Minimum Description Length
identification de modèles à partir de données

Maria-João Rendas

CNRS – I3S

Novembre 2006

Problème

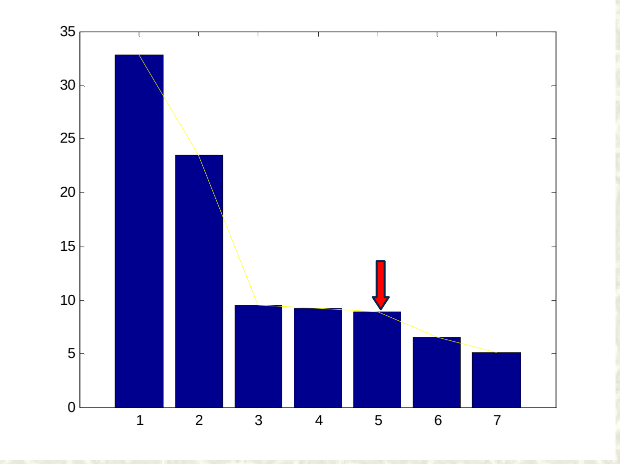
Étant données des observations $x^{(n)}$, choisir un modèle $\mathcal{H}$ qui exprime ses *propriétés intrinsèques*.

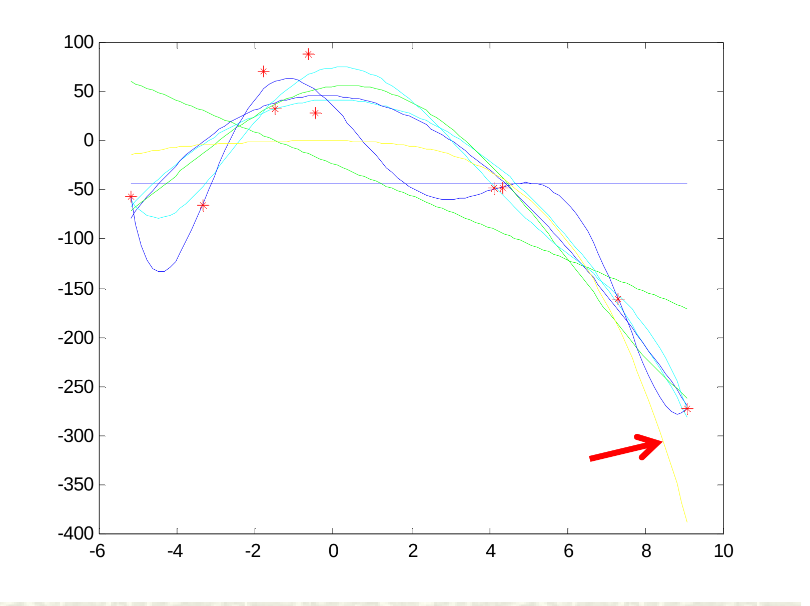
Exemples

- ajuste d'un modèle polynomial à des paires de valeurs réels
- segmentation non-supervisée (images, signaux,...)
- ajuste d'une distribution de probabilité à des échantillons

Ajuste d'un modèle polynomial

- Données
  - $x^{(n)} = [(x_1,y_1), (x_2,y_2),\ldots, (x_n,y_n)]$

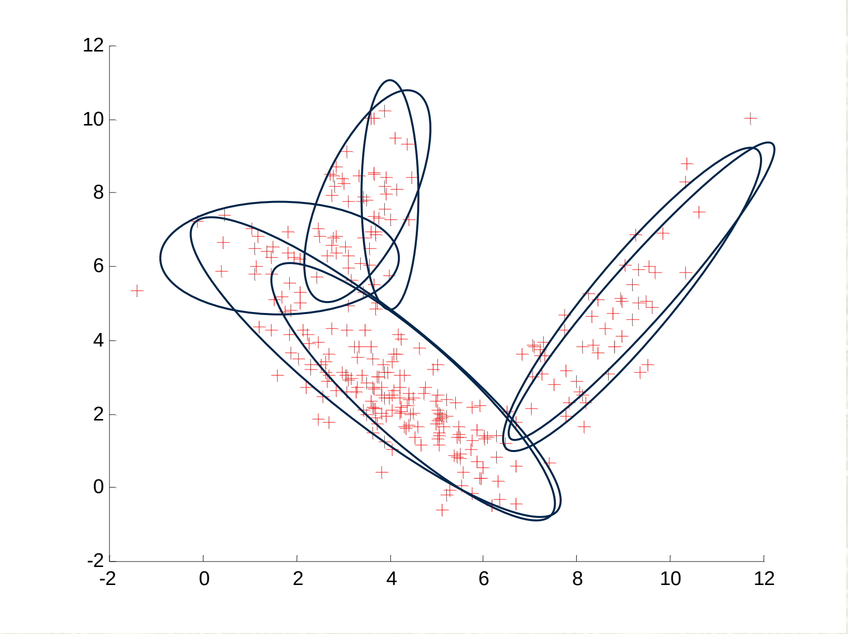


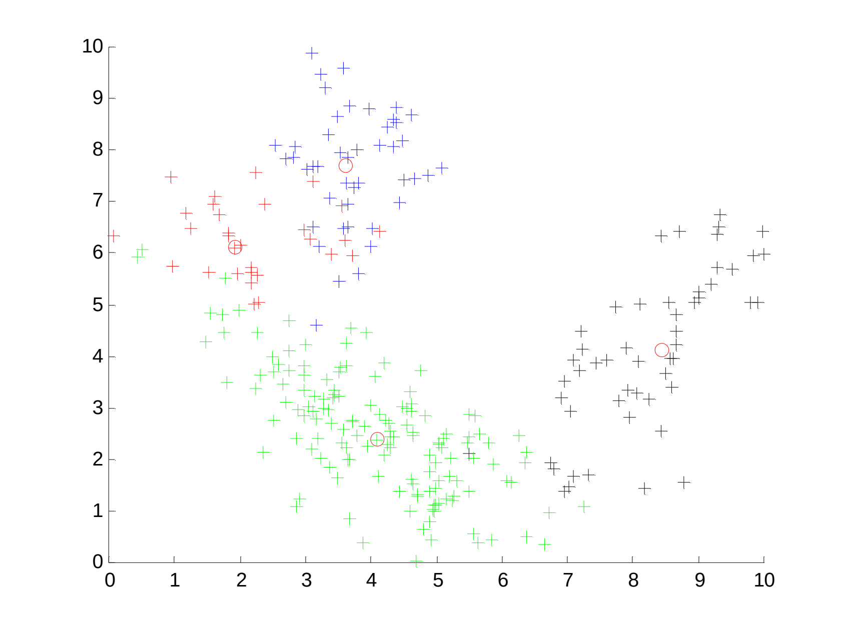


- Modèles candidats
  - $\mathcal{H}_k : y_i = a_0 + a_1 x_i + \ldots + a_k x_i^k, \quad k=0,1,2,\ldots$

Segmentation non-supervisée

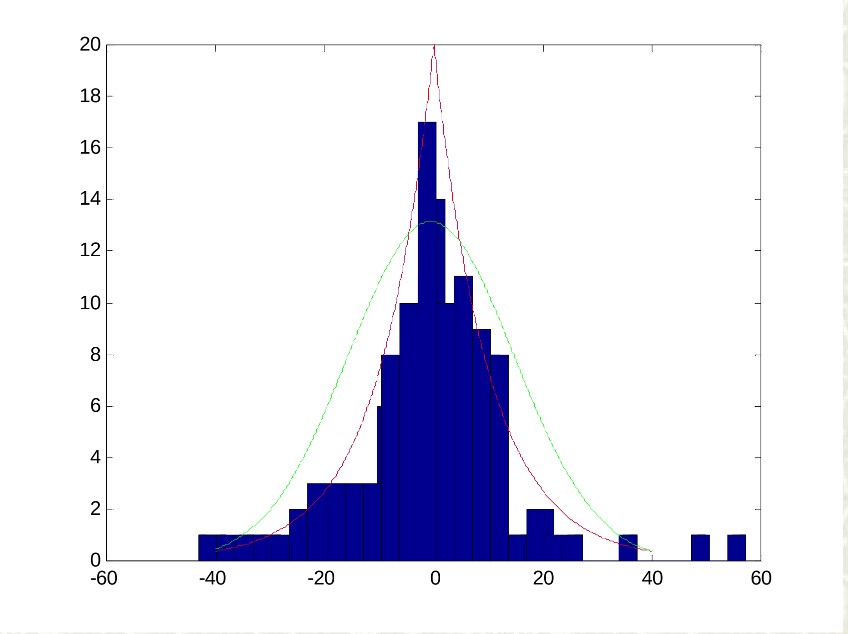
- Données
  - $x^{(n)} = [x_1, x_2, \ldots, x_n]$





Ajuste d'une distribution de probabilité

- Données
  - $x^{(n)} = [x_1, x_2, \ldots, x_n]$
- Modèles candidats
  - $\mathcal{H}_1 : x_i \sim \mathcal{N}(x_i : \mu, \sigma)$
  - $\mathcal{H}_2 : x_i \sim (2\lambda)^{-1} e^{-\lambda|x|}$



Principe de Longueur de Description Minimale

Choisir le modèle qui permet

la **codification la plus compacte** des données

Considère le problème de choix de modèles comme celui de déceler les *régularités* des données.

Basé sur *(i)* la relation intime entre (longueurs de) codes optimaux et lois de probabilité qui découle de l'**inégalité de Kraft,** et *(ii)* la notion de **code universel**

source p codeur C

$p_C(x) = 2^{-\lambda(C_p(x))}$ $\lambda(C_p(x)) = -\log p(x)$

Choisir le **code** optimal pour un ensemble de données est équivalent à trouver la **distribution** de probabilité de la source.

Définitions et notation

Modèle probabiliste

$\mathcal{H}=\{p_\gamma(x^n), \gamma\in\Gamma\}$ Γ peut être : fini, dénombrable, continu...

- ***Modèle paramétrique*** :

$\mathcal{H}^\Theta=\{p(x^n|\theta), \theta\in\Theta\}$ Ex: Gaussien, famille exponentielle,...

Estimateur du Maximum de Vraisemblance

$$\boxed{\hat{p}_{MV}(x^n) = \arg\max_{p\in\mathcal{H}} p(x^n)}$$

- ***Modèle paramétrique*** :

$$\boxed{\hat{\theta}_{MV}(x^n) = \arg\max_{\theta\in\Theta} p(x^n|\theta)} \quad \hat{p}_{MV}(x^n) = p(x^n|\hat{\theta}_{MV}(x^n))$$

Propriétés asymptotiques

- Estimateur consistant

$$x^n \in \mathcal{X}^\infty, x^n \sim p^* \Rightarrow \lim_{n\to\infty} \hat{p}_{MV}(x^n) = p^* \quad w.p.1$$

$$x^n \in \mathcal{X}^\infty, x^n \sim p(x^n|\theta^*) \Rightarrow \lim_{n\to\infty} \hat{\theta}_{MV}(x^n) = \theta^* \quad w.p.1$$

- Code universel (par rapport à un modèle)

$\mathcal{H}$: modèle probabiliste $\Leftrightarrow$ $\mathcal{L}$ ensemble de (longueurs de) codes (de préfixe)

$\underline{L}_{\mathcal{H}}$ est un *code universel pour* $\mathcal{H}$ ssi

$$\forall \; x^n \in \mathcal{X}^\infty \; \lim_{n\to\infty} 1/n \, \underline{L}_{\mathcal{H}}(x^n) = \lim_{n\to\infty} 1/n \min_{L\in\mathcal{L}} L(x^n)$$

Note: si $x^n \sim p^* \in \mathcal{H}$ $\lim_{n\to\infty} 1/n \min_{L\in\mathcal{L}} L(x^n) = H(p^*)$: taux d'entropie

Pénalité

d'un code/modèle (p) par rapport à un modèle $\mathcal{H}$ (ensemble de codes/modèles)

Pénalité

$$\boxed{\mathcal{P}_{p,\mathcal{H}}(x^n) = -\log p(x^n) - \min_{q \in \mathcal{H}} (-\log q(x^n))}$$

- ***Modèle paramétrique***

$$\mathcal{P}_{p,\mathcal{H}}(x^n) = -\log p(x^n) + \log p(x^n | \hat{\theta}_{MV}(x^n))$$

Pénalité au pire cas

$$\mathcal{P}_{p,\mathcal{H}} = \max_{x^n \in \mathcal{X}} \mathcal{P}_{p,\mathcal{H}}(x^n)$$

$$= \max_{x^n \in \mathcal{X}} [-\log p(x^n) - \min_{q \in \mathcal{H}} (-\log q(x^n))]$$

Code universel optimal
(par rapport à un modèle)

- Code universel optimal

$\underline{L}_{\mathcal{H}^*}$ est un *code universel optimal (pour le modèle $\mathcal{H}$)* ssi

$$\boxed{\mathcal{P}_{\underline{L}_{\mathcal{H}^*},\mathcal{H}} \leq \mathcal{P}_{L,\mathcal{H}}}$$

Solution: *Code (modèle) de Shtarkov:*

$$p_{nmv}(x^n) = p_{\mathcal{H}^*}(x^n) \propto p(x^n \mid \hat{\theta}_{MV}(x^n)), \quad \int p_{\mathcal{H}}^*(x^n)\, d\, x^n = 1$$

Pour ce code,

$$\forall\, x^n \in \mathcal{X}^\infty\ \mathcal{P}_{p_{nmv},\mathcal{H}}(x^n) = \mathcal{P}_{p_{nmv},\mathcal{H}} = -\log \int p(x^n \mid \hat{\theta}_{MV}(x^n))\, d\, x^n$$

Principe du MDL

- Choix entre deux modèles $\mathcal{H}_1$ et $\mathcal{H}_2$:

Choisir le modèle pour lequel le code universel optimal conduit à une longueur de code minimale:

$$\boxed{\begin{array}{l} \underline{L}_{\mathcal{H}_1}{}^{*}(x^n) < \underline{L}_{\mathcal{H}_2}{}^{*}(x^n) \Rightarrow \text{choisir } \mathcal{H}_1 \\ \underline{L}_{\mathcal{H}_1}{}^{*}(x^n) > \underline{L}_{\mathcal{H}_2}{}^{*}(x^n) \Rightarrow \text{choisir } \mathcal{H}_2 \end{array}}$$

Avec la définition de code optimal (de Shtarkov) nous sommes conduits à un critère du type « *codage en deux parties* » :

$$\underline{L}_{\mathcal{H}_1}{}^{*}(x^n) = -\log p(x^n|\, \hat{\theta}_1(x^n)) + \log \int p(x^n|\hat{\theta}_1(x^n))\, dx^n$$

Complexité paramétrique

- Complexité paramétrique d'un modèle

$$C_n(\mathcal{H}) = \log \int p(x^n | \hat{\theta}(x^n))\, dx^n$$

Avec cette définition

$$\underline{L}_{\mathcal{H}_1}{}^{\star}(x^n) = -\log p(x^n | \hat{\theta}_1(x^n)) + C_n(\mathcal{H}_1)$$

$C_n(\mathcal{H}_1)$: codage du modèle *(structure)*

$-\log p(x^n | \hat{\theta}^1(x^n))$: codage des détails *(bruit)*

Test MDL

- Choix entre deux modèles $\mathcal{H}_1$ et $\mathcal{H}_2$:

$$\underline{L}_{\mathcal{H}_1}{}^{\star}(x^n) < \underline{L}_{\mathcal{H}_2}{}^{\star}(x^n) \Rightarrow \text{choisir } \mathcal{H}_1$$

$$\underline{L}_{\mathcal{H}_1}{}^{\star}(x^n) > \underline{L}_{\mathcal{H}_2}{}^{\star}(x^n) \Rightarrow \text{choisir } \mathcal{H}_2$$

$$\Leftrightarrow$$

$$-\log p(x^n|\hat{\theta}_1(x^n)) + C_n(\mathcal{H}_1) \gtrless -\log p(x^n|\hat{\theta}_2(x^n)) + C_n(\mathcal{H}_2) \text{ choisir } \mathcal{H}_1$$

$$\Leftrightarrow \quad \boxed{\log \frac{p(x^n|\hat{\theta}_1(x^n))}{p(x^n|\hat{\theta}_2(x^n))} \gtrless C_n(\mathcal{H}_1) - C_n(\mathcal{H}_2)}$$

Complexité paramétrique
($\mathcal{H}$: ensemble fini)

Si $\mathcal{H} = \{p(x^n|\theta_i), i=1,2,\ldots,M\}$

$$\Rightarrow \mathcal{C}_n(\mathcal{H}) = \log \sum_{x^n} p(x^n|\hat{\theta}(x^n)) = \log \sum_j \sum_{x^n:\, \hat{\theta}(x^n) = \theta_j} p(x^n|\theta_j)$$

$$= \log \sum_j (1 - \sum_{x^n:\, \hat{\theta}(x^n) \neq \theta_j} p(x^n|\theta_j))$$

$$= \log \left(\mathrm{M} - \Pr\{\hat{\theta}(x^n) \neq \theta_j\}\right)$$

$$\leq \log \mathrm{M}$$

Ces expressions montrent que la complexité paramétrique d'un modèle mesure le nombre de distributions que le modèle contient qui sont *distinguables* avec un *certain volume de données*

Dans l'expression précédente, le terme d'erreur tend (pour des modèles non pathologiques, pour lesquels un estimateur consistant existe) vers zéro quand le nombre de données tend vers infini, et $\mathcal{C}_n(\mathcal{H}) \to \log M$

Exemple: Bernoulli

$$p(x|\theta)=\theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}, \quad x_i\in\{0,1\},\ \Theta=[0,1]$$

$S_n=\sum_i x_i, \quad \hat{\theta}(x^n)=\frac{S_n}{n}$ $\quad$ S_n: sufficient statistic for θ

$$C_n(\mathcal{H})=\log\sum_{x^n} p(x^n|\hat{\theta}(x^n))=\log\sum_{s=0}^{n} p(S_n=s)\sum_{x^n:\hat{\theta}(x^n)=s/n} p(x^n|S_n=s)$$

$$=\log\sum_{s=0}^{n} p(S_n=s)=\log\sum_{s=0}^{n}\binom{n}{s}\left(\frac{s}{n}\right)^{s}\left(1-\frac{s}{n}\right)^{n-s}\cong\log\sum_{s=0}^{n}\frac{\sqrt{n}}{\sqrt{2\pi s(n-s)}}$$ (Stirling app.)

$$\cong\log\frac{\sqrt{n}}{\sqrt{2\pi}}\int_0^n\frac{1}{s(1-s)}ds+o(1)=\log\sqrt{\frac{n\pi}{2}}+o(1)=\frac{1}{2}\log\frac{n\pi}{2}+o(1)$$

Principe du MDL et RVG

- MDL:

$$\log p_1(x^n|\hat{\theta}_1(x^n)) / p_2(x^n|\hat{\theta}_2(x^n)) <?> C_n(\mathcal{H}_1) - C_n(\mathcal{H}_2)$$

Le test du MDL est un test du rapport de vraisemblance généralisé, où le seuil de décision est automatiquement fixé par la complexité paramétrique des modèles.

**** RVG : rapport de vraisemblance généralisé***

Consistance

Le fait que le code optimal soit un *code universel* garanti que *quand* $n \to \infty$ *le "vrai" modèle* (si les données sont une réalisation d'une source avec une distribution de probabilité qui fait partie d'un des modèles) *est choisi*, *avec probabilité 1*.

Note: *cette propriété est maintenue même si le code utilisé n'est pas le code optimal (la distribution de Shtarkov)*

Approximation asymptotique (MDL)

Sous certaines conditions, pour des modèles paramétriques, (*k fixe*, $n \to \infty$)

$$C_n(\mathcal{H}_\Theta) = \frac{k}{2}\log\frac{n}{2\pi} + \log\int |I(\theta)|^{1/2} d\theta + o(1)$$

où

k est la dimension du modèle paramétrique $\mathcal{H}_\Theta$ (comme variété différentiable)

n est le nombre d'observations

$I(\theta)$ est la **matrice (asymptotique) de Fisher**:

$$\lim_{n\to\infty} \frac{o(1)}{n} = 0 \qquad I_{ij}(\theta) = -\lim_{n\to\infty} \frac{1}{n} E\left\{ \frac{\partial^2}{\partial\theta_i\,\partial\theta_j} \log p(x^n|\theta) \right\}$$

Conditions sufisantes

- $C_n(\mathcal{H}_\Theta) < \infty$, $\int |I(\theta)|^{1/2}\, d\theta < \infty$
- $\hat{\theta}(x^{(n)})$ reste eloigné de la frontière de Θ.
- $\mathcal{H}$ est une famille exponentielle :

$$p(x|\theta) = exp(\theta\, t(x)) f(x) g(\theta)$$

t: $\mathcal{X} \to \mathbb{R}$ est une fonction de x

Exemples: Bernoulli, Gaussienne, Multinomial, Poisson, Gamma, … (mais pas les modèles de mélange)

Interprétation

- Avec cette approximation

$$\underline{L}_{\mathcal{H}}^{\star}(x^n) = -\log p(x^n|\hat{\theta}(x^n)) + C_n(\mathcal{H})$$

$$= -\log p(x^n|\ \ (x^n))$$ (fit to data (noise): ≅ linear in n)

$$+\frac{k}{2}\log\frac{n}{2\pi}$$ (# ≠ models: ≅ log in n)

$$+\log\int|I(\theta)|^{1/2}d\theta$$ (model geometry: ≅ C^{te} in n)

$$+o(1)$$ (→0 when $n\to\infty$)

Good approximation if *n large*, $k << n$

MDL et Bayes

Pour des modèles paramétriques

$$\mathcal{H}_i^{\Theta}=\{p(x^n|\theta_i),\ \theta_i\in\Theta_i\},\ i=1,2$$

l'approche Bayesienne considère connues des distributions *a priori*, $w_i(\theta_i)$, pour les paramètres inconnus θ_i de chaque modèle $\mathcal{H}_i^{\Theta}$, et choisit le modèle pour lequel la *distribution marginale*

$$p_{\mathcal{H}_i^{\Theta}}(x^n)=\int p(x^n|\theta_i)w_i(\theta_i)d\theta_i$$

est la plus grande :

$$p_{\mathcal{H}_1}(x^n)>p_{\mathcal{H}_2}(x^n)\Rightarrow \qquad \text{choisir } \mathcal{H}_1$$

La marginale de Bayes est un code universel

$$p_{\mathcal{H}}(x^n) = \sum_{\theta_i \in \Theta} p(x^n|\theta_i)w(\theta_i)$$

(**countable Θ**)

$$\Rightarrow -\log p_{\mathcal{H}}(x^n) = -\log\left[\sum_{\theta_i \in \Theta} p(x^n|\theta_i)w(\theta_i)\right] \leq -\log\left[p(x^n|\theta_j)w(\theta_j)\right]$$

$$\Rightarrow -\log p_{\mathcal{H}}(x^n) \leq -\log\left[p(x^n|\hat{\theta}(x^n))w(\hat{\theta}(x^n))\right] = -\log p(x^n|\hat{\theta}(x^n)) - \log w(\hat{\theta}(x^n))$$

(Bayes better than 2-part coding!)

$$\Rightarrow -\lim_{n\to\infty}\frac{1}{n}\log p_{\mathcal{H}}(x^n) = -\lim_{n\to\infty}\frac{1}{n}\log p(x^n|\hat{\theta}(x^n)) - \lim_{n\to\infty}\frac{1}{n}\log w(\hat{\theta}(x^n))$$

$$= -\lim_{n\to\infty}\frac{1}{n}\log p(x^n|\hat{\theta}(x^n))$$

$$= -\lim_{n\to\infty}\frac{1}{n}\max_{\theta\in\Theta}\log p(x^n|\theta)$$

Comportement asymptotique de Bayes

Pour des familles exponentielles (*expansion de Laplace*)

$$-\log p_{\mathcal{H}_i}(x^n) = -\log p(x^n|\hat{\theta}_i(x^n)) + \frac{k}{2}\log\frac{n}{2\pi}$$

$$-\log w_i(\hat{\theta}_i(x^n)) + \log|I(\hat{\theta}_i(x^n))|^{1/2} + o(1)$$

Pour $n >> 1$

$$\boxed{-\log p_{\mathcal{H}_i}(x^n) \cong -\log p(x^n|\hat{\theta}_i(x^n)) + \frac{k}{2}\log\frac{n}{2\pi}}$$

Bayes et MDL coincident avec BIC (Bayesian Information Criterion, Schwartz)

Jeffrey’s prior, Bayes et MDL

Pour les distributions *a priori* de Jeffrey:

$$w_i(\theta_i) = \frac{|I(\theta_i)|^{1/2}}{\int |I(\phi)|^{1/2} d\phi}$$

alors :

MDL $\equiv$ Bayes *(up to order 1)*

Note: MDL et Bayes sont des approches différentes: MDL n’est pas basé sur des supposions sur la vraie distribution des données, ce que n’est pas le cas pour Bayes!

MDL et codage prédictif

La factorisation

$$p(x^n|\theta)=\prod p(x_i\,|x_1^{i-1},\theta)$$

implique

$$\underbrace{-\log p(x^n|\theta)}_{\text{longueur de code}}=\sum_{i=1}^{n}\boxed{-\log p(x_i\,|x_1^{i-1},\theta)}$$

longueur de code

pénalité de prédiction accumulée

$-\log p_{\mathcal{H}}(x_i\,|x_1^{i-1}) \rightarrow -\log p(x_i\,|\hat{\theta}(x_1^{i-1}))$: “coût” de la prédiction de x_i basée sur l’observation de $x_1 \ldots x_{i-1}$

$$-\log p_{\mathcal{H}}(x^n)=-\log p(x^n|\hat{\theta}_i\,(x^n))+\frac{k}{2}\log\frac{n}{2\pi}+o(1)$$

Pointers pour en savoir plus

- MDL “idéal” et Complexité de Kolomogorov
  - Vytanyi (Amsterdam, http://homepages.cwi.nl/~paulv/)
- MDL avec complexité paramétrique infinie
  - Rissanen (Helsinki), T. Cover (Stanford, http://yreka.stanford.edu/~cover), Grunwald (Amsterdam, http://homepages.cwi.nl/~pdg/)
- Interprétation géométrique de la complexité paramétrique
  - Balasubramanian ((UPenn, Philadelphia, http://perception.upenn.edu/faculty/pages/balasubramanian.php)

Code unniversel pour les entiers

Pour coder un entier $k \in \{1,..,M\}$ on a besoin de

$$n = \lceil \log k \rceil \text{ bits} : k \rightarrow c_n(k) \in \{0,1\}^n$$

Pour coder un entier $k \in \{1,\ldots\}$?

$$k \rightarrow C_u(k) = 0^{\lceil \log k \rceil} 1\, c_{\lceil \log k \rceil}(k) \in \{0,1\}^{2n+1}$$

C_u est un code universel pour les entiers