Classification par plus proches voisins, Optimalité sous hypothèse de marge

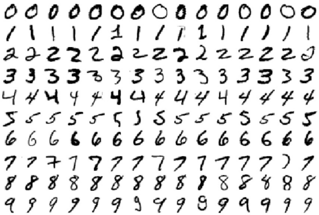
S. Gadat

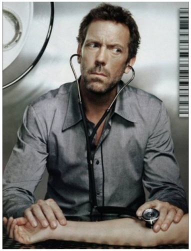
Institut de Mathématiques de Toulouse
Université Paul Sabatier
Issu de travaux avec T. Klein, K-A. Lê Cao, C. Marteau.

Toulouse, 22 Novembre 2013

I - 1 Motivations concrètes - Classification d'image - Diagnostic Médical

Problème : Classification automatique de chiffres manuscripts, base Mnist US Postals





Source : *Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition." Proc. of the IEEE, 86(11) :2278-2324, Nov. 1998.*

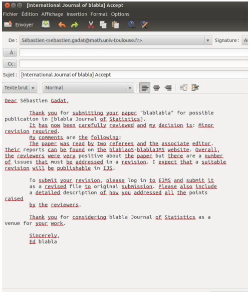
Nouvelle saisie :  Prédire la classe automatiquement ? Nouveau diagnostic ?
Approche statistique :

- Collecter les données digitales (24×24 pixels) $\hookrightarrow$ codage sur $\{0, \ldots, 255\}^{24\times 24}$.
- Réalisations de tests médicaux et saisie d'informations personnelles ($\hat{\text{A}}\textsc{ge}$, $\textsc{Sexe}$, $\textsc{Poids}$, $\ldots$,

Stocker les n données de la base d'apprentissage : $\mathcal{D}_n := (X_1, Y_1), \ldots, (X_n, Y_n)$,
Calculer un prédicteur à partir de $\mathcal{D}_n$, noté Φ_n (un chiffre / "Sain" vs "Malade").
On observe un nouvel X, comportement de $\Phi_n(X)$ avec beaucoup de données ?

I - 1 Motivations concrètes - Détection de Spam

Problème : Détection de Spam, Hp Database



Nouvelle saisie : Prédire la classe automatiquement ?

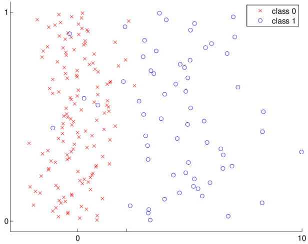
Approche statistique :

- Décrire les messages par simple comptage de p mots typiques.
- STATISTICS, PROBABILITY, $\$, !, \ldots$
- Stocker les n données de $\mathbb{Z}^p \times \{0,1\}$: $\mathcal{D}_n := (X_1, Y_1), \ldots, (X_n, Y_n)$.
- Calculer un prédicteur / algorithme/ classifieur à partir de $\mathcal{D}_n$, noté Φ_n pour décider "Spam" vs "non Spam".

On observe un nouvel X, comportement de $\Phi_n(X)$ avec beaucoup de données ?

I - 2 Cadre de la classification (binaire) supervisée

- On observe des données étiquettées d'un ensemble d'apprentissage $\mathcal{D}_n$. Ces données appartiennent à $\mathbb{R}^d \times \{0, 1\}$.



- On calcule un algorithme Φ_n à partir de $\mathcal{D}_n$ (algorithme 'off-line').
- On cherche à quantifier l'efficacité d'un algorithme de prédiction *via* une fonction de coût ℓ

Autres sources d'applications

- Traitement du signal, de l'image
- Classification de documents
- Bio-informatique
- Credit scoring

I - 3 Différentes formulations (pas totalement équivalentes)

Modèle de classification (simple)

- On observe n réalisations i.i.d. $(X_1, Y_1), \dots (X_n, Y_n)$.
- Les positions X et les labels Y sont décrits par une loi jointe : $(X, Y) \sim P$.
- X est un vecteur aléatoire d'un compact $K \subset \mathbb{R}^d$ et $Y \in \{0, 1\}$.
- La loi marginale P_Y est une Bernoulli (équilibrée $\mathcal{B}(1/2)$).
- On suppose que les lois conditionnelles sont a.c. w.r.t. $d\lambda_K(.)$, f (resp. g) est la densité de $X|Y = 0$ (resp. $X|Y = 1$).

Modèle d'Analyse discriminante (plus dur)

- On observe $n/2$ réalisations de loi de densité f : $(X_1, \dots, X_{n/2})$ et $(Y_1, \dots, Y_{n/2}) = (0, \dots, 0)$.
- On observe $n/2$ réalisations de loi de densité g : $(X_{n/2+1}, \dots, X_n)$ et $(Y_1, \dots, Y_{n/2}) = (1, \dots, 1)$.

Dans les deux cas, on définit la fonction de régression

$$\eta(x) = \frac{g(x)}{f(x) + g(x)} = \mathbb{P}(Y = 1|X).$$

I - 3 Différentes formulations (pas totalement équivalentes)

Modèle de classification (simple)

- Un algorithme Φ est une fonction de X et « prédit » 0 ou 1.
- On mesure le risque d'un algorithme Φ au travers de son *risque* définit par

$$\mathcal{R}(\Phi) = \mathbb{P}\left[\Phi(X) \neq Y\right] = \mathbb{E}_P\left[\mathbf{1}_{\Phi(X)\neq Y}\right]$$

- Il existe un algorithme optimal, le *classifieur bayésien*

$$\Phi_{Bayes}(X) := \mathbf{1}_{\eta(X)\geq 1/2}.$$

Modèle d'Analyse discriminante (plus dur)

- Une région de prédiction G permet de décider 1 si $X \in G$ (0 sinon).
- On mesure le risque basé sur la région G par

$$\mathcal{R}(G) = \frac{1}{2}\left[\int_G f + \int_{K\setminus G} g\right].$$

- Il existe une région optimale, la *région de Bayes*

$$G_{Bayes} := \{x \in K : g(x) \geq f(x)\}$$

Théorème (Györfi '78, Mammen & Tsybakov '99)

Dans tous les cas, l'excès de risque s'écrit

$$\mathcal{R}(\Phi) - \mathcal{R}(\Phi_{Bayes}) = \mathbb{E}_{P_X}\left[|2\eta(X) - 1|\,\mathbf{1}_{\Phi(X)\neq\Phi_{Bayes}(X)}\right]$$

Comme P est inconnue en pratique, Φ_{Bayes} n'est pas calculable.

I - 4 Un algorithme de classification classique

On considère un espace K **muni d'une distance** $\|.\|$ et pour $x \in K$, on ordonne les n observations par ordre croissant des distances à x :

$$\|X_{(1)}(x) - x\| \leq \|X_{(2)}(x) - x\| \leq \ldots \leq \|X_{(n)}(x) - x\|.$$

$X_{(m)}(x)$ est le m-ième voisin de x dans $\mathcal{D}_n$ et $Y_{(m)}(x)$ est le label correspondant.

$$\Phi_{n,k}(x) := \begin{cases} 1 & si \quad \dfrac{1}{k}\displaystyle\sum_{j=1}^{k} Y_{(j)}(x) > \dfrac{1}{2}, \\ 0 & sinon. \end{cases} \tag{1}$$

Un dessin vaut parfois mieux qu'un long discours ...

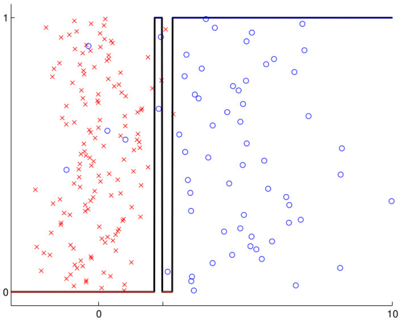


Fig. Gauche : décision par 3-NN

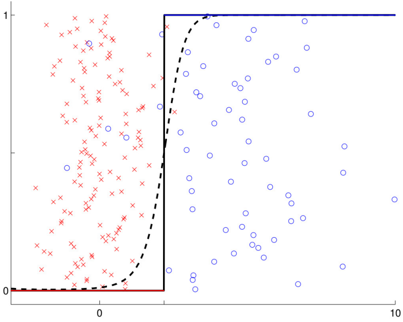
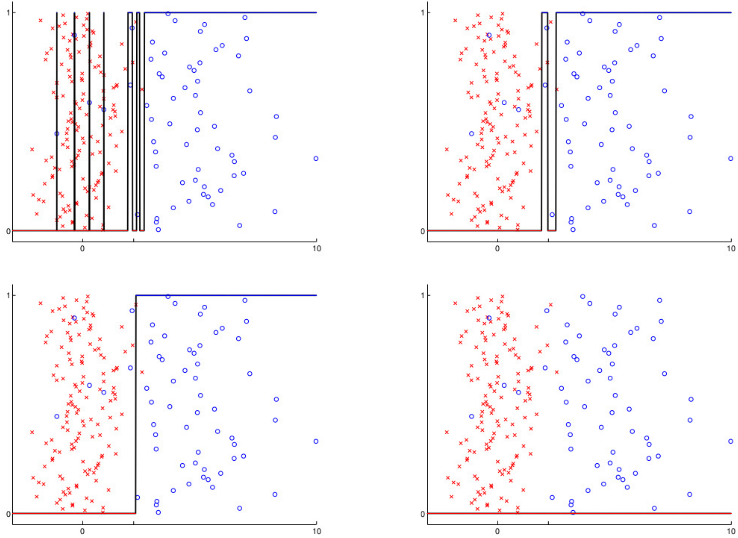


Fig. Droite : classifier Φ_{Bayes}

I - 4 Un algorithme de classification classique

Quelle est l'influence de k pour l'algorithme des k plus proches voisins ?



$k \in \{1, 3, 20, 200\}$, $k = 1 \hookrightarrow$ overfitting (pure variance), $k = 200 \hookrightarrow$ underfitting (pur biais).

II - 1 Hypothèses de travail

- Le statisticien a besoin de poser un cadre de travail . . .

Théorème (No Free Lunch Theorem, Wolpert 1996)

Si le support des lois K est infini, alors pour tout algorithme de classification Φ et tout entier $n \geq 1$:

$$\sup_{P \in \mathfrak{M}(K \otimes \{0;1\})} \mathbb{E}\left[\mathcal{R}(\Phi) - \mathcal{R}(\Phi_{Bayes})\right] \geq 1/2.$$

- On se placera soit dans le cadre de classification, soit dans celui de l'analyse discriminante.
- Il y a 2 sources d'aléa dans l'étude à mener : l'aléa d'échantillonage sur $\mathcal{D}_n$ et l'aléa de prédiction de Y en fonction de X. On notera $\mathbb{E}^X$ et $\mathbb{P}^X$ pour spécifier l'aléa sur l'échantillon lorsque le point de prédiction X est fixé.
- Les slides suivants se placent sous des hypothèses sur les lois des observations.

II - 2 Strong Density Assumption (H_{SDA}) et régularité de η

- **On fait l'hypothèse que la distribution de X est à support compact (noté K).**
  - Pour la classification, P_X a une densité μ par rapport à la mesure de Lebesgue et

$$\forall x \in K \quad \exists (\mu_{min}, \mu_{max}) \in \mathbb{R}_+^2 \qquad \mu_{min} \leq \mu(x) \leq \mu_{max}.$$

  - Pour l'analyse discriminante, f et g ont un support K_f et K_g compact, $K = K_f \cup K_g$ et

$$\forall x \in K \quad \exists (\mu_{min}, \mu_{max}) \in \mathbb{R}_+^2 \qquad \mu_{min} \leq f(x) + g(x) \leq \mu_{max}.$$

- **On fait l'hypothèse que le support K de la distribution de X est (c_0, r_0)-régulier :**

$$\forall x \in K \quad \forall r \leq r_0 \qquad \lambda\left(K \cap B(x,r)\right) \geq c_0 \lambda B(x,r).$$

  Cette hypothèse traduit la régularité de la frontière de K (il ne peut pas être fractal). **Ces deux hypothèses seront résumées par la notation H_{SDA}.**

- Enfin, on suppose que η **est L-Lispchitz** pour $\|.\|$:

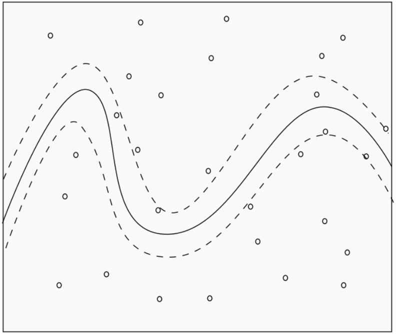
$$\exists L > 0 \quad \forall x \in K \quad \forall h \qquad |\eta(x+h) - \eta(x)| \leq L\|h\|.$$

II - 3 Hypothèse de Marge

Hypothèse de Marge $H_{MA}(\alpha)$ introduite par Mammen & Tsybakov ('99) : Il existe $\alpha \geq 0$, une constante $C > 0$ et ϵ_0 assez petit tel que

$$\forall \epsilon \leq \epsilon_0 \qquad \mathbb{P}_X\left[\left|\eta(X) - \frac{1}{2}\right| \leq \epsilon\right] \leq C\epsilon^{\alpha}$$

Trait plein : $\eta = 1/2$, pointillés : $\eta = 1/2 \pm \epsilon$.



- C'est une propriété locale autour de la frontière de décision $\eta = 1/2$.
- Si $\alpha = +\infty$, η a une discontinuité spaciale et "saute" au niveau $1/2$.
- Si η "traverse" la frontière $1/2$, alors $\alpha = 1$.
- Si η a r dérivées nulles sur l'ensemble $\eta = 1/2$, alors $\alpha = \frac{1}{r+1}$.

II - 4 Bornes inférieures générales de classification

Audibert & Tsybakov démontrent la borne inférieure de risque

Théorème (AT '07)

(a) *Si la fonction* η *est telle que* $\eta^{(\beta)}$ *est Lipschitz et sous l'hypothèse* H_{SDA} *et* $H_{MA}(\alpha)$*, alors si* $\alpha\beta < d$ *et pour tout algorithme de classification* Φ_n *:*

$$\mathcal{R}(\Phi_n) - \mathcal{R}(\Phi_{Bayes}) \geq C n^{-\beta(1+\alpha)/(2\beta+d)},$$

(b) *La minoration est optimale : on peut construire des algorithmes de classification atteignant cette vitesse asymptotique.*

- Cas standard : $\beta = 1$ et $\alpha = 1$, excès de risque $\sim n^{\frac{-2}{2+d}}$ où d est la dimension.
- Dans la situation où $\alpha = \infty$, on peut trouver des estimateurs tels que la consistance est obtenue asymptotiquement avec une vitesse exponentielle (!).
- Il y a de nombreuses valeurs de paramètres pour lesquels on obtient des vitesses plus rapides que $n^{-1/2}$ (classification plus facile que la régression).

II - 5 K plus proches voisins

La classif par K p.p.v. est une classif **plug-in** : étant donné $\mathcal{D}_n$, on définit

$$\forall x \in K \qquad \eta_{n,k}(x) := k_n^{-1} \sum_{j=1}^{k} Y_{(j)}(x)$$

On a alors

$$\Phi_{n,k}(x) = 1_{\eta_{n,k}(x) > 1/2}.$$

Si $\mathcal{B}_\epsilon := \left\{x \in \mathbb{R}^d : |\eta(x) - 1/2| \leq \epsilon\right\}$, :

$$\begin{aligned} \mathcal{R}(\Phi_{n,k}) - \mathcal{R}(\Phi_{Bayes}) \quad &= \quad \underbrace{\mathbb{E}\left[|2\eta(X) - 1|\mathbf{1}_{\{\Phi_{n,k}(X) \neq \Phi^*(X)\}}\mathbf{1}_{X \in \mathcal{B}_\epsilon}\right]}_{:= T_{1,\epsilon} \leq \epsilon^{1+\alpha} \qquad (H_{MA}(\alpha))} \\ &+ \underbrace{\mathbb{E}\left[|2\eta(X) - 1|\mathbf{1}_{\{\Phi_{n,k}(X) \neq \Phi^*(X)\}}\mathbf{1}_{X \in \mathcal{B}_\epsilon^c}\right]}_{:= T_{2,\epsilon}}. \end{aligned}$$

- Quantifier l'effet moyen de $\eta_{n,k}$ (biais autour de η et variance)
- Inégalité de concentration sur $\eta_{n,k}$.

Rôle de k et hypothèses sur les distributions cruciales

II - 5 K plus proches voisins - Décomposition du risque

Quantifier l'effet moyen de $\eta_{n,k}$ (proche de η ?) Pour une observation X quelconque

$$\begin{aligned}\Delta_n(X) &= \mathbb{E}_X|\eta_{n,k}(X)-\eta(X)| \\ &\leq \underbrace{\mathbb{E}_X\left(\left|\frac{1}{k}\sum_{i=1}^{k}Y_{(i)}(X)-\eta(X_{(i)})\right|\right)}_{:=S_1 \quad \textbf{variance}}+\underbrace{\mathbb{E}_X\left(\left|\frac{1}{k}\sum_{i=1}^{k}\eta(X_{(i)})-\eta(X)\right|\right)}_{:=S_2 \quad \textbf{biais}}\end{aligned}$$

Contrôle du biais : Intimement relié au volume (probabiliste) des boules $B(X,r)$.

$$\begin{aligned}S_2 \lesssim \mathbb{E}_X\left|X_{(k)}-X\right| &\lesssim \delta+\int_{\delta}^{+\infty}P\left(\left|X_{(k)}-X\right|>r\right)dr \\ &\lesssim \delta+\int_{\delta}^{+\infty}P\left(\frac{1}{n}\sum_{j=1}^{n}\mathbf{1}_{X_j\in B(X,r)}<\frac{k}{n}\right)dr\end{aligned}$$

Proposition

Sous l'hypothèse H_{SDA} et si η est L Lipschitz, on démontre que le biais est majoré en

$$S_2 \lesssim \left(\frac{k}{n}\right)^{1/d}+e^{-3k/8}$$

Majoration obtenue en ajustant δ avec $B(X,r)\sim C_d r^d$ pour r petit.

II - 5 K plus proches voisins - Cas de la classification

Contrôle de la variance : Les $(X_i)_{i=1\dots n}$ sont i.i.d. de loi P_X, puis cond. aux positions (X_i), les labels sont tirés indépendamment selon une $\mathcal{B}(\eta(X_i))$.

$$S_1 \leq k^{-1/2} \sqrt{\mathbb{E}\left\{\mathbb{E}_X\left(\frac{1}{k}\left[\sum_{i=1}^{k} Y_{(i)}(X) - \eta(X_{(i)})\right]^2 \quad \Big| (X_1, \dots, X_n)\right)\right\}}.$$

Une fois qu'on effectue le cond. $|(X_i)_{1 \leq i \leq n}$, l'inégalité de Hoeffding implique

Proposition

On a pour tout X dans K :

$$S_1 \leq \frac{1}{2\sqrt{k}},$$

et surtout :

$$\forall t \geq 0 \qquad \mathbb{P}_X\left(|\eta_{n,k}(X) - \mathbb{E}_X \eta_{n,k}(X)| > t\right) \lesssim e^{-Ckt^2},$$

où C est une constante universelle indépendante de X.

II - 5 K plus proches voisins - Cas de la classification

L'excès de risque s'optimise en choisissant k tel que $k_n^{-1/2} \simeq (k_n/n)^{1/d}$.

Théorème (Excès de risque du K-NN, Classif., GKM'13)

Pour $k_n \sim n^{2/(2+d)}$ et sous H_{SDA} et $H_{MA}(\alpha)$:

$$\mathcal{R}(\Phi_{n,k_n}) - \mathcal{R}(\Phi_{Bayes}) \lesssim n^{-(1+\alpha)/(2+d)}.$$

- État de l'art : vitesses en $n^{-1/(2+d)}$ selon les situations étudiées (Györfi'78, Devroye'81, Biau & Cerou & Guyader'10).
- Pour $d = 1$ et $\alpha = 1$, la vitesse est en $n^{-2/3}$, plus rapide que $n^{-1/2}$.
- Vitesse optimale sous les hypothèses H_{SDA} et $H_{MA}(\alpha)$ seulement lorsque η est Lipschitz (non optimale lorsque la régularité β de η augmente)

$$r_n(\alpha, \beta) \sim n^{-\beta(1+\alpha)/(2\beta+d)}.$$

- Les K-NN ne « lisse » pas assez pour exploiter la régularité de η.
- Dimension d dévastatatrice : plus d augmente, plus le problème est difficile.
- Le biais est de plus en plus mauvais ; il y a plus de $B(X, r)$ quand d est grand.

II - 6 K plus proches voisins - Cas de l'analyse discriminante

L'excès de risque se décompose toujours en

$$\Delta_n(X) \leq S_1 + S_2, \qquad \text{avec} \qquad S_2 \lesssim \left(\frac{k}{n}\right)^{1/d} + e^{-3k/8}.$$

Le terme S_1 est vraiment très ennuyeux

$$S_1 = \mathbb{E}_X \left(\left| \frac{1}{k} \sum_{i=1}^{k} Y_{(i)}(X) - \eta(X_{(i)}) \right| \right).$$

En A.D., conditionnellement aux positions $(X_{(i)})_{1 \leq i \leq n}$, les $(Y_{(i)}(X))_{1 \leq i \leq n}$ sont toutes dépendantes entre elles, contrairement au contexte de classification.
De même, la concentration de

$$\eta_{n,k}(X) = \frac{1}{k} \sum_{i=1}^{k} Y_{(i)}(X)$$

autour de sa moyenne est problématique.

II - 6 K plus proches voisins - Cas de l'analyse discriminante

- On procède à une étape de **Poissonisation** (Kac, '49) : considérons un modèle où les effectifs des deux classes sont (N_1, N_2), deux v.a. i.i.d. de loi $\mathcal{P}(n/2)$.
- **Avantage :** On tire profit de nombreuses propriétés des processus de Poisson.
- On définit $Z = \sigma \circ (X_1, \dots, X_{N_1}, X'_1, \dots, X'_{N_2})$ et l'algorithme des k ppv s'écrit alors à l'aide de $Z^{\mathcal{P}}_{(k)}(x)$ et

$$\Sigma_r := N_X([0, r[) = \# \left\{ i \in \{1, \dots N_1\} : \|x - X_{(i)}(x)\| \leq r \right\}.$$

- On sait alors que $\eta^{\mathcal{P}}_{n,k}(x) = \frac{1}{k} \Sigma_{Z_{(k)}(x)}$

Proposition (Application de la formule de Campbell & Mecke)

Si on note $z_{k,x} = \mathbb{E}_x Z_{(k)}(x)$, on a

$$\mathbb{E}_x \left[\eta^{\mathcal{P}}_{n,k}(x) \right] = \frac{1}{k} \left[\mathbb{E}_x \Sigma_{z_{k,x}} \right]$$

- On obtient "facilement" des inégalités de concentration pour $\eta^{\mathcal{P}}_{n,k}$ ...

II - 6 K plus proches voisins - Cas de l'analyse discriminante

- La *dépoissonisation* s'effectue en remarquant que

$$(\eta_{n,k}) \underbrace{=}_{\mathcal{L}} \left(\eta^{\mathcal{P}}_{n,k} \,|\, (N_1, N_2) = (n/2, n/2)\right) :$$

- Ainsi

$$\mathbb{P}\left(|\eta_{n,k}(x) - \mathbb{E}_x \eta_{n,k}(x)| \geq t\right) \quad \leq \quad \frac{\mathbb{P}\left(|\eta^{\mathcal{P}}_{n,k} - \mathbb{E}_x \eta^{\mathcal{P}}_{n,k}(x)| \geq t\right)}{\mathbb{P}\left[(N_1, N_2) = (n/2, n/2)\right]} \lesssim n e^{-Ckt^2}.$$

- On obtient le résultat de consistance

Théorème (Excès de risque du K-NN, An. Disc. ,GKM'13)

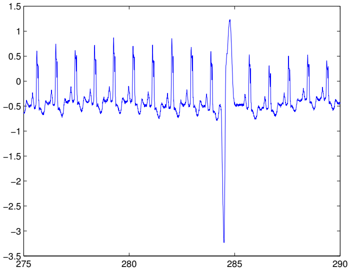
Pour $k_n \sim n^{2/(2+d)}$ *et sous* H_{SDA} *et* $H_{MA}(\alpha)$ *:*

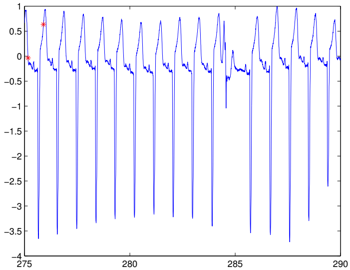
$$\mathcal{R}(\Phi_{n,k_n}) - \mathcal{R}(\Phi_{Bayes}) \lesssim \left(\frac{\log(n)}{n}\right)^{-(1+\alpha)/(2+d)}.$$

Perte d'un logarithme et résultat *presque* optimal.

III - 1 Autres types de données

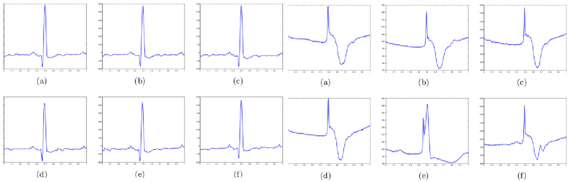
Deux exemples d'enregistrements d'ECG de 2 patients souffrant d'arythmie.





Question : segmenter les cycles, détecter l'arythmie :

Gauche : cycle normal, Droite : cycle arythmique.



III - 2 K plus proches voisins et variables fonctionnelles

- Dans l'exemple précédent, les données à classer **ne sont plus dans** $\mathbb{R}^d$ et se représentent *a priori* dans $\mathcal{H}_s$, espace de fonctions (régularité s).
- Modèle Gaussien de bruit blanc et dans un cadre de classification :

  $$(X|Y=0) \sim dX_t = f(t)dt + dW_t \iff (X|Y=0) \sim (x_k)_{k\in\mathbb{Z}} \text{ où } x_k = \theta_k(f) + \xi_k.$$

  $$(X|Y=1) \sim dX_t = g(t)dt + dW_t \iff (X|Y=1) \sim (x_k)_{k\in\mathbb{Z}} \text{ où } x_k = \theta_k(g) + \xi_k.$$

  $(\xi_k)_{k\in\mathbb{Z}}$ sont des v.a. i.i.d. $\mathcal{N}_{\mathbb{C}}(0,1)$, $(\theta_k(f))_{k\in\mathbb{Z}}$: coefficients de Fourier de f.
- Deux problèmes : le support des observations n'est plus compact et $d = "\infty"$.

Dans le cas gaussien avec support non compact, on a :

Théorème (GKM'13)

L'algorithme du kppv est "optimisé" pour $k_n(x) \sim [1 + n^{1/(2+d)} e^{-|x|^2/2}]$ et on a

$$\mathcal{R}(\Phi^d_{n,k_n}) - \mathcal{R}(\Phi^d_{Bayes}) \lesssim n^{-1/(2+d)}.$$

où Φ^d_{Bayes} est le classifieur Bayésien sur $\mathbb{R}^d$ et Φ^d_{n,k_n} le k_n ppv dans les d premières dimensions. On constate que $k_n(x) \longrightarrow 1$ lorsque $\|x\| \longrightarrow +\infty$.

Le nombre de voisins $k_n(x)$ **corrige le faible volume proba. des boules à l'infini.**

III - 3 Consistance fonctionnelle des K plus proches voisins

Quantifier l'approximation entre $\mathcal{H}_s$ et $\mathbb{R}^d$?

- La règle de classification fonctionnelle sera en réalité dans $\mathbb{R}^{d_n}$ avec $d_n \longrightarrow +\infty$.
- La résultat précédent quantifie l'écart entre $\Phi^{d_n}(n, k_n)$ et $\Phi^{d_n}_{Bayes}$ ...

Les données proviennent d'une fonction de $\mathcal{H}_s$, ainsi les $\theta(f)$ et $\theta(g)$ vérifient :

Hypothèse :

$$\sum_{j\geq} |\theta_j|^2 j^{2s} \leq A \qquad (\mathcal{H}_s(A))$$

Proposition (GKM'13)

Pour toute fréquence de coupure d, si $(f, g) \in \mathcal{H}_s(A)$:

$$\mathcal{R}(\Phi^d_{Bayes}) - \mathcal{R}(\Phi_{Bayes}) \leq d^{-s}$$

.

Dans ce contexte, on démontre alors le (dernier) résultat

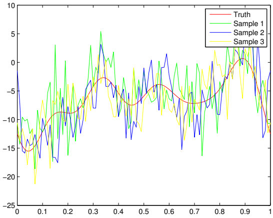
Théorème (GKM'13)

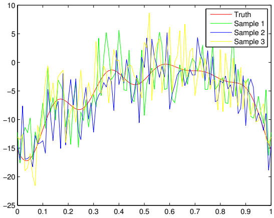
Dans le modèle gaussien sur $\mathcal{H}_s(A)$, le choix $d_n \sim C\frac{\log(n)}{\log\log(n)}$ avec $k_n(x)$ voisins assure que :

$$\mathcal{R}(\Phi^{d_n}_{n,k_n}) - \mathcal{R}(\Phi_{Bayes}) \lesssim \log(n)^{-s}.$$

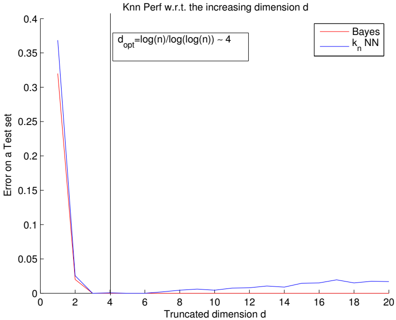
III - 4 Données simulées

Représentation de deux classes synthétiques :





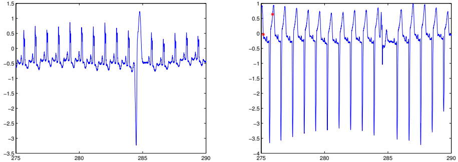
Évolution de l'erreur de classification : $n_{train} = 400, n_{test} = 2000, s = 1.$



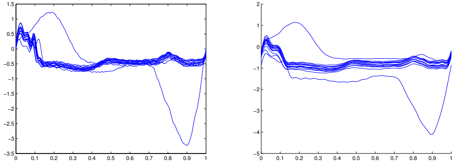
III - 4 Données ECG

Ne surtout pas faire avec un K ppv : application aveugle à la nature du pb ...

- Données



- Question : identifier la présence d'un ou plusieurs cycles arithmiques.
- Pre-process : Identifier les cycles (normalisés sur un intervalle de longueur 1)



- Benchmark : 10 individus, 20 cycles par individus et en moyenne 2 cycles sur 20 arythmiques. Erreur cv : $\simeq 25\%$ (!) en décomposant en Fourier
- Résultats numériques très peu convaincants, il faut repenser le probème ...

III - 5 K plus proches voisins et opérateur de déformations

- Si on note H l'opérateur de déformation (translation + homothétie)

$$(X|Y=0) \sim dX_t = (H \circ f)(t)dt + dW_t, \qquad H \perp\!\!\!\perp W.$$

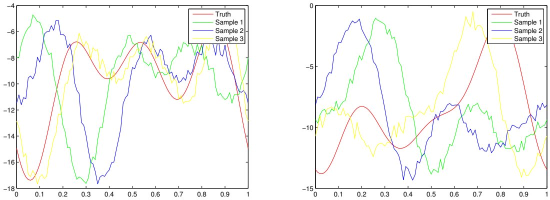
- Utiliser les k-ppv sans considérer l'action de H revient au "sacrifice" statistique (le biais dans ce modèle est naturellement important, les boules probabilistes nombreuses). Pour que 2 signaux $(X_{\omega_1}, X_{\omega_2})$ soient proches, il faut

$$H_{\omega_1} \simeq H_{\omega_2} \qquad W_{\omega_1} \simeq W_{\omega_2}.$$

- Exemple sur l'opérateur de translation aléatoire :

$$dX_t = f(t-\tau)dt + dW_t \qquad \tau_1 \simeq \tau_2$$

Deux classes synthétiques :



- Idée : placer le problème de classif dans la classe des orbites sous l'action de H.

III - 6 K plus proches voisins et opérateur de déformations

- Groupe $\mathbb{T}$, action de $\tau \in \mathbb{T}$: $f \longrightarrow f^{-\tau}$.
- Les modules en Fourier $|\theta_k|$ sont invariants, mais ne caractérisent pas les orbites.
- Étant données trois fréquences (k_1, k_2, k_3) sommant à 0, on considère :

$$S(f)_{k_1,k_2,k_3} = \theta_{k_1}(f)\theta_{k_2}(f)\theta_{k_3}(f).$$

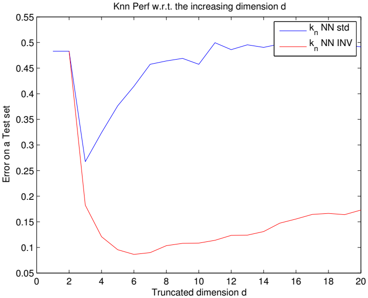
- Classification k-ppv sur les données transformées par S, seuillées en fréquences.

Proposition (GKM'13)

Les classes d'équivalences à translation près sont identifiables au travers de

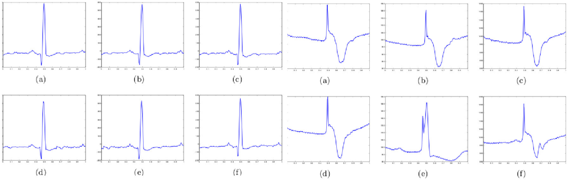
$$(S(f)_{k_1,k_2,k_3})_{(k_1,k_2,k3):k_1+k_2+k_3=0}.$$

K ppv "invariants" : classer les données sur $(S(f)_{k_1,k_2,k_3})_{(k_1,k_2,k3):k_1+k_2+k_3=0}$.
Classification Données simulées :



III - 6 K plus proches voisins et opérateur de déformations

Gauche : cycle normal, Droite : cycle arythmique.



Données ECG, $n = 200$ courbes, $d \sim 5$ fréquence de coupure.
Comparaison : K ppv standard - K ppv par Dynamic Time Warping
Performance : Précision - Temps de calcul

Algorithme	Erreur	Temps de calcul
K ppv	25 %	5 sec.
K ppv inv	2 %	10 sec.
DTW	2 %	5 min.

IV Conclusion

Remarques importantes :

- Importance du cadre statistiques (classification / analyse discriminante).
- L'hypothèse de Marge provoque une accéleration du risque.
- L'important est de comprendre la structure de voisinage et la taille des petites boules probabilistes autour de chaque observation.
- Les kppv sont à utiliser avec précaution (variables descriptives, aléa). Données ECG : erreur de classif. passe de 25% à moins de 2% en utilisant les invariants.

Extensions Mathématiques :

- Pas d'utilisation de la régularité de η. (Cf Samworth '12). Pondération ?
- Résultat non optimal en Gaussien (perte minimax en $\frac{d}{n}$ au lieu de $n^{2/(2+d)}$).
- Vitesse non optimale (mais presque) en A.D. avec une présence d'un $\log(n)$.
- Meilleure inégalité de concentration ? Approche alternative à la Poissonisation par des variables N.A. (c'est presque le cas pour $\eta_{n,k}$ lorsque $n \longmapsto +\infty$).
- Variables d'entrée sont perturbées par un opérateur partiellement connu/inconnu d'un point de vue stat. math, aspects semi-paramétriques ou non paramétriques.
- Borne inférieure en approche fonctionnelle...
- Coupler avec une sparse PCA (réduction de dimension et du biais dans les ppv)
- ...

Merci de votre attention !