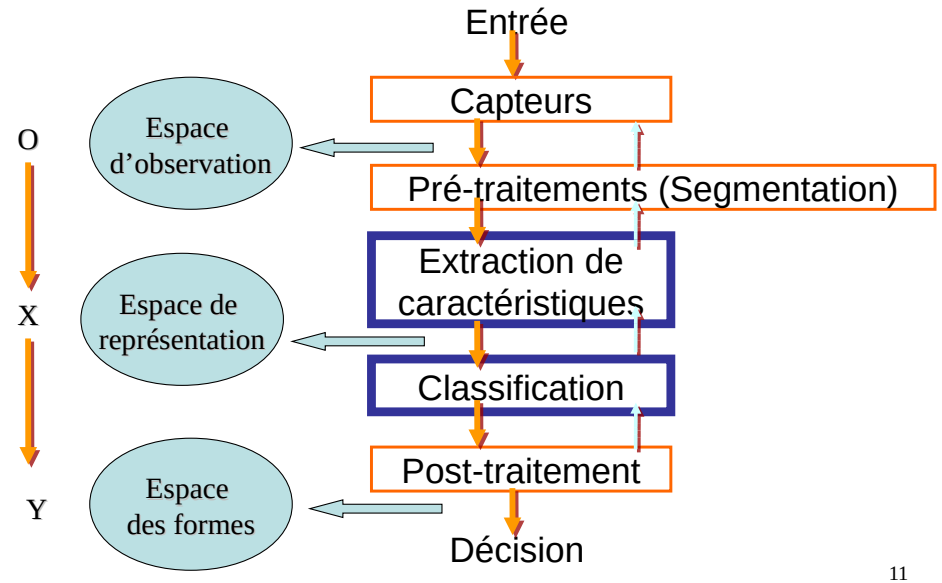
Exemples d'apprentissage $\{(x_1, y_1), \ldots, (x_l, y_l)\}$

Classifieur linéaire:

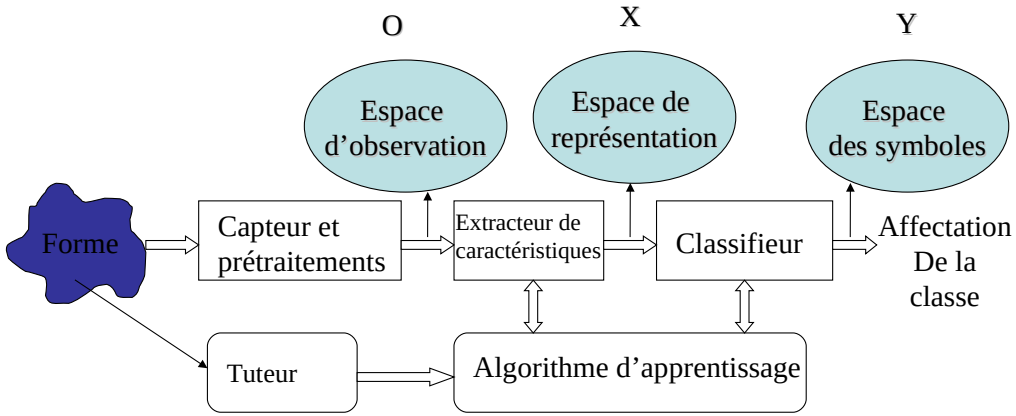
$$q(x) = \begin{cases} H & \text{si } (w \cdot x) + b \geq 0 \\ J & \text{si } (w \cdot x) + b < 0 \end{cases}$$



Fonctions d'un système de classification



Composants d'un système de classification



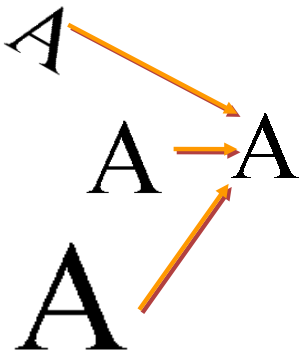
Pre-traitement : Normalisation

Image

- Rotation

- Translation

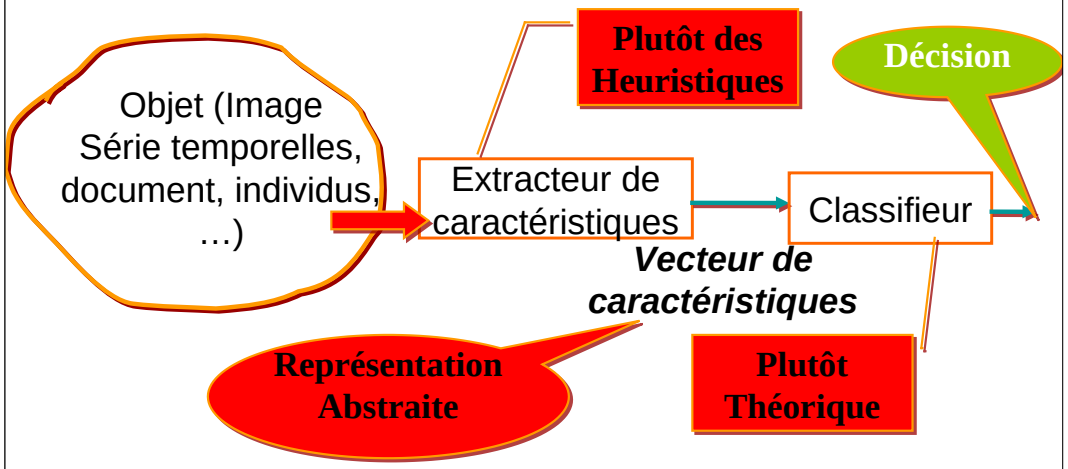
- Echelle (Scaling)



Parole

- Contrôle de gain automatique
- Alignement temporel

Extraction de caractéristiques et classifieur



Exemple de caractéristiques classiques pour les images

- Couleur
  - Systèmes de codage de la couleur (RGB, YUV, HSI, etc.)
  - Histogramme de couleur et ses moments
- Forme globale
  - Moments (Hu, Zernike)
  - Coefficients de Fourier
  - Coefficients type « ondelettes »
  - Descripteur de flux (flow) optique pour les vidéos
- Texture
  - Matrice de Cooccurrence
  - Transformée de Gabor et ondelettes
- Forme locale
  - Courbure, points d'inflexion, changement d'angle, …

Exemple de caractéristiques classiques pour le traitement de la parole

- Fréquence fondamentale (Pitch)
- Voisé / non voisé
- Formants (pics dans le spectre d'énergie)
- Silences
- Phonèmes

Exemple pour les flux IP

- Type d'application (ssh, http, ftp, icmp, ...)
- Longueur (#packets)
- Direction
- Ports
- Adresse IP
- « Flags » de la couche transport
- Time-stamps
- Histogrammes sur les octets de la "payload"
- ...

Extraction de caractéristiques

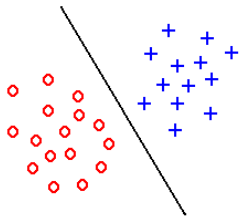
- Utilisation d'heuristiques
- dépend de l'application
- Il y a cependant quelques règles et outils généraux exploitables pour choisir des caractéristiques :
  - La dimension du vecteur de caractéristiques doit être très inférieure au nombre d'exemples disponibles pour l'apprentissage (The curse of dimensionality)

- **Réduction de dimensionalité**
  - Analyse en composantes principales
  - Analyse discriminante (Fisher Linear Discriminant)
  - Auto-encodeurs (Neural nets, deep learning)

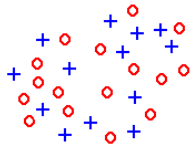
Extraction de caractéristiques

Tâche : extraire des caractéristiques adéquates pour la classification.

Bonnes caractéristiques:
- les objets appartenant à la même classe ont des caractéristiques similaires.
- les objets appartenant à des classes différentes ont des caractéristiques différentes.



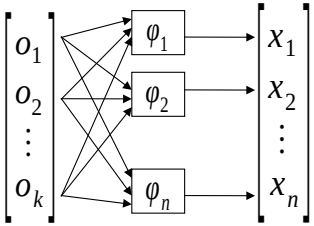
"Bonnes" caractéristiques



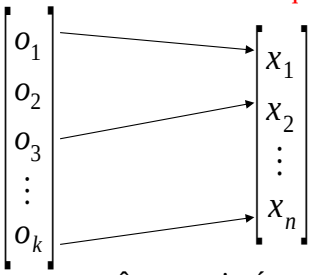
"Mauvaises" caractéristiques

Extraction de caractéristiques

Extraction des caractéristiques



Sélection des caractéristiques



Le problème d'extraction de bonnes caractéristiques peut être exprimé comme un problème d'optimisation sur la famille de fonctions paramétriques. $\varphi(\theta)$

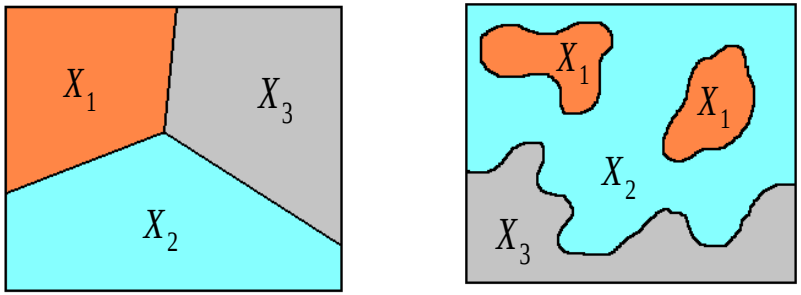
Méthodes supervisées : La fonction objectif est un critère de séparabilité (discrimination) des objets étiquetés (exemple : analyse discriminante, LDA).

Méthodes non supervisées : on recherche des représentations dans des espaces à faibles dimensions tout en préservant les caractéristiques des données d'entrées (par exemple analyse en composante principale (PCA).

Classifieur

Un classifieur partitionne l'espace de caractéristiques *X* en **régions étiquetées** par la classe d'appartenance de telle sorte que **:**

$X = X_1 \cup X_2 \cup \ldots \cup X_{|Y|}$ et $X_1 \cap X_2 \cap \ldots \cap X_{|Y|} = \{0\}$



La classification consiste à déterminer à quelle région le vecteur de caractéristique **x** appartient. Les frontières séparant les régions sont appelées les régions de décision..

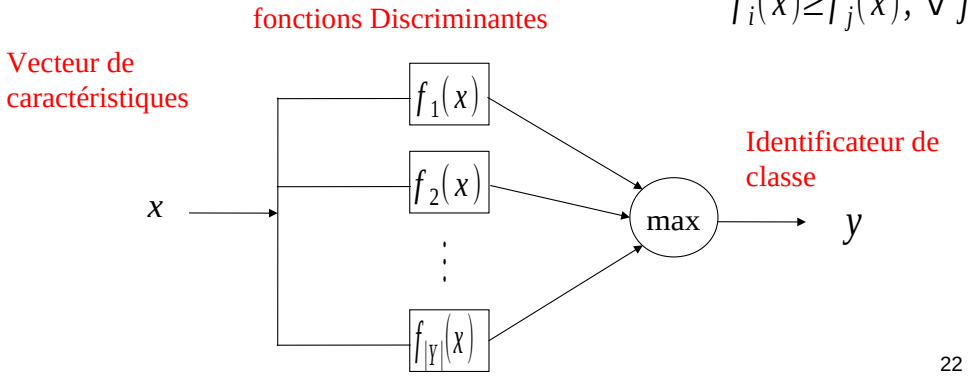
Representation d'un classifieur

Un classifieur est généralement représenté sous la forme d'un ensemble de fonctions discriminantes :

$$f_i(x): X \rightarrow \Re, i = 1, \ldots, |Y|$$

Le classifieur affecte l'objet associé au vecteur de caractéristique ***x*** à la classe ***i*** si :

$$f_i(x) \geq f_j(x), \forall j$$



Quelques notions de "Machine Learning"

Un petit pas pour l'homme, un grand pas pour l'IA ...

Test de turing : janvier 2018, les assistants personnels d'Alibaba et de Microsoft surpassent l'humain sur le test SQuAD de l'université de Stanford

Éléments bibliographiques

- R. Duda, P. Hart & D. Stork, ***Pattern Classification*** (2nd ed.), Wiley (Required)
- T. Mitchell, ***Machine Learning***, McGraw-Hill (Recommended)
- Articles scientifiques (Wikipedia)

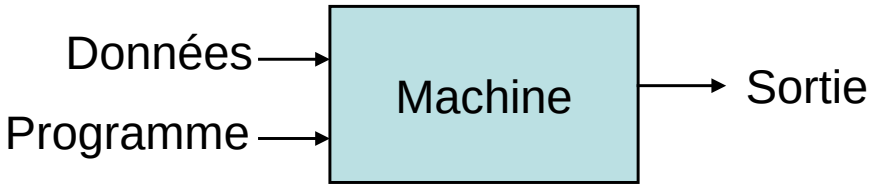
Quelques citations

- "A breakthrough in machine learning would be worth ten Microsofts" (Bill Gates, ex Chairman, Microsoft)
- "Machine learning is the next Internet" (Tony Tether, Director, DARPA)
- "Machine learning is the hot new thing" (John Hennessy, President, Stanford)
- "Web rankings today are mostly a matter of machine learning" (Prabhakar Raghavan, Dir. Research, Yahoo)
- Rise of the Robots Technology and the Threat of a Jobless Future (Martin Ford Book 2017)
- "On peut imaginer l'essor d'une nouvelle classe 'inutile', qui ne devra plus lutter contre son exploitation mais contre son insignifiance" (Yuval Noah Harari sept. 2018, auteur de "Sapiens")

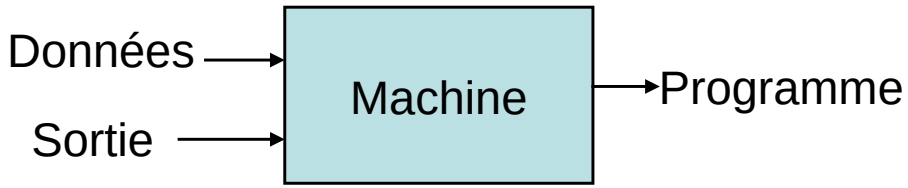
Qu'est-ce que le "Machine Learning" ?

- Le quoi:
  - L'automatisation de l'automate
  - Faire en sorte que les ordinateurs se programment eux-mêmes
- Le pourquoi
  - Écrire des programmes c'est le goulot d'étranglement
  - Laissons faire les données et quelques algorithmes de base !

Programmation traditionnelle



v.s. "Machine Learning"



Exemples d'applications

- Moteur de recherche sur le web, extraction d'information
- Biologie computationnelle
- Finance
- E-commerce
- Exploration spatiale
- Robotique
- Réseaux sociaux
- Programming, Debugging
- [autre (gaming)]

ML in a Nutshell

- Des dizaines d'algorithmes
- Des centaines de nouveaux algorithmes chaque année
- Chaque algorithme de ML s'appuie sur trois composantes :
  - **Modèle d'apprentissage**
  - **Optimisation**
  - **Évaluation**

Modèle d'apprentissage

- Arbres de décision
- Ensemble de règles / programmes logiques
- Basés instances (k-PPV)
- Modèles graphiques (Réseaux Bayésiens/Markoviens)
- Réseaux de neurones
- SVM (Support vector machines)
- Modèle d'ensembles
- Etc.

Évaluation (métriques)

- Taux de réussite (Accuracy)
- Précision et rappel (Precision and recall)
- Erreur quadratique (Squared error)
- Vraissemblance (Likelihood)
- Probabilité a posteriori (Posterior probability)
- Coût/Utilité (Cost / Utility)
- Marge (Margin)
- Entropie (Entropy)
- Kullback–Leibler divergence (K-L divergence)
- Etc.

Optimisation

- Optimisation combinatoire
  - Ex.: Greedy search (algorithme glouton)
- Optimisation convexe
  - Ex.: descente du gradient (gradient stochastique)
- Optimisation sous contraintes
  - E.x.: programmation linéaire (simplex)

Types d'apprentissage

- **Supervisé : (inductive) learning**

  Les données d'apprentissage comprennent les sorties désirées
- **Non supervisé: Unsupervised learning**

  Les données d'apprentissage ne comprennent pas les sorties désirées
- **Semi-supervised learning**

  Les données d'apprentissage comprennent quelques sorties désirées
- **Apprentissage par renforcement : Reinforcement learning**

  Récompenses sur des séquences d'actions prometteuses (robotique)

Apprentissage supervisé

- **Etant donnés** quelques exemples d'une fonction *(X, F(X))*
- **Prédiction:** déterminer la fonction *F(X)* pour de nouveaux exemples *X*
  - Si *F(X)* est discrète : Classification
  - Si *F(X) est c*ontinue : Régression
  - *F(X)* = Probabilité(*X*): Estimation de Probabilité
  -
- X: vecteur de caractéristiques (feature vector)
- y=F(X): **variable** de régression ou de classe

Ce que nous allons survoler

- **Apprentissage supervisé**
  - Arbre de décision (Decision tree induction)
  - Apprentissage de règles (Rule induction)
  - Apprentissage basé instance (Instance-based learning)
  - Apprentissage Bayésien (Bayesian learning)
  - Réseaux de neurones (Neural networks)
  - SVM (Support vector machines)
  - Modèles d'ensemble (Model ensembles)
  - Théorie de l'apprentissage (Learning theory)
- **Aprpentissage non supervisé**
  - Clustering (k-means)
  - Réduction de dimension (Dimensionality reduction)

ML en pratique

- Compréhension des principes de base
- Intégration de données, sélection, toilettage, pré-processing, etc.
- Modèles d'apprentissage
- Évaluation & Interprétation des résultats
- Consolidation et déploiement
- Bouclage (relevance feedback)

Apprentissage basé instances

- L'algorithme des **k**-plus proches voisins
- Algorithme très simple mais pas si mauvais dans la pratique
- Algorithme dit glouton (lazy) **O(N)** : reporte le calcul lors de la phase de classification

Apprentissage basé instances

- **Idée clé**
  - On stocke toutes les instances d'apprentissage **$\{(x_i, y_i=f(x_i))\}$, i=1..n**
- **Règle des k plus proches voisins**
  - Étant donnée un vecteur requête **x_q**, on recherche d'abord dans les données d'apprentissage les **k** plus proches voisins $\{x_j\}$, puis on décide :

$$\tilde{f}(x_q) \leftarrow h(\{f(x_j)\})$$

l'algorithme des k plus proches voisins

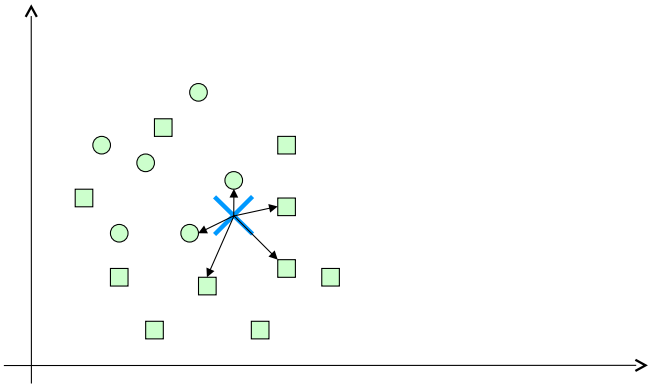
- On considère un ensemble **S** contenant **n** points (vecteurs) de dimension **p** suivant :

$$X : \begin{array}{c|cccccc} & X^1 & \dots & X^j & \dots & X^p \\ X_1 & x_1^1 & \dots & x_1^j & \dots & x_1^p \\ \dots & \dots & \dots & \dots & \dots & \dots \\ X_i & x_i^1 & \dots & x_i^j & \dots & x_i^p \\ \dots & \dots & \dots & \dots & \dots & \dots \\ X_n & x_n^1 & \dots & x_n^j & \dots & x_n^p \end{array}$$

- A chaque point X_i est associée une classe connue à l'avance y_i ($y_i=f(X_i)$)
- Soit $X_q=\left[x_q^1, x_q^2, \dots, x_q^q, \dots, x_q^p\right]$ un point que l'on souhaite classer
- On calcule toutes les distances entre le point X_j et les **n** points de l'ensemble **S**
- On conserve les **k** points les plus proches de X_j
- La classe majoritaire dans l'ensemble de ces k points est attribuée à X_j

l'algorithme des **k** plus proches voisins

Exemple :



Si **k=3** le nouveau point sera associé à la classe des **cercles**
Si **k=5** le nouveau point sera associé à la classe des **carrés**

l'algorithme des **k** plus proches voisins

Règle de décision :

- Si y=f(x) est une variable de classe (classification), on fait voter les **k** plus proches voisins (classification)
- Ou, si y=f(x) est continue (régression) dans un espace métrique, on prend la moyenne :

$$\tilde{f}(x_q) \leftarrow \frac{1}{k}\sum_{j=1}^{k} f(x_j)$$

Avantages et désavantages

Advantages

- L'apprentissage est très rapide (simple stockage)
- Apprentissage de fonction complexe facile (pas de limite)
- Pas de perte d'information

Désavantages

- Lent à l'exploitation (surtout si le nombre d'apprentissage est grand)
- Le taux d'erreur grimpe en présence de caractéristiques non pertinentes

Necessité de s'appuyer sur une mesure de **distance** ou de **similarité**

Similarité et Dissimilarité

- **Similarité**
  - Une valeur numérique qui mesure combien deux objets se ressemblent.
  - D'autant plus grande que deux objets se ressemblent.
  - On la normalise souvent dans **[0,1]**
- **Dissimilarité**
  - Une valeur numérique qui mesure combien deux objets se différencient (**distance**)

  D'autant plus petit que les deux objets sont proches
  - Le minimum de dissimilarité est souvent **0**
  - La borne supérieure est généralement quelconque
- **Proximité** se réfère a une similarité ou a une dissimilarité

Métrique (distance)

Une mesure de dissimilarité *d(p, q)* entre deux points *p* et *q est une* Distance si elle satisfait :

1. **Définie Positive :**
   $d(p, q) \geq 0$ pour tout p et q et
   $d(p, q) = 0$ seulement si $p = q$.
2. **Symétrie:** $d(p, q) = d(q, p)$ pour tout p et q.
3. **Inégalité triangulaire :**
   $d(p, r) \leq d(p, q) + d(q, r)$ for all points p, q, and r.

Examples:

Distance Euclidienne

Distance de Minkowski

Distance de Mahalanobis

Propriétés des mesures de similarité

- Les similarités, ont également des propriétés

1. ***s(p, q) = 1*** (ou un maximum de similarité) seulement si ***p = q***.
2. ***s(p, q) = s(q, p)*** pour tout ***p*** et ***q***. (symmétrie)

où ***s(p, q)*** est la similarity entre points (objects), ***p*** et ***q***.

Distance ou non ?

$$p=(p_1, p_2, \ldots, p_d) \in R^d \text{ et } q=(q_1, q_2, \ldots, q_d) \in R^d$$

$d(p,q) = \max_{1<j<d}(p_j, q_j)$ — **Non définie positive**

$d(p,q) = \max_{1<j<d}(p_j - q_j)$ — **Non symétrique**

Inégalité Triangulaire non satisfaite

$d(p,q) = \sqrt{\sum_{1<j<d}(p_j - q_j)^2}$ — **Métrique**

$d(p,q) = \min_{1<j<d}(p_j - q_j)$

Quelques exemples de métriques

Euclidienne

$$d(p,q)=\sqrt{\sum_{1<j<d}(p_j-q_j)^2}$$

Minkowski

$$d(p,q)=\sqrt[r]{\sum_{1<j<d}(p_j-q_j)^r}$$

r=1
Distance City Block
Manhatan distance
norme-L1

r=2
Distance Euclidienne
norme-L2

Mahalanobis

$$d(p,q)=(p-q)\Sigma^{-1}(p-q)^T$$

Similarité entre vecteurs binaires

- Une situation classique : les objets, ***p*** et ***q***, ont des attributs binaires
- On calcule les similarités en utilisant les grandeurs
  M_{01} = le nombre d'attributs où **p** est **0** et **q** est **1**
  M_{10} = le nombre d'attributs où **p** est **1** et **q** est **0**
  M_{00} = le nombre d'attributs où **p** est **0** et **q** est **0**
  M_{11} = le nombre d'attributs où **p** est **1** et **q** est **1**
- "Matching" simple et coefficients de Jaccard
  SMC = nombre de matches / nombre d'attributs

  SMC = $(M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

  J = nombre de matches **valeur-1-à-valeur-1 / nombre de non-deux-zéros matches (Jaccard)**

  $J = (M_{11}) / (M_{01} + M_{10} + M_{11})$

SMC versus Jaccard: Exemple

p = 1 0 0 0 0 0 0 0 0 0
q = 0 0 0 0 0 0 1 0 0 1

$M_{01} = 2$
$M_{10} = 1$
$M_{00} = 7$
$M_{11} = 0$

$SMC = (M_{11} + M_{00})/(M_{01} + M_{10} + M_{11} + M_{00}) = (0+7) / (2+1+0+7) = 0.7$

$J = (M_{11}) / (M_{01} + M_{10} + M_{11}) = 0 / (2 + 1 + 0) = 0$

Similarité Cosinus

Si d_1 et d_2 sont deux vecteurs de Rd, alors

$\cos(d_1, d_2) = (d_1 \bullet d_2) / (||d_1|| \, ||d_2||)$, où :

- $\bullet$ indique le produit scalaire et

$||d||$ est la norme (L2) du vecteur d.

Exemple:

d_1 = 3 2 0 5 0 0 0 2 0 0
d_2 = 1 0 0 0 0 0 0 1 0 2 **$\cos(d_1, d_2) = .3150$**

$d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 =$ **5**

$||d_1|| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{1/2} = (42)^{1/2} =$ **6.481**

$||d_2|| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{1/2} = (6)^{1/2} =$ **2.245**

Pour aller plus loin ...

On peut (doit) s'appuyer sur les probabilités

Base des Probabilités

Une expérience aléatoire est une expérience qui peut conduire à plusieurs résultats. Le résultat courant est déterminé par un mécanisme obscure que l'on qualifie de chance, de risque, de providence, de hasard, etc.

Si l'on reproduit de manière répétée une expérience aléatoire, on produira des résultats différents (non déterminisme), i.e. des données aléatoires.

TERMINOLOGIE

A ***espace témoin***, *S*, est l'ensemble de tous les résultats possibles issus d'une expérience aléatoire.

Un ***événement*** est un sous ensemble de l'espace témoin, un ensemble de résultats. Un évènement, *A*, ***se produit*** lorsque le résultat courant de l'expérience aléatoire est un élément de *A*. Puisque *S* est l'ensemble de tous les résultats possible, *S* se produit ***toujours***. $A \subseteq S$

La ***probabilité*** d'un évènement *A* est un nombre entre *0* et *1* qui exprime (mesure, quantifie) la « chance » pour *A* de se produire. On la dénote ***P(A)***.

Base des Probabilités

- Quelle que soit l'interprétation que l'on donne aux probabilités, on suppose que les probabilités satisfont toujours ***Les définitions axiomatiques des probabilités***
- ***Soit S*** un espace témoin, A un évènement, et P(A) un nombre réel associé à chaque évènement A; On considérera que *P*(*A*) est la ***probabilité*** de l'évènement *A*, si les axiomes suivants sont satisfaits.

Axiome 1: $P(A) \geq 0$

Axiome 2: $P(S)=1$

Axiome 3: Si *A* et *B* sont mutuellement exclusif, alors

$$P(A \cup B) = P(A) + P(B)$$

Les théorèmes élémentaires

Quelques théorèmes simples découlent des définitions axiomatiques.

1. $P(\bar{A}) = 1 - P(A)$
2. $P(\emptyset) = 0$
3. Une séquence d'évènement $A_1, A_2, \ldots, A_n$ telle que pour tout $i \neq j$ $A_i \cap A_j = \emptyset$, est appelé mutuellement exclusive. Dans ce cas,
$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$$
4. Si $A \subset B$, alors $P(A) \leq P(B)$ et $P(B \cap \bar{A}) = P(B) - P(A)$

5. Pour tous évènements A et B,
$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$
6. $0 \leq P(A) \leq 1$.

Indépendance des évènements

Deux évènements *A* et *B* sont ***indépendants*** si l'occurrence ou la non-occurrence de *A* n'est pas affectée par l'occurrence ou la non-occurrence de *B* et *vice versa*.

Une collection d'événements $A_1, A_2, \ldots, A_n$ sont ***mutuellement indépendants*** ssi pour tous les sous ensembles de cette collection :

$$A_{i_1}, A_{i_2}, \ldots, A_{i_k}, k \leqslant n$$

$$P(\cap_{j=1}^{k} A_{i_k}) = \prod_{j=1}^{k} P(A_{i_k})$$

Par exemple, *A*, *B*, et *C* sont indépendants ssi

$$P(A \cap B) = P(A)P(B)$$
$$P(A \cap C) = P(A)P(C)$$
$$P(B \cap C) = P(B)P(C)$$
$$P(A \cap B \cap C) = P(A)P(B)P(C).$$

Probabilité Conditionnelle

- La probabilité d'un évènement mesure sa fréquence d'occurrence.
- Une ***probabilité conditionnelle*** prédit la fréquence d'occurrence d'un évènement étant données certaines conditions.
- Notation: P(*A*|*B*) = la probabilité conditionnelle que l'évènement A se produise, étant donné que l'évènement B s'est produit.

La "condition" contient
Une connaissance partielle.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Règle de Bayes

- Si $P(B) > 0$, alors

Probabilité *a priori*

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})}$$

Probabilité *a posteriori*

- Si $P(B) > 0$ et $A_1, A_2, \ldots, A_n$ constituent une partition de l'espace témoin Ω, alors :

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{n} P(B|A_j)P(A_j)}$$

Rapport de Vraisemblance (likelihood ratio)

$$O(H) = \frac{p(H)}{p(\bar{H})} = \frac{p(H)}{1 - p(H)}$$

Apprentissage Bayésien

Le classifieur naïf de Bayes

- On note :

$x=[x^1,x^2,\ldots,x^i,\ldots,x^d]$ un vecteur de descripteurs

$C=\{c_1,\cdots,c_q\}$ l'ensemble des classes possibles

S : ensemble fini de couples de la forme (x,c_k)

n : nombre d'observations de S

n_k : nombre d'éléments de S appartenant à la classe c_k

- La règle de classification de Bayes recommande de classer le vecteur x dans la classe c_k pour laquelle $P(c_k/x)$ est maximal.

Ce qui revient à maximiser : $\frac{p(x/c_k)\,p(c_k)}{p(x)}$

soit encore $p(x/c_k)\,p(c_k)$, $p(x)$ ne dépendant pas de c_k

Pour pouvoir appliquer la règle de Bayes il faut donc pouvoir estimer : $p(x/c_k)$ et $p(c_k)$

Le classifieur naïf de Bayes

- On estime $p(c_k)$ par : $\hat{p}(c_k)=\frac{n_k}{n}$
- et $p(x/c_k)$ par : $\hat{p}(x/c_k)=\prod_{i=1}^{d} p(x^i/c_k)$

Ce qui revient à considérer les attributs comme indépendants les uns des autres (caractère naïf du classifieur)

- La règle de classification de Bayes devient alors : classer le vecteur x dans la classe c_k qui maximise :

$$c_k=\underset{j=1..c}{ArgMax}\prod_{i=1}^{d} p(x^i/c_j)\,p(c_j)$$

On maximise ainsi la probabilité a posteriori (h_{MAP}: maximum a posteriori hypothesis)

Exemple

Est ce qu'un patient à la maladie de Lyme ?

1) Un patient effectue un examen qui consiste en un test qui s'avère positif.
2) Le test retourne un résultat positif correct dans seulement 98% des cas où la maladie est bien présente
3) Le test retourne un résultat négatif correct dans seulement 97% des cas où la maladie n'est pas présente
4) D'autre part, on estime a seulement 43 cas de maladie de Lyme pour 100000 habitants (France)

P(Lyme) =

P(non Lyme) =

P(+| Lyme) =

P(- | Lyme) =

P(+ | non Lyme) =

P(-| non Lyme) =

P(Lyme | +) =

Estimation des paramètres pour les textes

- **Modèle Binomial :**

$$\hat{P}(X_w = t | c_j) =$$ fraction des documents de la classe $\boldsymbol{c_j}$ dans lesquels le mot $\boldsymbol{w}$ apparaît (ou n'apparaît pas)

- **Modèle Multinomial :**

$$\hat{P}(X_i = w | c_j) =$$ Fréquence d'occurrence du mot **w** dans l'ensemble des documents De la classe $\boldsymbol{c_j}$

  - Revient à créer un mega-document pour la classe *j* en concatenant tous les documents de cette classe
  - Puis à évaluer la fréquence d'occurrence de *w* dans ce mega-document

Le classifieur naïf de Bayes version "binomiale"

- Exemple

$d = 5$

$S = S_1 \cup S_2$

$S_1 = \{01100, 11001, 10110, 10101, 10010\}$

$S_2 = \{01010, 11111, 11010, 11101, 10101\}$

$C = \{c_1, c_2\}$

- On demande de classer x=00111

$$\hat{p}(c_1) = \frac{n_1}{n} = \frac{5}{10} = \frac{1}{2}$$

$$\hat{p}(c_2) = \frac{n_2}{n} = \frac{5}{10} = \frac{1}{2}$$

Le classifieur naïf de Bayes

- On suppose les attributs indépendants

	c_1	c_2		c_1	c_2
$\hat{p}(x_1 = 0/c_j)$	1/5	1/5	$\hat{p}(x_1 = 1/c_j)$	4/5	4/5
$\hat{p}(x_2 = 0/c_j)$	3/5	1/5	$\hat{p}(x_2 = 1/c_j)$	2/5	4/5
$\hat{p}(x_3 = 0/c_j)$	2/5	2/5	$\hat{p}(x_3 = 1/c_j)$	3/5	3/5
$\hat{p}(x_4 = 0/c_j)$	3/5	2/5	$\hat{p}(x_4 = 1/c_j)$	2/5	3/5
$\hat{p}(x_5 = 0/c_j)$	3/5	2/5	$\hat{p}(x_5 = 1/c_j)$	2/5	3/5

$$\hat{p}(00111/c_1) = \frac{1}{5} \times \frac{3}{5} \times \frac{3}{5} \times \frac{2}{5} \times \frac{2}{5} = \frac{36}{5^5} = 11.5\ 10^{-3}$$

$$\hat{p}(00111/c_2) = \frac{1}{5} \times \frac{1}{5} \times \frac{3}{5} \times \frac{3}{5} \times \frac{3}{5} = \frac{27}{5^5} = 8.6\ 10^{-3}$$

- On classera donc x=00111 en classe c_1

Exercice classifieur naïf de Bayes

- Objectif : Classifier des phrases selon leur thème : la radio ou la télévision
- Échantillon :
  - Classe télévision :
    - Le programme TV n'est pas intéressant.
    - La TV m'ennuie.
    - Les enfants aiment la TV.
    - On reçoit la TV par onde radio.
  - Classe radio :
    - Il est intéressant d'écouter la radio.
    - Sur les ondes, les programmes pour enfants sont rares.
    - Les enfants vont écouter la radio; c'est rare.
- Vocabulaire : V={TV, programme, intéressant, enfants, radio, onde, écouter, rare}

- En utilisant un classifieur de Bayes, à quel thème serait associée la phrase : « J'ai vu la radio de mes poumons à la TV »

Exercice classifieur naïf de Bayes

–Codage des informations

–Échantillon :

TV	programme	intéressant	enfants	radio	onde	écouter	rare	
x1	x2	x3	x4	x5	x6	x7	x8	Classe
1	1	1	0	0	0	0	0	c1
1	0	0	0	0	0	0	0	c1
1	0	0	1	0	0	0	0	c1
1	0	0	0	1	1	0	0	c1
0	0	1	0	1	0	1	0	c2
0	1	0	1	0	1	0	1	c2
0	0	0	1	1	0	1	1	c2

Phrase à classer :
x=10001000

Exercice classifieur naïf de Bayes (version binomial)

$$\hat{p}(c_1)=\frac{4}{7} \qquad \hat{p}(c_2)=\frac{3}{7}$$

$$\hat{p}(x_1=0/c_1)=\frac{0}{4} \quad \hat{p}(x_1=0/c_2)=\frac{3}{3} \quad \hat{p}(x_1=1/c_1)=\frac{4}{4} \quad \hat{p}(x_1=1/c_2)=\frac{0}{3}$$
$$\hat{p}(x_2=0/c_1)=\frac{3}{4} \quad \hat{p}(x_2=0/c_2)=\frac{2}{3} \quad \hat{p}(x_2=1/c_1)=\frac{1}{4} \quad \hat{p}(x_2=1/c_2)=\frac{1}{3}$$
$$\hat{p}(x_3=0/c_1)=\frac{3}{4} \quad \hat{p}(x_3=0/c_2)=\frac{2}{3} \quad \hat{p}(x_3=1/c_1)=\frac{1}{4} \quad \hat{p}(x_3=1/c_2)=\frac{1}{3}$$
$$\hat{p}(x_4=0/c_1)=\frac{3}{4} \quad \hat{p}(x_4=0/c_2)=\frac{1}{3} \quad \hat{p}(x_4=1/c_1)=\frac{1}{4} \quad \hat{p}(x_4=1/c_2)=\frac{2}{3}$$
$$\hat{p}(x_5=0/c_1)=\frac{3}{4} \quad \hat{p}(x_5=0/c_2)=\frac{1}{3} \quad \hat{p}(x_5=1/c_1)=\frac{1}{4} \quad \hat{p}(x_5=1/c_2)=\frac{2}{3}$$
$$\hat{p}(x_6=0/c_1)=\frac{3}{4} \quad \hat{p}(x_6=0/c_2)=\frac{2}{3} \quad \hat{p}(x_6=1/c_1)=\frac{1}{4} \quad \hat{p}(x_6=1/c_2)=\frac{1}{3}$$
$$\hat{p}(x_7=0/c_1)=\frac{4}{4} \quad \hat{p}(x_7=0/c_2)=\frac{1}{3} \quad \hat{p}(x_7=1/c_1)=\frac{0}{4} \quad \hat{p}(x_7=1/c_2)=\frac{2}{3}$$
$$\hat{p}(x_8=0/c_1)=\frac{4}{4} \quad \hat{p}(x_8=0/c_2)=\frac{1}{3} \quad \hat{p}(x_8=1/c_1)=\frac{0}{4} \quad \hat{p}(x_8=1/c_2)=\frac{2}{3}$$

$$\hat{p}(c_1)\times\hat{p}(10001000/c_1)=\frac{4}{7}\times\frac{4}{4}\times\frac{3}{4}\times\frac{3}{4}\times\frac{1}{4}\times\frac{3}{4}\times\frac{4}{4}\times\frac{4}{4}=\frac{20736}{458752}=0{,}0452$$
$$\hat{p}(c_2)\times\hat{p}(10001000/c_2)=\frac{3}{7}\times\frac{0}{3}\times\frac{2}{3}\times\frac{2}{3}\times\frac{1}{3}\times\frac{2}{3}\times\frac{2}{3}\times\frac{1}{3}\times\frac{1}{3}=\frac{0}{45927}=0$$

Exercice classifieur naïf de Bayes

Estimateur de Laplace $\hat{p}(x_i/c_k)=\frac{1+\text{Nombre d'occurences de } x_i \text{ de } V \text{ dans l'ensemble des textes de classe } k}{Card(V)+\text{Nombre total d'occurences de mots dans l'ensemble des textes de classe } k}$

$$\hat{p}(c_1)=\frac{4}{7} \qquad \hat{p}(c_2)=\frac{3}{7}$$

$$\hat{p}(x_1=1/c_1)=\frac{1+4}{8+9} \quad \hat{p}(x_1=1/c_2)=\frac{1+0}{8+11}$$
$$\hat{p}(x_2=1/c_1)=\frac{1+1}{8+9} \quad \hat{p}(x_2=1/c_2)=\frac{1+1}{8+11}$$
$$\hat{p}(x_3=1/c_1)=\frac{1+1}{8+9} \quad \hat{p}(x_3=1/c_2)=\frac{1+1}{8+11}$$
$$\hat{p}(x_4=1/c_1)=\frac{1+1}{8+9} \quad \hat{p}(x_4=1/c_2)=\frac{1+2}{8+11}$$
$$\hat{p}(x_5=1/c_1)=\frac{1+1}{8+9} \quad \hat{p}(x_5=1/c_2)=\frac{1+2}{8+11}$$
$$\hat{p}(x_6=1/c_1)=\frac{1+1}{8+9} \quad \hat{p}(x_6=1/c_2)=\frac{1+1}{8+11}$$
$$\hat{p}(x_7=1/c_1)=\frac{1+0}{8+9} \quad \hat{p}(x_7=1/c_2)=\frac{1+2}{8+11}$$
$$\hat{p}(x_8=1/c_1)=\frac{1+0}{8+9} \quad \hat{p}(x_8=1/c_2)=\frac{1+2}{8+11}$$

Version “multinomial” et Estimateur de Laplace

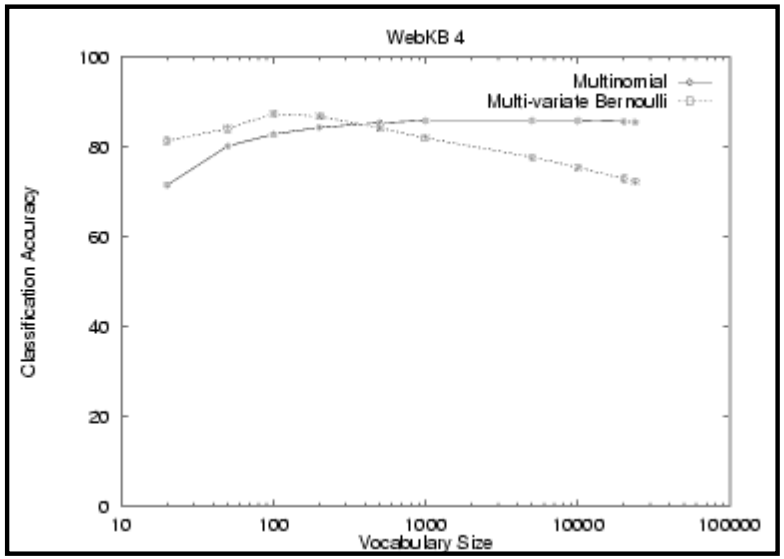
$$\hat{p}(c_1)\times\hat{p}(10001000/c_1)=\frac{4}{7}\times\frac{5}{17}\times\frac{2}{17}\times\frac{3}{17}=\frac{40}{2023}=0{,}020$$
$$\hat{p}(c_2)\times\hat{p}(10001000/c_2)=\frac{3}{7}\times\frac{1}{19}\times\frac{3}{19}=\frac{9}{2527}=0{,}0035$$

Experimentation : Multinomial vs Binomial

- [1] ont effectué un test pour determiner quel modèle est le meilleur
- Tache : Classer une page web d’un site universitaire dans l’une des catégories {étudiant, prof, autre}
- Apprentissage sur ~5000 pages web “étiquetées à la main”
  - Cornell, Washington, U.Texas, Wisconsin

[1] Andrew McCallum and Kamal Nigam, A Comparison of Event Models for Naive Bayes Text Classification, aaaiws98, 1998.

Multinomial vs. Binomial



Conclusions sur l'expérience

- Le modèle multinomial est meilleur en grande dimension (cas des données textuelles)
- Pour le modèle binomial, il est très important d'effectuer une sélection préalable des descripteurs (ne retenir que les descripteurs discriminants)
- Beaucoup d'autres expériences abondent en ce sens

Classifieur de Bayésien Naïf en pratique

En grande dimension, $c_k = \underset{j=1..c}{ArgMax} \prod_{i=1}^{d} p\left(x^i / c_j\right) p\left(c_j\right)$ tend vers 0

La solution consiste à prendre le logarithme

$$c_k = \underset{j=1..c}{ArgMax} \prod_{i=1}^{d} p\left(x^i / c_j\right) p\left(c_j\right) = \underset{j=1..c}{ArgMax} \log\left\{\prod_{i=1}^{d} p\left(x^i / c_j\right) p\left(c_j\right)\right\}$$

$$c_k = \underset{j=1..c}{ArgMax} \sum_{i=1}^{d} \log\left(p\left(x^i / c_j\right)\right) + \log\left(p\left(c_j\right)\right)$$

Bayésien naïf : conclusions

- Les résultats issus d'une classification de type Bayésien naïf (la classe qui maximise la probabilité a posteriori) sont souvent bons à très bons, mais d'autres approches (SVM, NN) font mieux en général.
- Cependant, du fait que l'hypothèse d'indépendance conditionnelle sur les caractéristiques n'est pas vérifiée en générale, les estimations sur les probabilité a posteriori ne sont pas correctes.

  Ces probabilités sont souvent très proches de 0 ou de 1.

Quelques bons points sur le modèle BN

- Théoriquement **optimal si** l'hypothèse d'indépendance est vérifiée
- **Rapide**
- Fait le **tri** entre **descripteurs** robustes et non discriminants (à vérifier)
- **Peu de méta paramètre** (valeur par défaut ok)
- Très bon dans les situations caractérisées par un grand nombre de descripteurs d'importance similaire
- Méthode très utile et efficace pour effectuer des petits tests (utilisé en tant que "**baseline**")