

Cours 0 : Introduction à l'Estimation Statistique

Apprentissage Statistique / Statistiques Avancées

Résumé

Ce cours introductif pose les fondations mathématiques des statistiques paramétriques et non paramétriques. Nous rappelons les principaux paradigmes de l'inférence statistique, passons en revue les méthodes classiques d'estimation paramétrique (Maximum de Vraisemblance et Méthode des Moments), et introduisons les concepts centraux des statistiques non paramétriques et de la théorie de l'approximation.

1 Rappels et Cadre Général

Soit $(X_i)_{1 \leq i \leq N}$ un ensemble de variables aléatoires indépendantes et identiquement distribuées (i.i.d.). Nous supposons que les données sont générées par un processus dont la densité de probabilité (ou fonction de masse) est notée p_θ .

Ici, le paramètre d'intérêt est $\theta \in \Theta$, où $\Theta \subseteq \mathbb{R}^d$ ($d < \infty$). Le problème fondamental de l'estimation statistique est le suivant : *Étant donné l'observation du jeu de données $\{x_1, \dots, x_N\}$, comment trouver un estimateur de θ , noté $\hat{\theta}$?*

1.1 Résumé des Paradigmes Statistiques

Selon la nature de Θ et selon que θ est considéré comme déterministe ou comme une variable aléatoire, différents cadres mathématiques s'appliquent. Ces notions sont généralement abordées dans les cours d'introduction aux statistiques et à l'analyse numérique. Les principaux résultats sont résumés dans le Tableau ??.

	θ est déterministe	θ est aléatoire (possède une distribution a priori)
θ dans un ensemble discret ou fini (ex : $\theta \in \{0, 1\}$)	Tests d'hypothèses Lemme de Neyman-Pearson (Maximiser P_D sous la contrainte $P_{FA} \leq \alpha$)	Théorie de la décision - Maximum de Vraisemblance (MV) si θ est équiprobable. - Maximum A Posteriori (MAP) si les a priori sont inégaux.
θ dans un ensemble continu, dimension finie (ex : $\theta \in [0, 1]$)	Théorie de l'estimation - Aucun estimateur universellement optimal n'existe. - Les performances sont bornées inférieurement par la <i>Borne de Cramér-Rao</i> (BCR).	Approche bayésienne - Estimateur de l'Erreur Quadratique Moyenne Minimale (MMSE) : $\hat{\theta} = \mathbb{E}[\theta X]$ - Les performances sont bornées par la BCR bayésienne.

TABLE 1 – Aperçu des cadres statistiques

2 Statistiques Paramétriques

En statistiques paramétriques, nous supposons que la distribution sous-jacente appartient à une famille connue régie par un paramètre déterministe de dimension finie $\theta \in \mathbb{R}^d$.

2.1 Estimation par le Maximum de Vraisemblance (MV)

La fonction de vraisemblance $L(\theta; x_1, \dots, x_N)$ représente la probabilité conjointe d'observer les données sachant le paramètre θ . En raison de l'hypothèse i.i.d., elle se factorise ainsi :

$$L(\theta; x_1, \dots, x_N) = p_\theta(x_1, \dots, x_N) = \prod_{n=1}^N p_\theta(x_n) \quad (1)$$

L'Estimateur du Maximum de Vraisemblance (EMV) est la valeur du paramètre qui maximise cette fonction :

$$\hat{\theta}_{MV} = \arg \max_{\theta \in \mathbb{R}^d} L(\theta; x_1, \dots, x_N) \quad (2)$$

En pratique, il est strictement équivalent et numériquement beaucoup plus stable de maximiser la log-vraisemblance, $l(\theta) = \log L(\theta)$ (en utilisant le logarithme népérien, de base e) :

$$\hat{\theta}_{MV} = \arg \max_{\theta \in \mathbb{R}^d} l(\theta; x_1, \dots, x_N) \quad \text{où} \quad l(\theta; x_1, \dots, x_N) = \sum_{n=1}^N \log p_\theta(x_n) \quad (3)$$

Exemple : Loi de Bernoulli

Soit p_θ une loi de Bernoulli de paramètre $\theta \in [0, 1]$. Soient $x_1, \dots, x_N \in \{0, 1\}$ des réalisations i.i.d. La fonction de masse pour une observation unique est $p_\theta(x_n) = \theta^{x_n}(1 - \theta)^{1-x_n}$.

La fonction de vraisemblance est :

$$L(\theta; x_1, \dots, x_N) = \prod_{n=1}^N \theta^{x_n}(1 - \theta)^{1-x_n} = \theta^{S_N}(1 - \theta)^{N-S_N}$$

où $S_N = \sum_{n=1}^N x_n$ est le nombre de succès.

La log-vraisemblance est :

$$l(\theta) = S_N \log(\theta) + (N - S_N) \log(1 - \theta)$$

Pour trouver le maximum, nous dérivons par rapport à θ et annulons la dérivée :

$$\frac{\partial l}{\partial \theta} = \frac{S_N}{\theta} - \frac{N - S_N}{1 - \theta} = 0 \implies S_N(1 - \theta) = \theta(N - S_N) \implies S_N = N\theta$$

Ainsi, l'estimateur du MV est la moyenne empirique : $\hat{\theta}_{MV} = \frac{1}{N} S_N$.

2.2 Méthode des Moments

La méthode des moments consiste à exprimer les moments théoriques de la distribution (qui sont des fonctions de θ) et à les évaluer aux moments empiriques de l'échantillon.

Exemple utilisant la mesure empirique

Soit X une variable aléatoire dont la distribution de probabilité peut s'écrire à l'aide de masses de Dirac : $p_\theta(x) = \theta \delta_1(x) + (1 - \theta) \delta_0(x)$. Le premier moment théorique (l'espérance) est :

$$\mathbb{E}_{p_\theta}[X] = \int x p_\theta(x) dx = 1 \cdot \theta + 0 \cdot (1 - \theta) = \theta$$

La distribution empirique basée sur N échantillons est donnée par $\widehat{p}(x) = \frac{1}{N} \sum_{n=1}^N \delta_{x_n}(x)$. L'espérance empirique est :

$$\widehat{\theta} = \mathbb{E}_{\widehat{p}}[X] = \int x \left(\frac{1}{N} \sum_{n=1}^N \delta_{x_n}(x) \right) dx = \frac{1}{N} \sum_{n=1}^N x_n$$

ce qui représente intuitivement le ratio du nombre de 1 sur le nombre total d'échantillons.

3 Statistiques Non Paramétriques

En statistiques non paramétriques, l'hypothèse selon laquelle le mécanisme générateur des données appartient à une famille paramétrique de dimension finie est abandonnée. L'objet d'intérêt est plutôt une fonction f appartenant à un espace fonctionnel de dimension infinie \mathcal{F} (par ex., $f \in \mathcal{F}$).

3.1 Travailler dans des Espaces de Dimension Infinie

Pour manipuler rigoureusement les espaces de dimension infinie, nous restreignons généralement \mathcal{F} à un **Espace de Hilbert**. Un espace de Hilbert généralise la notion d'espace euclidien ; il est muni d'un produit scalaire $\langle \cdot, \cdot \rangle$ qui induit une distance, et il possède une base dénombrable. Cela permet de représenter les fonctions via des décompositions sur une base (ex : séries de Fourier, ondelettes).

3.2 Théorie de l'Approximation et Compromis d'Erreur

Lorsque l'on tente d'estimer une fonction $f \in \mathcal{F}$ en utilisant un espace d'hypothèses restreint ou fini \mathcal{H} à partir de N échantillons finis, nous rencontrons deux sources principales d'erreur :

1. **L'erreur d'approximation (Biais) :** L'erreur introduite en restreignant notre recherche à un espace plus petit et plus simple \mathcal{H} plutôt qu'au véritable espace de dimension infinie \mathcal{F} . Elle mesure à quel point le meilleur modèle possible dans \mathcal{H} peut approcher la vraie fonction f .
2. **L'erreur d'estimation (Variance) :** L'erreur provenant du fait que nous ne disposons que d'un nombre fini N d'échantillons pour trouver la fonction optimale au sein de \mathcal{H} .

Ces deux erreurs conduisent au fondamental **compromis Biais-Variance**. Augmenter la complexité de l'espace d'hypothèses diminue l'erreur d'approximation mais augmente l'erreur d'estimation, et inversement.

3.3 Approche Minimax

Pour dériver des bornes sur ces erreurs lorsque N est fini et que la dimensionnalité est potentiellement élevée, les statisticiens utilisent souvent l'**approche Minimax**. Cela implique de trouver l'estimateur qui minimise le risque maximum possible (erreur espérée) sur la pire distribution possible dans la classe \mathcal{F} .

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}[L(\widehat{f}, f)]$$

4 Problèmes Classiques en Non Paramétrique

4.1 Estimation de Densité

Soit X une variable aléatoire ayant une Fonction de Répartition (FR) notée F et une Densité de Probabilité (DP) notée f . Par définition :

$$F(A) = \mathbb{P}(X \in A) = \int_A f(x)dx$$

Pour $A = (-\infty, x]$, la fonction de répartition est :

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f(u)du$$

Objectif : Étant donné des échantillons $\{x_1, \dots, x_N\}$, trouver un estimateur $\hat{f}(x; x_1, \dots, x_N)$ qui approche la vraie densité $f(x)$ partout. (Les méthodes classiques incluent les histogrammes et l'estimation par noyaux).

4.2 Régression Non Paramétrique

Considérons des observations appariées (X, Y) où la relation est régie par :

$$Y = f(X) + \varepsilon$$

Ici, ε est un bruit aléatoire centré (de moyenne nulle) et indépendant de X , tel que $\mathbb{E}[\varepsilon] = 0$. **Objectif :** Étant donné un jeu de données de paires $\{(X_n, Y_n)\}_{1 \leq n \leq N}$, estimer la fonction inconnue $f(x)$ pour prédire les valeurs futures : $\hat{y} = \hat{f}(x)$.