MVA : 3e séance

9.02.2012
A. Dalalyan

I. Séance Précédente :

$X_1, \dots, X_n$ iid de densité f sur $[0,1]$

$K \geq 1$ un entier, $h = 1/K$, $C_k = [(k-1)h ; kh[$.

Estimateur par histogramme :

$$\hat{f}_{n,h}(x) = \frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}_{C_k}(X_i) \qquad \text{si } x \in C_k \ ; k = 1, \dots, K.$$

On introduit les notations

$$p_k = \int_{C_k} f(x)\,dx \qquad \hat{p}_k = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{C_k}(X_i)$$

On a vu que

$$\bullet\ \mathbb{E}\left[\|\hat{f}_{n,h} - f\|_2^2\right] = \underbrace{\|f\|_2^2 - \frac{1}{h}\sum_{k=1}^{K} p_k^2}_{\text{Carré du biais}} + \underbrace{\frac{1}{nh}\sum_{k=1}^{K} p_k(1-p_k)}_{\text{Variance}}$$

De plus si $f \in \mathrm{Lip}(L)$, alors

$$\mathbb{E}\left[\|\hat{f}_{n,h} - f\|_2^2\right] \leq L^2 h^2 + \frac{1}{nh} \qquad (*)$$

II. Remarques

1. Le minimum par rapport à h de $(*)$ est atteint lorsque $h = h_{opt} = (2nL^2)^{-1/3}$. Dans ce cas, le risque est majoré par $3\,(L/2)^{2/3}\, n^{-2/3}$

2. Si on considère des fonctions f appartenant à la classe de Hölder : $H(\beta, L) = \{ f : |f(x) - f(y)| \leq L|x-y|^{\beta} \}$

alors la borne (*) se transforme en

$$\mathbb{E}\left[\|\hat{f}_{n,h} - f\|_2^2\right] \leq L^2 h^{2\beta} + \frac{1}{nh}$$

Le minimum par rapport à h de cette expression est atteint lorsque $h_{opt} = C\, n^{-\frac{1}{2\beta+1}}$. Le risque optimal est alors de l'ordre de $n^{-2\beta/2\beta+1}$.

(3) On voit que le risque optimal ainsi que la fenêtre optimale dépendent des valeurs (L,β) qui sont inconnues en pratique. Pour pallier ce défaut, on introduit des estimateurs adaptatifs. L'idée est de choisir h en fonction des données.

(4) Les résultats précédents se généralisent facilement au cas multidimensionnel : $X_1, \dots, X_n$ iid de densité $f : [0,1]^d \to \mathbb{R}$. La vitesse optimale est alors obtenue en minimisant par rapport à h l'expression $L^2 h^{2\beta} + \frac{1}{nh^d}$. Cela donne $h_{opt} = n^{-1/2\beta+d}$ et un risque de l'ordre de $n^{-2\beta/2\beta+d}$.

Cette vitesse est d'autant plus lente que la dimension est grande. De plus, ce ralentissement se fait de façon exponentielle, car $n^{-2\beta/2\beta+d} = \exp\left\{-\frac{2\beta \ln n}{2\beta+d}\right\}$ et donc si l'on augmente d par Δd, il faut augmenter n t.q.

$$\ln n_{new} \approx \ln n_{old} + \Delta d \implies n_{new} = n_{old}\, e^{\Delta d}.$$

On appelle ce phénomène "curse of dimensionality"

On y reviendra plus tard dans ce cours.

III. Adaptation par minimisation de l'estimateur sans biais du risque

① Idée principale

Supposons maintenant qu'on a M fenêtres potentielles:

$$h_1, h_2, \dots, h_M$$

Comment trouver celle qui donne le meilleur estimateur?

Notons $L(h, f) = \| \hat{f}_{n,h} - f \|_2^2$

On a vu que

$$L(h,f) = \|f\|_2^2 - \frac{1}{h} \sum_{k=1}^{K} p_k^2 + \frac{1}{nh} \sum_{k=1}^{K} p_k(1-p_k)$$

$$= \|f\|_2^2 + J(h).$$

La meilleure fenêtre est celle qui minimise le risque, ce qui équivaut à minimiser $J(h)$:

$$h^* \in \arg\min_{h \in \{h_1, \dots, h_M\}} J(h).$$

Or h^* est impossible à calculer à cause de non-disponibilité de la densité f et donc des p_k.

Pour contourner cette difficulté, on remplace J par un estimateur sans biais.

<u>Proposition</u> : Pour toute fenêtre déterministe h, l'expression

$$\hat{J}_n(h) = \frac{2}{(n-1)h} - \frac{n+1}{(n-1)h} \sum_{k=1}^{K} \hat{p}_k^2$$

fournit un estimateur sans biais de $J(h)$.

Preuve.

Comme on l'a déjà remarqué $n\hat{p}_k \sim \mathcal{B}(n, p_k)$

[la variable aléatoire $n\hat{p}_k$ suit la loi binomiale de paramètres n et p_k.]

Par conséquent $\mathbb{E}[n\hat{p}_k] = n p_k$ et $\mathrm{Var}[n\hat{p}_k] = np_k(1-p_k)$.

On en déduit que

$$\mathbb{E}[\hat{p}_k^2] = \mathrm{Var}[\hat{p}_k] + \left(\mathbb{E}[\hat{p}_k]\right)^2 = \frac{p_k(1-p_k)}{n} + p_k^2 .$$

Par conséquent,

$$\mathbb{E}[\hat{J}_n(h)] = \frac{2}{(n-1)h} - \frac{n+1}{(n-1)h}\sum_{k=1}^{K}\left(\frac{p_k - p_k^2}{n} + p_k^2\right)$$

$$= \frac{2}{(n-1)h} - \frac{n+1}{n(n-1)h}\underbrace{\sum_{k=1}^{K} p_k}_{=1} - \frac{n+1}{(n-1)h}\sum_{k=1}^{K}\frac{n-1}{n}p_k^2$$

$$= \frac{1}{nh} - \frac{n+1}{nh}\sum_{k=1}^{K} p_k^2 = J(h) \quad \blacksquare$$

2) Algorithme :

- On a $X_1, \dots, X_n$.
- On se donne un paramètre $a > 1$ (typiquement $a = 1{,}2$)
- On pose $h_1 = \frac{1}{n}$, $h_m = \frac{1}{[n a^{1-m}]}$ $m = 1, \dots, M$
- On calcule $\hat{J}_n(h_1), \dots, \hat{J}_n(h_M)$

  et $\hat{m} \in \arg\min_m \hat{J}_n(h_m)$
- On pose $\hat{f}(x) = \hat{f}_{n,h_{\hat{m}}}(x)$.

③ Pourquoi ça marche ?

Ecrivons la perte sous la forme :

$$L(h,f) = \|f\|_2^2 + \widehat{J}_n(h) + \xi_n(h).$$

On a

$$\mathbb{E}[\xi_n(h)] = \mathbb{E}[L(h,f)] - \|f\|_2^2 - \mathbb{E}[\widehat{J}_n(h)]$$
$$= \mathbb{E}[L(h,f)] - \|f\|_2^2 - J(h) = 0.$$

Donc $\{\xi_n(h) : h \in \{h_1,\dots,h_M\}\}$ est un vecteur aléatoire de moyenne nulle. Posons

$$V_n(h) = \mathrm{Var}[\xi_n(h)]$$

__Lemme 1.__ Quelle que soit la variable aléatoire $\tilde{h}$ à valeurs dans $\{h_1,\dots,h_M\}$, on a

$$\mathbb{E}[\xi_n(\tilde{h})] \leq \sqrt{\mathbb{E}[V_n(\tilde{h})]} \times M^{1/2}.$$

__Preuve.__

$$\mathbb{E}[\xi_n(\tilde{h})] = \mathbb{E}\left[V_n(\tilde{h})^{\frac{1}{2}} \times \frac{\xi_n}{\sqrt{V_n}}(\tilde{h})\right]$$
$$\leq \mathbb{E}\left[V_n(\tilde{h})^{\frac{1}{2}} \times \max_m \frac{|\xi_n|}{\sqrt{V_n}}(h_m)\right]$$
$$\overset{\text{Cauchy-Schw.}}{\leq} \mathbb{E}[V_n(\tilde{h})]^{1/2} \times \mathbb{E}\left[\max_m \frac{\xi_n^2}{V_n}(h_m)\right]^{1/2}$$
$$\leq \mathbb{E}[V_n(\tilde{h})]^{1/2} \times \Big(\sum_{m=1}^{M} \underbrace{\mathbb{E}\left[\frac{\xi_n^2}{V_n}(h_m)\right]}_{=1}\Big)^{1/2} \quad ▨$$

On a donc

$$\mathbb{E}\left[L(\hat{h},f)\right] = \|f\|_2^2 + \mathbb{E}\left[\hat{J}_n(\hat{h})\right] + \mathbb{E}\left[\xi_n(\hat{h})\right]$$

$$\leqslant \|f\|_2^2 + \mathbb{E}\left[\hat{J}_n(h^*)\right] + \sqrt{M \cdot \mathbb{E}\left[V_n(\hat{h})\right]}\,.$$

Cette inégalité est vraie pour tout $h^* \in \{h_1, \dots, h_M\}$, car $\hat{h}$ est par définition le minimiseur de $\hat{J}_n$. Par conséquent, l'inégalité ci-dessus reste vraie pour

$$h^* = \arg\min_{h \in \{h_1,\dots,h_M\}} J(h) = \arg\min_h \mathbb{E}\left[L(h,f)\right]$$

On a donc

$$\mathbb{E}\left[L(\hat{h},f)\right] \leqslant \|f\|_2^2 + \mathbb{E}\left[\hat{J}_n(h^*)\right] + \sqrt{M\,\mathbb{E}\left[V_n(\hat{h})\right]}$$

$$= \mathbb{E}\left[L(h^*,f)\right] - \mathbb{E}\left[\xi_n(h^*)\right] + \sqrt{M\,\mathbb{E}\left[V_n(\hat{h})\right]}$$

$$\leqslant \mathbb{E}\left[L(h^*,f)\right] + \sqrt{M\,\mathbb{E}\left[V_n(h^*)\right]}$$

Cela nous conduit vers l'inégalité oracle suivante :

$$\mathbb{E}\left[L(\hat{h},f)\right] \leqslant \underbrace{\min_{h \in \{h_1,\dots,h_M\}} \mathbb{E}\left[L(h,f)\right]}_{\text{risque de l'oracle}} + \underbrace{\sqrt{M} \cdot \sqrt{\max_m V_n(h_m)}}_{\text{terme résiduel}}$$

Le choix de $h_1, \dots, h_M$ proposé dans l'algorithme garantit que $M = \text{Const} \cdot \ln n$ est à croissance lente. On peut également démontrer que $V_n(h_m) \leqslant C/n$ pour une constante C bien choisie.

4. A quoi ressemble $\xi_n(h)$?

$$\xi_n(h) = L(h,f) - \|f\|_2^2 - \widehat{J}_n(h)$$

$$= \|\widehat{f}_{n,h}\|_2^2 - 2\langle \widehat{f}_{n,h}, f\rangle - \widehat{J}_n(h)$$

$$= \sum_{k=1}^{K} \frac{1}{h}\,\widehat{p}_k^{\,2} - 2\sum_{k=1}^{K} \int_{C_k} \frac{\widehat{p}_k}{h} f(x)\,dx - \widehat{J}_n(h)$$

$$= \frac{1}{h}\sum_{k=1}^{K} \widehat{p}_k^{\,2} - \frac{2}{h}\sum_{k=1}^{K} \widehat{p}_k p_k - \frac{2}{(n-1)h} + \frac{n+1}{(n-1)h}\sum \widehat{p}_k^{\,2}$$

$$= \frac{2n}{(n-1)h}\sum_{k=1}^{K} \widehat{p}_k^{\,2} - \frac{2}{h}\sum_{k=1}^{K}\Big[(\widehat{p}_k - p_k)p_k + p_k^2\Big] - \frac{2}{(n-1)h}\sum_{k=1}^{K} p_k$$

$$= \frac{2n}{(n-1)h}\sum_{k=1}^{K}\Big[\widehat{p}_k^{\,2} - \underbrace{\frac{n-1}{n}p_k^2 - \frac{1}{n}p_k}_{\mathbb{E}[\widehat{p}_k^{\,2}]}\Big] - \frac{2}{h}\sum_{k=1}^{K} p_k(\widehat{p}_k - p_k)$$

Posons

$$\xi_n^{(1)}(h) = \frac{2n}{(n-1)h}\sum_{k=1}^{K}\left(\widehat{p}_k^{\,2} - \mathbb{E}\left[\widehat{p}_k^{\,2}\right]\right)$$

$$\xi_n^{(2)}(h) = \frac{2}{h}\sum_{k=1}^{K} p_k(\widehat{p}_k - p_k).$$

__Lemme__ Quelle que soit la valeur $h > 0$, on a

$$\mathrm{Var}\left[\xi_n^{(2)}(h)\right] \quad \frac{4}{nh^2}\sum_{k=1}^{K} p_k^3(1-p_k)$$

$$\leq (4/n) \times \sup_{x\in[0,1]} f(x)^2.$$

__Preuve__ Il suffit de remarquer que $\xi_n^{(2)}(h) = \frac{2}{nh}\sum_{i=1}^{n} Z_i$ où les variables aléatoires Z_i sont iid de variance $\sum_{k=1}^{K} p_k^3(1-p_k)$. Pour conclure, on utilise $1-p_k \leq 1$ et $p_k \leq h\cdot\sup_x f(x)$.