Parameter estimation in supervised learning

M. Ndaoud

Previous session

- tools : LLN, CLT, Slutsky Lemma
- estimators : empirical cumulative function, empirical quantile
- graphical statistics : boxplot, qq-plot, heatmap
- Results convergence a.s. and speed of convergence of:

$$\widehat{F}_n, \quad \widehat{q}_{n,p}$$

So far we have note used the statistical model to construct estimators.

Statistical model (1/2)

Question : A model is a prior knowledge on data. How can we leverage this information in order to construct and study estimators that are "more efficient" than model-free estimators as $\widehat{F}_n, \widehat{q}_{n,p}, \ldots$?

Example of a statistical model (2/2)

Problem : A physicist observes the lifetime of radioactive atoms which he decides to model by random variables $X_1, \ldots, X_n$ i.i.d. He wishes to use these data to estimate their underlying law. He can choose between two approaches:

Example of a statistical model (2/2)

Problem : A physicist observes the lifetime of radioactive atoms which he decides to model by random variables $X_1, \ldots, X_n$ i.i.d. He wishes to use these data to estimate their underlying law. He can choose between two approaches:

- "model-free" : by estimating the cumulative function of X_i through $\widehat{F}_n$

Example of a statistical model (2/2)

Problem : A physicist observes the lifetime of radioactive atoms which he decides to model by random variables $X_1, \ldots, X_n$ i.i.d. He wishes to use these data to estimate their underlying law. He can choose between two approaches:

- "model-free" : by estimating the cumulative function of X_i through $\widehat{F}_n$
- "model-based" : he knows that lifetimes follow an exponential law $\in \{\mathcal{E}xp(\theta) : \theta > 0\}$. In this case, it is enough to estimate θ by an estimator $\widehat{\theta}_n$ and to approximate the distribution function of X_i by $F_{\widehat{\theta}_n}$ where

$$F_\theta(x) = \mathbb{P}[\mathcal{E}xp(\theta) \leq x] = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 - \exp(-\theta x) & \text{else.} \end{cases}$$

Statistical paradigm

1) Starting point : data (ex.: real numbers)

$$\mathtt{x}_1, \ldots, \mathtt{x}_\mathtt{n}$$

2) **Statistical modeling** :

- data are realizations

$$X_1(\omega), \ldots, X_n(\omega) \text{ of r.v. } X_1, \ldots, X_n.$$

(in other words, for a certain ω, $X_1(\omega) = \mathtt{x}_1, \ldots, X_n(\omega) = \mathtt{x}_\mathtt{n}$)

- The **distribution** $\mathbb{P}^{(X_1,\ldots,X_n)}$ of $(X_1, \ldots, X_n)$ is unknown, but belongs to a given family (a priori)

$$\boxed{\left\{ \mathbb{P}_\theta^n, \theta \in \Theta \right\}} \ : \text{ the model}$$

We believe that there exists $\theta \in \Theta$ such that $\mathbb{P}^{(X_1,\ldots,X_n)} = \mathbb{P}_\theta^n$.

- θ is the parameter and Θ the set of parameters.

Statistical paradigm (Cont'd)

Problem: from the "observation" $X_1, \ldots, X_n$

- **Modeling:** which model to choose ?

Statistical paradigm (Cont'd)

Problem: from the "observation" $X_1, \dots, X_n$

- **Modeling:** which model to choose ?
- Estimation : construct a function $\phi_n(X_1, \dots, X_n)$ that approximates the best θ

Statistical paradigm (Cont'd)

Problem: from the "observation" $X_1, \dots, X_n$

- **Modeling:** which model to choose ?
- Estimation : construct a function $\phi_n(X_1, \dots, X_n)$ that approximates the best θ
- Test : Establish a decision $\varphi_n(X_1, \dots, X_n) \in \{\text{set of decisions}\}$ concerning a hypothesis about θ.

Statistical paradigm (Cont'd)

Problem: from the "observation" $X_1, \ldots, X_n$

- **Modeling:** which model to choose ?
- Estimation : construct a function $\phi_n(X_1, \ldots, X_n)$ that approximates the best θ
- Test : Establish a decision $\varphi_n(X_1, \ldots, X_n) \in \{\text{set of decisions}\}$ concerning a hypothesis about θ.
- Prediction : Guess the unobserved value X_{n+1} based on $X_1, \ldots, X_n$

Example of head or tail

- We toss a coin 18 times and observe $(H = 0,\ T = 1)$

$$0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0$$

- statistical model : we observe $n = 18$ independent random variables X_i, Bernoulli of unknown parameter $\theta \in \Theta = [0, 1]$.

Example of head or tail

- We toss a coin 18 times and observe $(H = 0,\ T = 1)$

$$0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0$$

- statistical model : we observe $n = 18$ independent random variables X_i, Bernoulli of unknown parameter $\theta \in \Theta = [0, 1]$.
    - Estimation. Estimator $\bar{X}_{18} = \frac{1}{18}\sum_{i=1}^{18} X_i \overset{\text{here}}{=} 8/18 = 0.44$. What precision ?

Example of head or tail

- We toss a coin 18 times and observe $(H = 0,\ T = 1)$

$$0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0$$

- statistical model : we observe $n = 18$ independent random variables X_i, Bernoulli of unknown parameter $\theta \in \Theta = [0, 1]$.
  - Estimation. Estimator $\bar{X}_{18} = \frac{1}{18} \sum_{i=1}^{18} X_i \overset{\text{here}}{=} 8/18 = 0.44$. What precision ?
  - Test. Decision to make : "is the coin balanced ?". For example: we compare $\bar{X}_{18}$ to 0.5. If $|\bar{X}_{18} - 0.5|$ "small", we accept the hypothesis "the coin is balanced". Otherwise, we reject.
  - Prediction. If we toss the same coin a new time, is the outcome more likely to be head or tail?

Maximum Likelihood Estimation (MLE)

Sampling model (in $\mathbb{R}$)

- We observe a sample of size n of random variables $X_1, \ldots, X_n$.
- The distribution of X_i belongs to the parametric family $\{\mathbb{P}_\theta,\ \theta \in \Theta\}$ (family of distrubtions $\mathbb{R}$). We denote the densities : $\forall \theta \in \Theta, x \in \mathbb{R},\ f(\theta, x)$.
- The distribution of $(X_1, \ldots, X_n)$ is given by : $\forall x_1, \ldots, x_n \in \mathbb{R}$,

$$\boxed{\prod_{i=1}^{n} f(\theta, x_i)}$$

Example 1 : the normal model

$X_i \sim \mathcal{N}(m, \sigma^2)$, avec $\theta = (m, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+ \setminus \{0\}$.

- The normal density is given by:

$$f(\theta, x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(x-m)^2}{2\sigma^2} \Big)$$

- The corresponding distribution is given by : for all $x_1, \ldots, x_n \in \mathbb{R}$,

$$\prod_{i=1}^{n} f(\theta, x_i) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(- \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - m)^2 \Big)$$

Example 2 : Bernoulli model

$X_i \sim \text{Bernoulli}(\theta)$, with $\theta \in \Theta = [0, 1]$

- For all $x \in \{0, 1\}$

$$\boxed{f(\theta, x) = (1 - \theta)I(x = 0) + \theta I(x = 1) = \theta^x (1 - \theta)^{1-x}}$$

- The distribution of the observations has density:

$$\prod_{i=1}^{n} \theta^{x_i} (1 - \theta)^{1-x_i},$$

for $x_1, \ldots, x_n \in \{0, 1\}$

Maximum likelihood

- Fundamental and essential principle in statistics. Known special cases since the 18th century. General definition: Fisher (1922).
- Provides a first systematic method of constructing an estimator.
- Optimal procedure (in what sense?) under assumptions of regularity of the family $\{\mathbb{P}_\theta, \theta \in \Theta\}$.
- Sometimes difficult to implement in practice $\rightarrow$ optimization problem.

The likelihood function

Definition

Under de sampling model (in $\mathbb{R}$) with densities $f(\theta, x)$ the *likelihood function* *of the n-sample $(X_1, \ldots, X_n)$ associated to the family $\{f(\theta, \cdot), \theta \in \Theta\}$ is given by :*

$$\boxed{\theta \in \Theta \mapsto \mathcal{L}_n(\theta, X_1, \ldots, X_n) = \prod_{i=1}^{n} f(\theta, X_i)}$$

- A random function
- The distribution of the observations

Examples

- Example 1: Poisson model. We observe

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{Poisson}(\theta),$$

$\theta \in \Theta = \mathbb{R}_+ \setminus \{0\}$.

- The density is given by

$$f(\theta, x) = \frac{\theta^x}{x!} e^{-\theta}, \quad x = 0, 1, 2, \ldots.$$

- The associated likelihood function is

$$\begin{aligned} \theta \mapsto \mathcal{L}_n(\theta, X_1, \ldots, X_n) &= \prod_{i=1}^n e^{-\theta} \frac{\theta^{X_i}}{X_i!} \\ &= \frac{1}{\prod_{i=1}^n X_i!} e^{-n\theta} \theta^{\sum_{i=1}^n X_i} \end{aligned}$$

Examples

- Example 2 The Cauchy model. We observe

$$X_1, \dots, X_n \overset{i.i.d.}{\sim} \text{Cauchy centered around } \theta,$$

  $\theta \in \Theta = \mathbb{R}$.
- We have
$$f(\theta, x) = \frac{1}{\pi\left(1 + (x - \theta)^2\right)}$$
- The associated likelihood function is given by

$$\theta \mapsto \mathcal{L}_n(\theta, X_1, \dots, X_n) = \frac{1}{\pi^n} \prod_{i=1}^{n} \frac{1}{\left(1 + (X_i - \theta)^2\right)}$$

The maximum likelihood principle

1. Case 1 : "θ_1 is more likely than θ_2" if

$$\prod_{i=1}^{n} f(\theta_1, X_i) \geq \prod_{i=1}^{n} f(\theta_2, X_i)$$

2. Case 2 : "θ_2 is more likely than θ_1" if

$$\prod_{i=1}^{n} f(\theta_2, X_i) > \prod_{i=1}^{n} f(\theta_1, X_i)$$

The maximum likelihood principle:

$$\widehat{\theta}_{\mathrm{n}}^{\mathtt{mv}} = \begin{cases} \theta_1 & \text{when } \theta_1 \text{ is more likely} \\ \theta_2 & \text{when } \theta_2 \text{ is more likely} \end{cases}$$

Maximum Likelihood Estimation

- Situation : $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathbb{P}_\theta$, $\{\mathbb{P}_\theta, \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^d$, $\theta \mapsto \mathcal{L}_n(\theta, X_1, \ldots, X_n)$ the associated likelihood.

Definition

We call maximum likelihood estimator every estimator $\widehat{\theta}_{\mathrm{n}}^{\mathtt{mv}}$ satisfying

$$\mathcal{L}_n(\widehat{\theta}_{\mathrm{n}}^{\mathtt{mv}}, X_1, \ldots, X_n) = \max_{\theta \in \Theta} \mathcal{L}_n(\theta, X_1, \ldots, X_n).$$

- Questions : Existence, uniqueness, statistical properties?

Remarks

- Log-likelihood:

$$\theta \mapsto \ell_n(\theta, X_1, \ldots, X_n) = \log \mathcal{L}_n(\theta, X_1, \ldots, X_n) = \sum_{i=1}^{n} \log f(\theta, X_i).$$

Well-defined if $f(\theta, \cdot) > 0$.

Max. likelihood = max. log-likelihood.

(log-likelihood is usually easier to maximize)

- Likelihood equation :

$$\nabla_\theta \ell_n(\theta, X_1, \ldots, X_n) = 0$$

Example : Poisson model

- Likelihood

$$\mathcal{L}_n(\theta, X_1, \ldots, X_n) = \frac{1}{\prod_{i=1}^n X_i!} e^{-n\theta} \theta^{\sum_{i=1}^n X_i}$$

- Log-likelihood

$$\ell_n(\theta, X_1, \ldots, X_n) = c(X_1, \ldots, X_n) - n\theta + \sum_{i=1}^n X_i \log \theta$$

- Likelihood equation

$$-n + \sum_{i=1}^n X_i \frac{1}{\theta} = 0, \quad \text{soit} \quad \boxed{\widehat{\theta}_{\mathrm{n}}^{\,\mathtt{mv}} = \frac{1}{n} \sum_{i=1}^n X_i = \overline{X}_n}$$

Example : Cauchy model

- Likelihood

$$\mathcal{L}_n(\theta, X_1, \ldots, X_n) = \pi^{-n} \prod_{i=1}^{n} \frac{1}{1 + (X_i - \theta)^2}$$

- Log-likelihood

$$\ell_n(\theta, X_1, \ldots, X_n) = -n \log \pi - \sum_{i=1}^{n} \log\left(1 + (X_i - \theta)^2\right)$$

- Likelihood equation

$$\boxed{\sum_{i=1}^{n} \frac{X_i - \theta}{1 + (X_i - \theta)^2} = 0}$$

does not have an explicit solution and may have more than one solution in general.

Connection with supervised learning

Example 1: Gaussian Linear regression

Assume that we observe $(X_1, Y_1), \ldots, (X_n, Y_n)$ following the model

$$Y_i = \langle X_i, \beta \rangle + \sigma \xi_i,$$

where ξ_i are i.i.d. random standard normal variables.

- The distribution of $Y|X$ is given by $\mathcal{N}(\langle X, \beta \rangle, \sigma^2)$, where β is the parameter.
- Likelihood

$$\mathcal{L}_n(\beta, (X_1, Y_1), \ldots, (X_n, Y_n)) = C \exp\left(-\tfrac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \langle X_i, \beta \rangle)^2\right).$$

- Log-likelihood

$$\ell_n(\beta, (X_1, Y_1), \ldots, (X_n, Y_n)) = \log(C) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \langle X_i, \beta \rangle)^2.$$

Example 1: Gaussian Linear regression

- The optimization problem to solve becomes:

$$\min_\beta \sum_{i=1}^{n} (Y_i - \langle X, \beta \rangle)^2 = \min_\beta \| Y - X\beta \|^2.$$

- Maximizing the likelihood is equivalent in this case to minimizing the least squares.

Example 2: Logistic regression

Assume that we observe $(X_1, Y_1), \ldots, (X_n, Y_n)$, where $Y \in \{-1, +1\}$, following the model

$$\mathbb{P}(Y_i = 1 | X_i) = \frac{1}{1 + e^{-\langle X_i, \beta \rangle}}.$$

- The distribution of $Y|X$ is a Bernoulli distribution depending on a parameter β.
- Log-likelihood

$$\ell_n(\beta, (X_1, Y_1), \ldots, (X_n, Y_n)) = -\sum_{i=1}^{n} \log\left(1 + e^{-Y_i \langle X_i, \beta \rangle}\right).$$

Empirical risk minimization

In both cases, the estimation problem boils down to minimization of convex functions.

- Regression:

$$\min_{\beta} \sum_{i=1}^{n} (Y_i - \langle X, \beta \rangle)^2.$$

- Classification:

$$\min_{\beta} \sum_{i=1}^{n} \log \left(1 + e^{-Y_i \langle X_i, \beta \rangle} \right).$$

Convex optimization

- Goal: Find the minimizer of $f(x)$ where $f : \mathbf{R}^n \to \mathbf{R}$ is convex and smooth.

- In this problem, a necessary and sufficient condition for the optimal solution $\hat{x}$ is
$$\nabla f(\hat{x}) = 0.$$

Gradient Descent Method

- Idea: relies on the fact that $-\nabla f(x^k)$ is a descent direction.

- The update
$$x^{k+1} = x^k - \eta_k \nabla f(x^k)$$
leads to $f(x^{k+1}) < f(x^k)$.

- η_k is the step size: η_k cannot be too small (slow convergence), nor too big (divergence).

Gradient Descent Method

Algorithm:

- Given x_0 a starting point.
- Repeat: $x^{k+1} = x^k - \eta_k \nabla f(x^k)$ (until stopping criterion is satisfied).

Usual stopping criterion $\|\nabla f(x)\| \leq \epsilon$.

Pros and Cons

- Pros:
  - Can be applied to every dimension and space (even possible to infinite dimension)
  - Easy to implement
- Cons:
  - Local optima problem
  - Relatively slow close to minimum
  - Gradient methods are ill-defined for non-differentiable functions