

Identification de structures de dépendance statistique

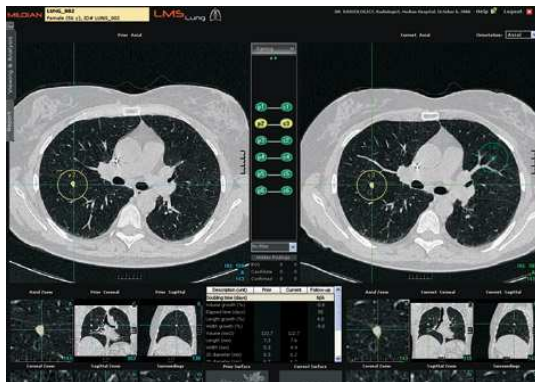
João RENDAS, ERIC THIERRY, GILLES MENEZ

6 décembre 2007

Motivation

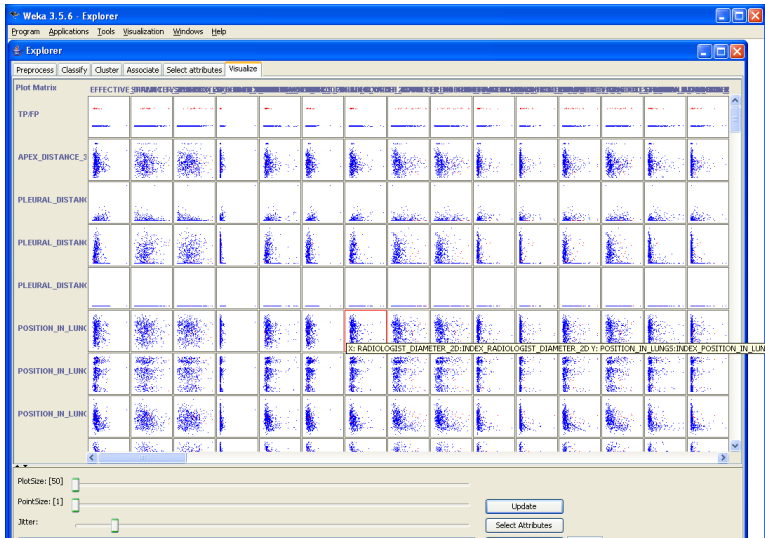
Collaboration avec Median Tech (Sophia Antipolis)

- ▶ Système d'aide au diagnostique du cancer du poumon
- ▶ analyse de clichés radiologiques
- ▶ caractérisation de régions suspectes par un ensemble d'attributs ($\vec{X}_k \in \mathfrak{R}^{d=104}$).



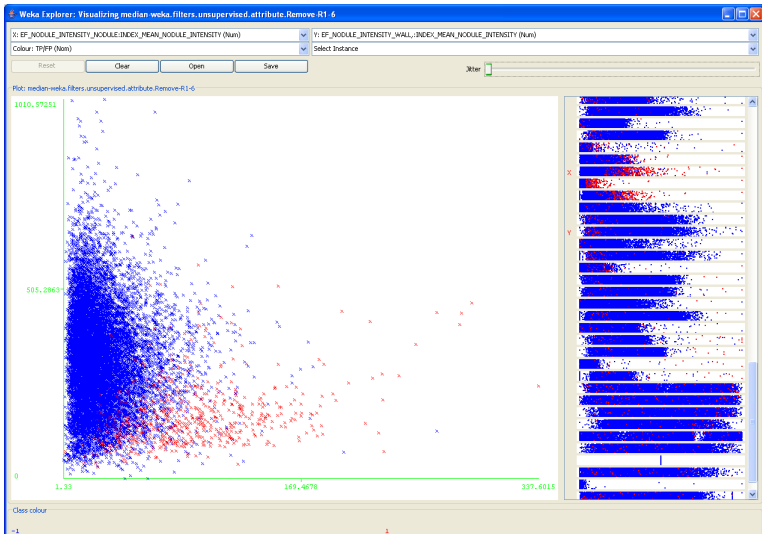
Données fournies par Median

Deux classes : Sains et Malades



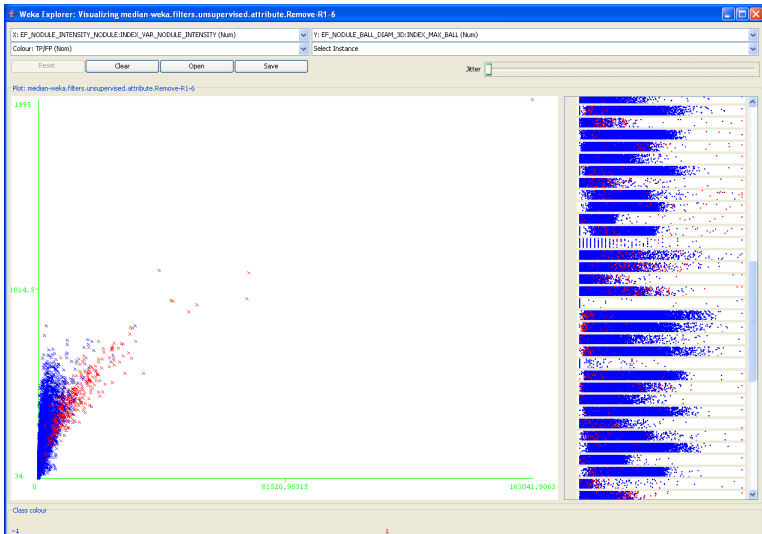
Données fournies par Median

Deux classes : Sains et Malades



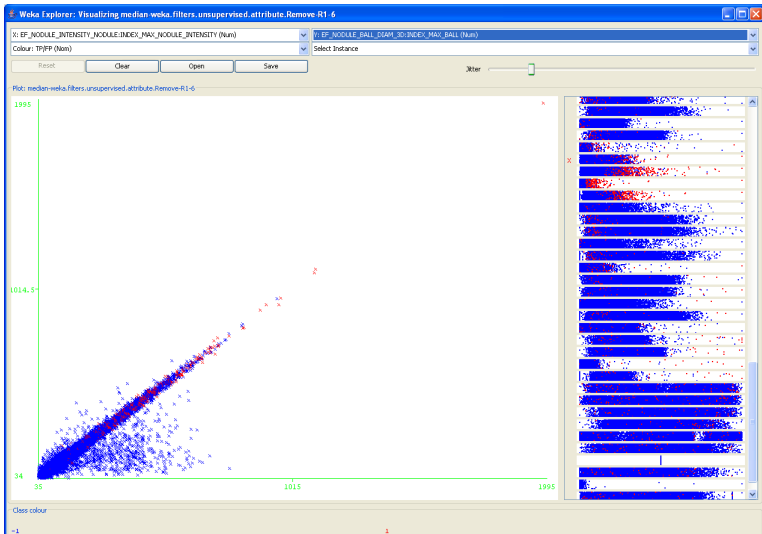
Données fournies par Median

Deux classes : Sains et Malades



Données fournies par Median

Deux classes : Sains et Malades



Nos objectifs

- ▶ Faire une analyse des données utilisées actuellement (*redondance, suffisance,...*)
- ▶ Comprendre les *limites de performance* qui peuvent être attendus
- ▶ Permettre apprentissage *incrémentale*
- ▶ Faire un classifieur plus performant que l'actuel (SVM) ...
- ▶ Travailler sur l'apprentissage **statistique** (!!)

Approches au problème d'apprentissage de classifieurs

$$\{\mathcal{A}_{\text{sains}}, \mathcal{A}_{\text{malades}}\} \longrightarrow ?$$

1. Apprendre la **fonction de décision** (e.g. SVM - Support Vector Machine)

Apprendre les régions de décision

$$? \equiv \mathcal{R}_{\text{sains}}, \mathcal{R}_{\text{malades}}$$

$$\vec{x}_k \in \mathcal{R}_{\text{sains}} \implies x_k \in \text{sains} \quad \vec{x}_k \in \mathcal{R}_{\text{malades}} \implies x_k \in \text{malades}$$

2. Apprendre les **caractéristiques** (distributions) de **chaque classe**

Apprendre les densités de chaque ensemble de données

$$? \equiv p_{\text{sains}}(\vec{x}), p_{\text{malades}}(\vec{x})$$

$$\frac{p_{\text{sains}}(\vec{x}_k)}{p_{\text{malades}}(\vec{x}_k)} > \gamma \implies x_k \in \text{sains} \quad \frac{p_{\text{sains}}(\vec{x}_k)}{p_{\text{malades}}(\vec{x}_k)} < \gamma \implies x_k \in \text{malades}$$

Avantages de la modélisation des classes

(approche 2)

- ▶ outlier detection
- ▶ Interprétabilité
- ▶ analyse de performance
- ▶ apprentissage incrémentale (e.g. augmenter le nombre de classes considérées)
- ▶ pas de contraintes imposées à la géométrie des régions de décision
pas de définition *a priori* des variables de décision (comme pour les méthodes à noyaux)
- ▶ mesure de la confiance dans la classification

Le problème

$$\mathcal{A} = \{\vec{x}_1 \dots \vec{x}_n\}, x_i \in \mathbb{R}^d \implies p(\vec{x})$$

Difficultés

- ▶ $d \gg 1$
- ▶ variables **continues**

Identification de densités dans des espaces de grandes dimensions

- ▶ méthodes à noyaux (vitesse de convergence (minimax) très lente)
- ▶ modèles de mélange(semi-paramétriques, même problème)
- ▶ “projection poursuit”
- ▶ log-spline models
- ▶ vraisemblance pénalisée
- ▶ “sparsity arguments” (dans les variables, et/ou dans des fonctions paramétriques des variables)
- ▶ *kd*-trees
- ▶ imposer une structure de (in)dépendance statistique (factorisation)

Objectif

“*Divide and conquer*”

$$p(\vec{x}) = p(x_1, \dots, x_d) = \prod_{i=1}^q p(x_{\pi_1^i}, \dots, x_{\pi_{\ell_i}^i} | V_i)$$

où

- ▶ $\{\pi^1, \dots, \pi^q\}$ est une partition de $\{1, \dots, d\}$
- ▶ $V_i \subset \bigcup_{j \neq i} \pi^j$

Si $\ell_i \ll d$ nous devons résoudre (plusieurs !!) problèmes plus simples, au lieu d'un seul problème très complexe.

Possibilité d'identifier des composantes *non-informatives*

Naïf Bayes

Cas extrême :

$$p(\vec{x}) = \prod_{i=1}^d p(x_i)$$

$$V_i = \emptyset, \quad \pi^i = x_i.$$

Admet que toutes les observations sont *statistiquement indépendantes*

Modélisation de la dépendance statistique

- ▶ Réseaux Bayésiens
en général $p(x_{\pi_i} | V_i)$ est modélisée paramétriquement
structure de dépendance (cliques) *imposée*
algorithmes *gloutons*

Indépendance statistique

Mesure universelle d'indépendance : l'*information mutuelle*

$$I(X; Y) = \int \int p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy = D(p(x, y) || p(x)p(y))$$

$I(X; Y) = 0 \Leftrightarrow X$ statistiquement indépendant de Y

Invariante par rapport à des transformations 1-to-1.

Tests d'indépendance

$\{\vec{x}_1, l, \dots, \vec{x}_{n_x}\}, \{\vec{y}_1, l, \dots, \vec{y}_{n_y}\} \longrightarrow \vec{X}$ indépendant de \vec{Y} ?

$\{\vec{x}_1, l, \dots, \vec{x}_{n_x}\}, \{\vec{y}_1, l, \dots, \vec{y}_{n_y}\} \longrightarrow \hat{l}(X; Y)$

- ▶ Variables discrètes
algorithmes rapides basées sur des tests locaux de
équi-distribution
- ▶ Variables Continues (Correlation de Pearson, Spearman rank
correlation, Kendall's tau, Hoeffding...)
- ▶ Indépendance *versus* indépendance *conditionnelle*
coefficient de corrélation partiel - corrélation des erreurs de
modèles de régression
test basé sur les propriétés d'uniformité et indépendance des
copula (partielles) $u(x, y) = \Pr\{Y \leq y | X = x\}$

Définition

Information mutuelle :

$$I(X, Y) = H(X) + H(Y) - H(X, Y) \geq 0$$

avec $H(X)$ entropie de la variable X et $H(X, Y)$ entropie conjointe du couple (X, Y)

Si X est indépendante de Y alors $I(X, Y) = 0$

Loi normale

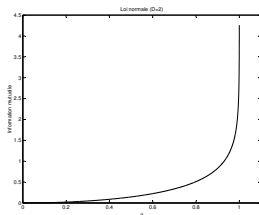
Soit $X \in \mathbb{R}^d$ $X \sim \mathcal{N}(\mathbf{m}, \Sigma)$ alors

$$H(\mathbf{X}) = \frac{d}{2}(1 + \log(2\pi)) + 0.5 \log(|\Sigma|)$$

Cas $d = 2$

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$$

$$I(X, Y) = 0.5 \log(1 - \rho^2)$$



Estimation d'entropie

Estimer $I \implies$ estimer H :

$$\hat{I}(X, Y) = \hat{H}(X) + \hat{H}(Y) - \hat{H}(X, Y)$$

Estimation de l'entropie

Observations : $\vec{x}_1, \dots, \vec{x}_n$ et $\vec{y}_1, \dots, \vec{y}_n$

Modèle : $f_X(x) \in \mathcal{M}$

- ▶ Estimation paramétrique

$$\mathcal{M} = \{f(x, \theta), \theta \in \Omega\}$$

- ▶ Estimation non-paramétrique

$$\mathcal{M} = \{f(x), \text{une loi suffisamment régulière}\}$$

Estimation paramétrique

$$\mathcal{M} = \{f(\mathbf{x}, \theta), \theta \in \Omega\}$$

Il faut estimer les paramètres θ de la loi

Exemple :

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{|2\pi\Sigma|^{1/2}} \exp(-0.5(\mathbf{x})^t \Sigma^{-1}(\mathbf{x}))$$

alors l'UMVUE de l'entropie est

$$\hat{H}(\mathbf{X}) = \frac{d}{2} \log(e\pi) + \frac{1}{2} \log |V| - 0.5 \sum_{i=1}^d \psi\left(\frac{n+1-i}{2}\right)$$

avec $V = \mathbf{X}\mathbf{X}'$ et $\psi = \frac{d[\log(\Gamma(u))]}{du} \Big|_{u=z}$ la fonction digamma .

Estimation non-paramétrique à noyaux

Estimation non-paramétrique de la densité

$$\hat{f}(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right)$$

où $K(\cdot)$ est un noyau tel que $\int K = 1$ et possédant propriétés de régularité.

Drawbacks :

- ▶ complexité quadratique (si utilisé naïvement) dans le nombre d'échantillons
- ▶ *curse of dimensionality* : la complexité est exponentielle dans la dimension de l'espace

Estimation non-paramétrique à noyaux

Plusieurs estimateurs "plug-in" :

$$\blacktriangleright \hat{H} = - \int \hat{f}(\mathbf{x}) \log \hat{f}(\mathbf{x}) d\mathbf{x}$$

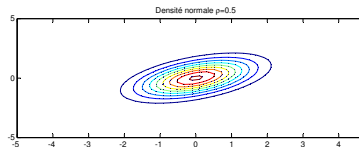
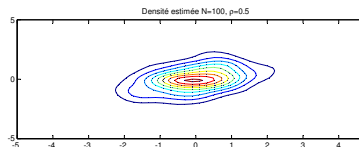
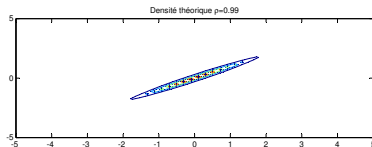
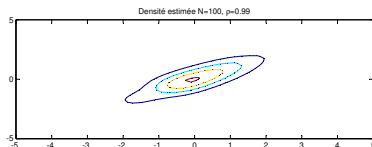
$$\blacktriangleright \hat{H} = -\frac{1}{n} \sum_{i=1}^n \log \hat{f}_{-i}(\mathbf{x}_i)$$

$$\text{avec } \hat{f}_{-i}(\mathbf{x}) = \frac{1}{(n-1)h^d} \sum_{j=1, j \neq i}^n K\left(\frac{\mathbf{x}-\mathbf{x}_j}{h}\right)$$

Choix de h

La forme du *noyau*, $K(\cdot)$ n'est pas primordiale.

Le choix du *paramètre d'échelle* h est *critique* en particulier pour l'estimation de la densité conjointe



En général : **cross-validation** \implies complexité croît linéairement avec le nombre de valeurs de h considérés

Estimation de la densité

Critère :

$$MISE(h) = \int E[(\hat{f}_h(\mathbf{x}) - f(\mathbf{x}))^2] = \int \text{Var}[\hat{f}_h(\mathbf{x})] d\mathbf{x} + \int \text{Bias}^2[\hat{f}_h(\mathbf{x})] d\mathbf{x}$$

avec $\text{Bias}[\hat{f}_h(\mathbf{x})] = (K_h * f - f)$

Pour que l'estimateur soit *consistent* il faut que

$$h \rightarrow 0 \text{ et } hn \rightarrow \infty \text{ lorsque } n \rightarrow \infty$$

Le h minimisant le critère asymptotique lorsque $X \sim$ normale est

$$h = \hat{\sigma}_X n^{-1/(d+4)}$$

Mais ce choix est-il adapté à l'estimation de l'entropie ? **NON**

Loi Normale Bidimensionnelle

Pour une loi normale bidimensionnelle et un noyau gaussien on peut montrer que le paramètre h optimum est

$$h_i^* = \sigma_i (1 - \rho^2)^{5/12} (1 + \rho^2/2)^{-1/6} n^{-1/6}$$

et

$$AMISE^* = \frac{3}{8\pi} (\sigma_1 \sigma_2)^{-1} (1 - \rho^2)^{-5/6} (1 + \rho^2/2)^{1/3} n^{-2/3}$$

donc

$$AMISE^* \rightarrow \infty \text{ lorsque } \rho \rightarrow 1$$

C'est gênant !!

Choix de h pour l'estimation d'entropie

Une solution est donnée par Hall et Morton 1993 pour le choix de h . Ils proposent de choisir le h qui minimise

$$\hat{H}_h = \frac{1}{n} \sum_{i=1}^n -\log \hat{f}_{-i}(\mathbf{x}_i)$$

Ils montrent que théoriquement

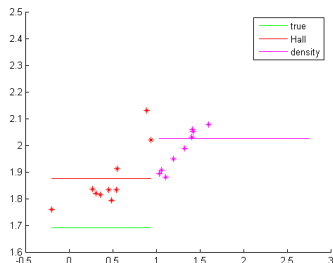
$$\hat{H}_h = \bar{H} + (C_1(nh^d)^{-1+(d/\alpha)} + C_2h^4) + o_P(C_1(nh^d)^{-1+(d/\alpha)} + C_2h^4)$$

α étant un paramètre lié à la vitesse de décroissance de la queue de la loi f et $\bar{H} = -\frac{1}{n} \sum_{i=1}^n \log f(x_i)$.

Cette étude montre que les queues du noyau $K(\cdot)$ doivent être au moins aussi lourdes que celles de la loi $f(\cdot)$

Résultats de simulation

Résultats pour 10 réalisations de $x_i = y_{i1}^2 + y_{i2}^2, y_{i,j} \stackrel{iid}{\sim} \mathcal{N}(0, 1), i = 1, \dots, 200$



— valeur exacte de l'entropie

rouge : estimées avec la valeur de h qui minimize \hat{H}_h

magenta : valeurs de l'entropie choisis par la règle optimale pour des variables normales

Note : Des test sur des variables avec d'autres distributions confirment la suprémacie de cet estimateur.

Histogrammes de bins variables

Pour des densités de support borné

$$x \in [a, b]$$

nous avons considéré le modèle (histogramme avec “bins” variables)

$$\hat{p}(x) = \hat{\nu}_i, \in [\hat{c}_{i-1}, \hat{c}_i], i = 1, \dots, \hat{M}$$

où $c_{i+1} > c_i$, $\hat{c}_0 = a$ et $\hat{c}_{\hat{M}} = b$ (connus).

Un estimateur de l'entropie est obtenu par l'approche “plug-in” :

$$\hat{H} = H(\hat{p}(\cdot))$$

MDL

Maximum de vraisemblance pour les paramètres $\{M, \{c_i\}, \{\nu_i\}\}$

$$\Rightarrow \hat{M} = m \leq n$$

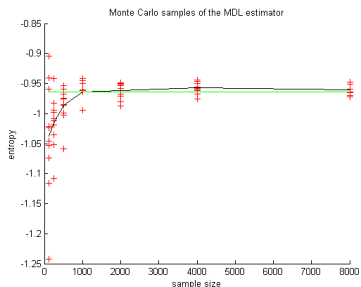
et une *densité discrète* dans les m valeurs distincts des données.
Nous avons considéré la pénalisation de la fonction de vraisemblance par la version asymptotique du MDL

$$\mathcal{C}(X, M) = -\frac{1}{n} \log \hat{p}_M + \frac{k(M)}{2} \log n$$

où $k(M) = 2(M - 1)$ est le nombre de paramètres “libres” du modèle.

Résultats

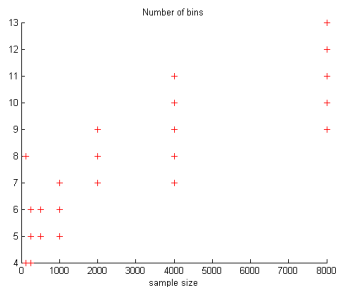
Variable normale tronquée à l'intérieur de l'intervalle unitaire ($a \geq 0, b \leq 1$)



Algorithme de *programmation dynamique* (on peut démontrer que les points \hat{c}_i doivent nécessairement coïncider avec une des observations) Version sous-optimale : si $n \gg 1$ nous considérons un sous-échantillon des données (ici de taille 500) comme candidats pour les limites des intervalles.

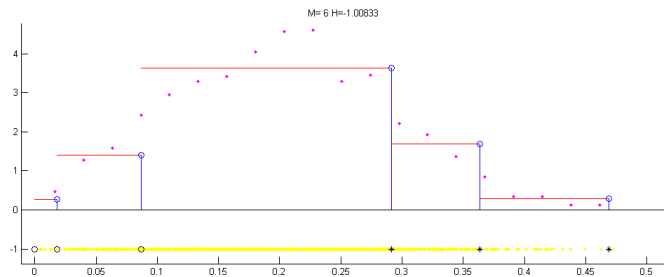
Résultats

Evolution du nombre d'intervalles (M)



Résultats

Densité estimée (exemple)



MDL-NML

Note :

une meilleure performance doit être obtenue avec la pénalité MDL-NML (Normalized Maximum Likelihood)

$$\mathcal{P}(M) = \log \int_X \hat{p}_M(X) dX$$