Notes du CM6

Statistique inférentielle non asymptotique

Scribes : **Alexis HORDE, Ioana IVAN**
UE : Mathématiques pour l'intelligence artificielle
Date : 23/09/2022
Sources :
CM6, Thomas Schatz

1. Notations

— Ensemble des observations : O
— Les observations O suivent une distribution P : $O \sim P$
— La distribution P est un modèle statistique de $\mathcal{P}$: $P \in \mathcal{P}$
— En apprentissage, on cherche à définir un coût sur un objet d'intérêt (à optimiser), c'est la fonctionnelle : $f : \mathcal{P} \to E$
— Fonction estimée : $\hat{f}$

2. Estimations ponctuelle

2.1. Formalisme

$f : \Omega \to E$ tel que $\hat{f} \approx f(P)$

2.2. Exemple : estimation ponctuelle de la moyenne

RAS

2.3. Evaluation de l'estimation ponctuelle

2.3.1. Biais, variance, risque

— Biais : $b_P(\hat{f}) = E_{O\sim P}\left[\hat{f}(O)\right] - f(P)$
— Variance : $var_P(\hat{f}) = var_{O\sim P}\left[\hat{f}(O)\right]$
— Risque $\ell : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ avec $R_P(\hat{f}) = E_{O\sim P}\left[\ell\left(f(P), \hat{f}(O)\right)\right]$

2.4. Exemple : calcul de biais, variance, risque

2.4.1. Énoncé

Soit un échantillon de $\mathbb{R}^n$: $(X_1, ..., X_n) \sim_{i.i.d.} \mathcal{P}$. Prenons deux caractéristiques de cette distribution qu'on aimerait calculer, qu'on va noter f_1 et f_2. La fonction $f_1(P)$ estime la moyenne empirique de la distribution $\mathcal{P}$ et $f_2(P)$ estime la variance empirique.

$$f_1(P) = E[P] = \mu$$
$$f_2(P) = var(P) = \sigma^2$$

Les expressions de la moyenne empirique et de la variance empirique sont données ci-dessous :

$$\hat{\mu}(X_1, ..., X_n) = \frac{1}{n}\sum_{i=1}^{n} X_n$$

$$\hat{s}(X_1, ..., X_n) = \frac{1}{n}\sum_{i=1}^{n} (X_i - \hat{\mu}(X_1, ..., X_n))^2$$

On cherche à calculer le biais, la variance et le risque pour $\hat{\mu}$ et $\hat{s}$.
Le risque est calculé comme suit :

$$l\left(f(P), \hat{f}(O))\right) = E_{O\sim P}\left[f(P) - \hat{f}(O)\right]^2$$

On admet :

$$var(\hat{s}) = \frac{X_{X\sim P}(X-\mu)^4}{n} - var(P)\frac{n-3}{n(n-1)}$$

2.4.2. Pour la moyenne empirique

Calcul du biais

$$\begin{aligned} b_P(\hat{\mu}) &= E_{(X_1,...,X_n)\sim_{i.i.d.}\mathcal{P}}\left[\hat{\mu}(X_1, ..., X_n)\right] - E\left[P\right] \\ &= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] - \mu \\ &= \frac{1}{n}\sum_{i=1}^{n} E[X_i] - \mu \\ &= \left(\frac{1}{n}\sum_{i=1}^{n} \mu\right) - \mu \\ &= 0 \end{aligned}$$

Calcul de la variance

$$var_P(\hat{\mu}) = var\left(\hat{\mu}\right) = var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n} var\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, X_j)$$
$$= \frac{1}{n^2}\sum_{i=1}^{n} var(X_i) = \frac{\sigma^2}{n}$$

Calcul du risque

$$R_p(\hat{\mu}) = E\left[(\mu - \hat{\mu})^2\right]$$

$$b_P(\hat{\mu})^2 = E\left[\mu - \hat{\mu}\right]^2$$

$$var(\mu) = E\left[\hat{\mu}^2\right] - E\left[\hat{\mu}\right]^2$$

$$b_P(\hat{\mu})^2 = \left(E\left[\mu\right] - E\left[\hat{\mu}\right]\right)^2 = E\left[\mu\right]^2 + E\left[\hat{\mu}\right]^2 - 2E\left[\mu\right]E\left[\hat{\mu}\right]$$

$$\begin{aligned}
R_p(\hat{f}) &= E_{O\sim P}\left[\left(f(P) - \hat{f}(O)\right)^2\right] \\
&= E_{O\sim P}\left[\left(f(P) - E\left[\hat{f}(O)\right] + E\left[\hat{f}(O)\right] - \hat{f}(O)\right)^2\right] \\
&= E_{O\sim P}\left[\left(f(P) - E\left[\hat{f}(O)\right]\right)^2\right] + E\left[\left(E[\hat{f}(O)] - \hat{f}(O)\right)^2\right] \\
&+ 2E\left[f(P) - E[\hat{f}(O)]\right]\left(E[\hat{f}(O)] - \hat{f}(O)\right) (= 0 \text{ par linéarité de l'espérance})
\end{aligned}$$

En distribuant l'espérance ($E[E[X] = E[X]]$) et en utilisant que l'espérance d'un produit est égal au produit d'espérances ($E[AB] = E[A]E[B]$), on montre que le risque se décompose en biais et variance :

$$R_p(\hat{f}) = b_P(\hat{f})^2 + var_p(\hat{f})$$

2.4.3. Pour la variance empirique

Calcul du biais

$$b_P(\hat{s}) = E[\hat{s}] - \sigma^2$$

On applique la formule de la variance $var(x) = E[x^2] - E[x]^2 = \sigma^2$

$$b_P(\hat{s}) = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right] - \sigma^2$$

Le premier terme est égal à $E_{O\sim P}[\hat{f}(O)]$ et le deuxième à $f(P)$.

$$b_P(\hat{s}) = \frac{1}{n}\sum_{i=1}^{n} E[X_i^2] - E\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j\right] - \sigma^2$$

Mais X_iX_j sont i.i.d. sauf si $i = j$:

$$\begin{aligned} b_P(\hat{s}) &= E[X_1^2] - \left(\frac{1}{n^2} \sum_{i=1, i\neq j}^{n} \sum_{j=1, i\neq j}^{n} \mu^2 + \frac{1}{n^2} \sum_{i=1}^{n} E[X_1^2] \right) - \sigma^2 \\ &= E[X_1^2] - \frac{\mu^2}{n^2} n(n-1) - \frac{E[X_1^2]}{n} - \sigma^2 \\ &= E[X_1^2] \left(1 - \frac{1}{n}\right) - \frac{n-1}{n} \mu^2 - \sigma^2 \\ &= \left(1 - \frac{1}{n}\right) \left(E[x_i^2] - \mu^2\right) - \sigma^2 \\ &= \left(1 - \frac{1}{n}\right) \sigma^2 - \sigma^2 = -\frac{\sigma^2}{n} < 0 \text{ Toujours une sous-estimation} \end{aligned}$$

On peut travailler avec l'estimateur non-biaisé :

$$\hat{s}_U(X_1, ..., X_n) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \mu)^2$$

$$b_P(\hat{s}) = E[k\hat{s}] - \sigma^2 = kE[\hat{s}] - \sigma^2 = \left[k \frac{n-1}{n} - 1 \right] \sigma^2$$

$$k = \frac{n}{n-1}$$

$$var(k\hat{s}) = k^2 var(\hat{s}) = \frac{n^2}{(n-1)^2} var(\hat{s})$$

À savoir qu'un estimateur non-biaisé n'a pas nécessairement un risque minimal.

Calcul de la variance
Non fait en cours

Calcul du risque
Non fait en cours

2.5. Remarques

- En grande dimension, on peut :
  - Soit travailler sur des distributions plus petites
  - Soit contraindre (c'est-à-dire se restreindre à un espace de fonction pré-défini par exemple)
- Compromis biais-variance :
  - En grande dimension, des estimateurs naturels peuvent être biaisés
  - On préfère utiliser des estimateurs non biaisés pour définir des estimateurs optimaux
  - On va travailler sur un compromis de biais/variance afin de limiter le risque
    - Phénomène de Stein (régularisation) : on introduit du biais pour réduire le risque
    - Souvent, on part d'un estimateur non biaisé optimal puis on rajoute du biais pour diminuer la variance

2.6. Outils pour la construction d'estimateurs ponctuels

2.6.1. Suffisance

Intuition

Si on nous donne la statistique suffisante, on sait tout ce qu'il faut sur la distribution de probabilité $\mathcal{P}$ (par exemple, la moyenne empirique).

Définition

Soit X un exemple d'une population inconnue $P \in \mathcal{P}$ où P est une famille de populations. Une statistique $T(X)$ est suffisante pour $P \in \mathcal{P}$ (ou pour $\theta \in \Theta$ si $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$) si et seulement si les distributions conditionnelles de X sachant T est connue (indépendantes de P ou θ).

2.6.2. Exemple

Soit $(X_1, ..., X_n) \sim_{iid} \mathcal{P}$. On définit les statistiques d'ordre (le plus petit $X_{(1)}$, puis le deuxième plus petit $X_{(2)}$, puis le dernier plus petit $X_{(n)} = \max(X_1, ..., X_n)$)

On montre que les statistiques d'ordre (T) sont des statistiques suffisantes.

En effet, soit $P(X_1 = b1, ..., X_n = b_n | X_{(1)} = a_1, ..., X_{(n)} = a_n)$. Alors $b_1, ..., b_n$ doivent être des permutations de $a_1, ..., a_n$ et toutes les permutations sont équiprobables car iid. De ce fait, la probabilité conditionnelle est indépendante de $X_{(1)} = a_1, ..., X_{(n)} = a_n$. [A approfondir ...]

Intérêt : on obtient un estimateur bête $\hat{f}(X_1, ..., X_n)$ que l'on peut moyenner afin d'obtenir un meilleur estimateur $E\left[\hat{f}(X_1, ..., X_n)|T\right] = \frac{1}{n}\sum_{i=1}^n X_i$

2.6.3. Maximum de vraisemblance

Définitions et propriétés

Pour $X \sim P_\theta$, $\mathcal{P} = \{P_\theta | \in \Theta\}$:

— Maximum de vraisemblance : $\ell(\theta) = P_\theta(X)$

— Estimateur du maximum de vraisemblance : $\theta^* \in \arg\max_{\theta \in \Theta} \ell(\theta)$

— Comme log est croissante, trouver le maximum de vraisemblance avec θ est équivalent à résoudre le problème avec $\ell\ell(\theta) = \log \ell(\theta)$

Exemple

Soient $(X_1, ..., X_n)$ suivant une distribution iid. $\mathcal{N}(\sigma, \mu^\in) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. Les paramètres sont $\theta = (\mu, \sigma)$

On a :

1. $\ell_X(\theta) = P_{\mu,\sigma}(X) = \prod_{i=1}^n \mathcal{N}(X_i, \mu, \sigma^2)$
2. $\log \ell_X(\theta) = \sum_{i=1}^n \frac{1}{2}\log(\sqrt{\pi}) \log(\sigma) - \frac{(x_i - \mu)^2}{2\sigma^2}$

Par la suite, on va calculer le gradient de $\log \ell_X(\theta)$ puis l'annuler afin d'optenir les paramètres optimaux (qui sont les estimateurs du maximum de vraisemblance).

3. Estimations par intervalles

3.1. Modélisation

$R : \Omega \to 2^{Im(f)}$ tel que $p_{O\sim P}\left(f(P) \in R(O)\right) \approx 1 - \alpha$

Remarque : le classifieur optimal ne donne pas nécessairement un risque nul (exemple : erreur de Bayes)

3.2. Exemple : estimation par intervalle de la moyenne

On reprend le même exemple que dans 2.2.
On peut définir un intervalle de confiance pour la moyenne du type :
$R(O) = \left[\hat{\mu} - k(\alpha\sqrt{\hat{s}(O)}); \hat{\mu} + k(\alpha\sqrt{\hat{s}(O)})\right]$
Avec :

- Estimation de la variance empirique : $\hat{s}(O) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \hat{\mu}(O)\right)^2$
- $k(alpha)$: quantile de la distribution gaussienne

Idée de la démonstration :

- Utiliser le théorème central limite sur la variable centrée réduite de la moyenne puis encadrer à droite et à gauche dans $P(... \leq ... \leq ...)$ au risque α
- Transformer l'expression pour encadrer la moyenne : on se retrouve avec l'expression définie ci-dessus pour le cas asymptotique
- Adaptations : étudier aussi en passant par Student

4. Etude des queues de distribution

On définit ici trois inégalités de concentration :

4.1. Inégalité de Markov

$$\forall t \geq 0, p\left(X \geq t\right) \leq \frac{E[X]}{t}$$

4.2. Inégalité de Hoeffding

4.2.1. Formule

Pour $X_i \in [a, b]$ iid, $\forall t \geq 0, p\left(|\sum_{i=1}^{n}\left(X_i - \mu\right)| \geq t\right) \leq 2\exp -\frac{2t^2}{n(b-a)^2}$

4.2.2. Exemple

Avec $t = ns$ (n taille de l'échantillon), on rerouve :

$$p\left(\underbrace{\frac{1}{n}\left|\sum_{i=1}^{n}(X_i-\mu)\right|}_{\text{Moy. emp. } \mu} \geq s\right) \leq 2\exp -\frac{2(ns)^2}{n(b-a)^2}$$

Lecture de l'inégalité de concentration : les queues tendent vers 0 avec une vitesse en $\exp(-n)$.