Théorie de l'Information et Statistique

Plan

- Méthode des types
- Tests d'hypothèses : le Lemme de Stein
- estimation paramétrique

Méthode des types

$X_1, \dots X_n$, $X_i \in \mathcal{X} = \{1, \dots, m\}$, $|\mathcal{X}| = m$

Type

P_x, le type de la séquence $x = x_1, \dots, x_n$ mesure la fréquence de chaque symbole de $\mathcal{X}$ dans la séquence x:

$$P_x(i) = \frac{N_i(x)}{n}, \qquad N_i(x) = |\{x_k = i\}|, \qquad i = 1, \dots, m$$

P_x est une loi de probabilité (aussi appelé ”loi empirique" de x).

$\mathcal{P}_n$

$\mathcal{P}_n$ est l'ensemble de types qui peuvent être obtenus à partir d'une séquence de longueur n :

$$\begin{aligned}\mathcal{P}_n &= \Big\{P = (p(1), \cdots, p(m)) \ : \\ &\quad p(i) = \frac{n_i}{n}, n_i \in \{0, \ldots, m\}, \sum_i n_i = n\Big\}\end{aligned}$$

$$|\mathcal{P}_n| \leq (n+1)^m.$$

Le nombre de types est polynomial en $n \Rightarrow$ le nombre de séquences avec un type donné doit être exponentiel

Exemple

Pour $\mathcal{X} = \{a, b, c\} (m = 3)$, et $n = 3$, les possibles types sont

$$
\begin{aligned}
\mathcal{P}_n &= \{[1\ 0\ 0]\ , \\
&= [2/3\ 1/3\ 0]\ , \\
&= [2/3\ 0\ 1/3]\ , \\
&= [1/3\ 1/3\ 1/3]\ , \\
&= [1/3\ 2/3\ 0]\ , \\
&= [1/3\ 0\ 2/3]\ , \\
&= [0\ 1\ 0]\ , \\
&= [0\ 2/3\ 1/3]\ , \\
&= [0\ 1/3\ 2/3]\ , \\
&= [0\ 0\ 1]\}
\end{aligned}
$$

Classe du type P, $T(P)$

$T(P)$ est l'ensemble de séquences qui ont le type P :

$$T(P) = \{x \in \mathcal{X}^n \ : P_x = P\}$$

Exemple

Pour $\mathcal{X} = \{a, b, c\}(m = 3)$, et $n = 3$, la classe du type $p = [2/3\ 0\ 1/3]$ est

$$T([2/3\ 0\ 1/3]) = \{aac, aca, caa\}$$

Theorème

La probabilité d'une séquence iid dépend uniquement de son type: si $x_1, \ldots x_n \sim Q$

$$Q(x) = 2^{-n(H(P_x)+D(P_x||Q))}$$

$\Rightarrow x \in T(Q) \Rightarrow Q(x) = 2^{-nH(Q)}$.

Theorème

$\forall P \in \mathcal{P}_n$, $x_i \in \mathcal{X}$,

$$\frac{1}{(n+1)^m} 2^{nH(P)} \leq |T(P)| \leq 2^{nH(P)}$$

La taille des classes de chaque type a une dépendance exponentielle en n.

Exemple (cont.)

Pour $\mathcal{X} = \{a, b, c\}(m = 3)$, et $n = 3$, nous avons vu que la taille de la classe du type $p = [2/3\ 0\ 1/3]$ est égale à 3. Pour cette loi,

$$H(p) = \frac{2}{3}\log_2(3/2) + \frac{1}{3}\log_2(3) = 0.9183$$

et nous pouvons vérifier

$$0.1055 = \frac{1}{(n+1)^m}2^{nH(P)} \leq 3 = |T(P)| \leq 2^{nH(P)} = 6.75$$

Theorème

$\forall P \in \mathcal{P}_n,\, x_i \in \mathcal{X},\, x_i \sim Q$

$$\frac{1}{(n+1)^m} 2^{-nD(P||Q)} \leq \Pr\{T(P)\} \leq 2^{-nD(P||Q)}$$

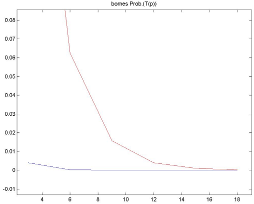
$\Rightarrow$ pour $P \neq Q, \Pr\{T(P)\} \rightarrow 0$

Exemple (cont.)

Pour $\mathcal{X} = \{a, b, c\}(m = 3)$, et $n = 3$, soit $T(p)$ la classe du type $p = [2/3\ 0\ 1/3]$. Si $X_i \sim Q = [1/3 1/3 1/3]$, alors

$$D(p||Q) = \frac{2}{3}$$

et nous obtenons la variations suivante pour les bornes inférieure et supérieur du théorème du transparent précédant :



Theorème de Sanov

$x_i \in \mathcal{X}$, $x_i \sim Q$ (iid). $E \subset \mathcal{P}$ un ensemble de lois de probabilité. Alors

$$Q^n(E) = Q^n(E \cap \mathcal{P}_n) \leq (n+1)^m 2^{-nD(P^\star||Q)}$$

où

$$P^\star = \arg\min_{P \in E} D(P||Q)$$

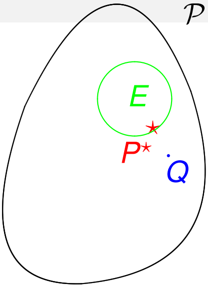
Si E coincide avec la fermeture de son intérieur, alors

$$\frac{1}{n} \log Q^n(E) \to -D(P^\star||Q)$$

Ce théorème affirme que la probabilité d'un ensemble E est dominée par l'entropie relative entre l'élément de E plus proche (au sens de l'entropie relative) de la vraie distribution des données.

Le résultat reste valable pour des variables aléatoires continues.

Illustration du théorème de Sanov



$$Q^n(E) \leq (n+1)^m 2^{-nD(P^\star||Q)}$$

Theorème

$E \subset \mathcal{P}$ un ensemble *convexe* de lois de probabilité, et $Q \notin E$. Soit

$$P^\star = \arg\min_{P \in E} D(P||Q)$$

Alors, $\forall P \in \mathcal{P}$,

$$D(P||Q) \geq D(P||P^\star) + D(P^\star||Q)$$

Ce théorème est une sorte de ""théorème de Pythagore" pour l'entropie relative (qui n'est pas une distance!).

Lemme

$$D(P||Q) \geq \frac{1}{2\ln 2}\|P - Q\|_1^2$$

Ce Lemme affirme que convergence en entropie relative implique convergence dans la norme L_1.

Théorème

$E \subset \mathcal{P}$ un ensemble *convexe* de lois de probabilité, et $Q \notin E$. $x_i \in \mathcal{X}$, $i = 1, \ldots, n$, $x_i \sim Q$ (iid). Soit

$$P^\star = \arg\min_{P \in E} D(P||Q)$$

Alors,

$$\Pr\{x_1 = a | P_x \in E\} \to P^\star(a)$$

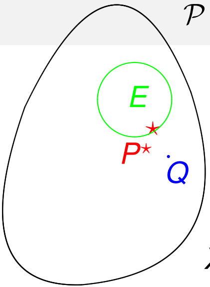
en probabilité quand $n \to \infty$.

Une version plus forte de ce théorème affirme que

$$\Pr\{x_1 = a_1 \cdots x_m = a_m | P_x \in E\} \to \prod_{i=1}^{m} P^\star(a_i)$$

en probabilité, quand $n \to \infty$.

Illustration du théorème conditionnel



$$X \sim Q, \qquad X \in \mathcal{X}$$

$$\Pr\left(x_1 = a | P_x \in E\right) = P^{\star}(a), \qquad a \in \mathcal{X}$$

Problème

$x_i \in \mathcal{X}, i = 1, \ldots, n$, $x_i \sim Q$ (iid).

$$\begin{aligned} H_1 &: Q = P_1 \\ H_2 &: Q = P_2 \end{aligned}$$

Déterminer $A \subset \mathcal{X}^n$ tel que

$$g_A(x) = \begin{cases} 1, & x \in A \\ 2, & x \in A^c \end{cases}$$

minimize

$$\alpha_g = \Pr\{g(x) = 2 | H_1\} = P_1(A^c)$$

sous la contrainte

$$\beta_g = \Pr\{g(x) = 1 | H_2\} = P_2(A) \leq \beta^\star$$

Solution

Soit $A_n(T)$ le sous-ensemble de $\mathcal{X}^n$

$$A_n(T) = \left\{ x \in \mathcal{X}^n \ : \ \frac{P_1(x)}{P_2(x)} > T \right\}$$

et $g_{A_n(T)}$ la fonction de décision qui lui est associée. Alors, $\forall A \subset \mathcal{X}^n$,

$$\alpha_{g_A} \leq \alpha_{g_{A_n(T)}} \to \beta_{g_A} \geq \beta_{g_{A_n(T)}}$$

Ce théorème affrirme que le détecteur optimal (de Neyman-Pearson) est un test du rapport de vraisemblance.

Solution (en fonction d'entropies relatives)

L'ensemble $A_n(T)$ qui défini le test topimal peut être défini comme

$$A_n(T) = \left\{ x \in \mathcal{X}^n \ : \ D(P_x||P_2) - D(P_x||P_1) > \frac{1}{n} \log T \right\}$$

Cette expression montre que le test optimal peut être identifié avec une région de décision dans l'espace des types.

lemme de Stein

$x_i \in \mathcal{X}, i = 1, \ldots, n$, $x_i \sim Q$ (iid). Admettons que dans le test de H_1 *vs* H_2, $D(P_1||P_2) < \infty$. Soit $A_n \subset \mathcal{X}^n$, et pour $\epsilon \in]0, 0.5[$ soit

$$\beta_n^\epsilon = \min_{\alpha_{g_{A_n}} < \epsilon} \beta_{g_{A_n}}$$

Alors

$$\lim_{\epsilon \to 0} \lim_{n \to \infty} \frac{1}{n} \log \beta_n^\epsilon = -D(P_1||P_2)$$

Ce théorème nous donne le meilleur taux de diminution de l'erreur (β) quand l'autre erreur (α) tend lentement vers zéro.

Borne de Chernoff

La borne de Chernoff caractérise le comportement des *test Bayésiens*.
$x_i \in \mathcal{X}, i = 1, \ldots, n$, $x_i \sim Q$ (iid). Admettons que dans le test de H_1 *vs* H_2, $D(P_1||P_2) < \infty$), et $P(H_1) = \pi_1$, $P(H_2) = \pi_2$.
Soit

$$D^\star = \lim_{n\to\infty} \min_{A\subset\mathcal{X}^n} -\frac{1}{n}\log P_e$$

où P_e est la probalité d'erreur.
Alors, le meuilleur taux de diminution de l'erreur est

$$D^\star = D(P_{\lambda^\star}||P_1) = D(P_{\lambda^\star}||P_2)$$

Dans le théorème précédent

$$P_\lambda(x) = \frac{P_1^\lambda(x)P_2^{1-\lambda}(x)}{\sum_{a\in\mathcal{X}} P_1^\lambda(a)P_2^{1-\lambda}(a)}, \qquad x \in \mathcal{X}$$

et $\lambda^\star$ est déterminé par (information de Chernoff)

$$D(P_{\lambda^\star}||P_1) = D(P_{\lambda^\star}||P_2)$$