Statistical Modeling

Monday, Course Week 3

Gherardo Varando

Exam

30 hours take home exam in Week 3

- Start: Wednesday 15 January 2019 at 09:00
- Deadline: 16 January 2019 at 15:00
- You can submit just a pdf (it can be generated with Rmarkdown)

Statistical inference

- **Statistical inference** or **learning** is the process of using data to discover the distribution that generated the data
- Suppose we have a sample $X_1, \ldots, X_n \sim F$ how de we infer the distribution F?
- Suppose we have a sample $X_1, \ldots, X_n \sim X$ how we infer the $\mathbb{E}(X)$? Here we already saw the empirical mean

Statistical models

What is a statistical model?

A **statistical model** $\mathcal{M}$ is a set of distributions (or densities or regression functions).

A discrete case

The set of all the symmetric PMF $f(x)$ over the values $A = \{-2, -1, 1, +2\}$.

$$\mathcal{M} = \left\{ f(x) \geq 0 \text{ s.t. } f(x) = f(-x) \forall x \in A, \sum_{x \in A} f(x) = 1 \right\}$$

Continuous case

$$\mathcal{M} = \{f \text{ continuous densities in } [0, 1]\}$$

Parametric and non-parametric models

Parametric model

A parametric model is a statistical model that can be **parametrized** by a **finite** number of parameters.

- The Gaussian distributions $\{N(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma > 0\}$
- The Bernoulli distributions $\{f(x|p) = p^x(1-p)^{1-x}, p \in [0,1]\}$

Non-parametric model

A model is **non**-parametric if can not be parametrized by a finite number of parameters.

- $\mathcal{M} = \{\text{All CDF}\}$
- $\{f \text{ continuous densities in } [0,1]\}$

We already saw some examples of parametric models and selection of the parameters, that is **learning**.

- The ISI data, we tried to fit visually and with the Q-Q plot the exponential and gamma distribution
- In the Brain cell data set we saw the log-normal family and the Gaussian family

Point estimation

Point estimation refers to providing a single “best guess” of some quantity of interest.

- A parameter of the model
- $\mathbb{E}(X)$ or $\mathbb{V}(X)$
- A regression function, density or CDF

In general a point estimator is some function of the random observations $X_1, \ldots, X_n$

$$g(X_1, \ldots, X_n)$$

- The empirical mean $\bar{X}$ is a point estimator of the true mean value $\mathbb{E}(X)$
- The empirical CDF $\hat{F}$ is a point estimator of the true CDF F_X

!!!

A point estimator is a function of random variables $X_1, X_2, \ldots, X_n$ (usually i.i.d.), hence it is a **random variable**

- Let $\hat{\theta}$ the point estimator of the parameter θ in a parametric model
- Since $\hat{\theta}$ is a random variable, we can define $\mathbb{E}(\hat{\theta})$ and more importantly $\mathbb{V}(\hat{\theta})$
- The standard deviation of $\hat{\theta}$ is called the **standard error** of $\hat{\theta}$ and is a measure of the error of the estimator
- Since in general the real distribution is unknown we can only estimate $\hat{se} \approx se = \sqrt{\mathbb{V}(\hat{\theta})}$

Empirical mean and sem

The empirical mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is a **point estimator** of the true mean value $\mathbb{E}(X)$. We already saw that the standard error of the mean

$$sem \approx se(\bar{X}) = \sqrt{\mathbb{V}(\bar{X})}$$

is a measure of the error that we perform using $\bar{X}$ to estimate $\mathbb{E}(X)$

Bernoulli distributions

Suppose we observe $X_1, \ldots, X_n \sim Bernoulli(p)$. We already saw how we can estimate the only parameter p.

Bernoulli distributions

Suppose we observe $X_1, \ldots, X_n \sim Bernoulli(p)$. We already saw how we can estimate the only parameter p.

- $\mathbb{E}[X_i] = p$
- So we can use $\hat{p} = \bar{X}$ as a point estimator of p
- And we can obtain an estimation of the error

$$se = \sqrt{p(1-p)/n}$$

$$\hat{se} = \sqrt{\hat{p}(1-\hat{p})/n}$$

Methods of moments

The Bernoulli estimation we have seen before can be interpreted as an example of **method of moments**

Method of moments

The method of moments estimator $\hat{\theta}$ is obtained from the equations,

$$\mathbb{E}(X) = \int_{\mathbb{R}} x f(x|\theta) dx = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$\mathbb{E}(X^2) = \int_{\mathbb{R}} x^2 f(x|\theta) dx = \frac{1}{n} \sum_{i=1}^{n} X_i^2$$

$$\vdots = \vdots$$

$$\mathbb{E}(X^k) = \int_{\mathbb{R}} x^k f(x|\theta) dx = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

The method of moments can be easy to implement, and under appropriate condition the corresponding estimator converge to the true value of the parameter

- Moreover the method of moments estimator is asymptotically normal

$$\sqrt{n}(\hat{\theta}_n - \theta) \to N(0, \Sigma)$$

If we have only one parameter we just need one equation

$$\mathbb{E}(X) = \bar{X}$$

- This method is also used as a first estimation to initialize other algorithms

The likelihood of the data

Likelihood

Given a random sample $X_1, X_2, \dots, X_n \sim X$, where $X \sim f(x|\theta)$ the likelihood function is defined as,

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$

where $f(x|\theta)$ is a probability density function or a probability mass function

- The log-likelihood function is $\ell_n(\theta) = \ln(\mathcal{L}_n(\theta))$
- Intuitively the likelihood is the probability of observing the data under the given parameter θ
- We can think of selecting the θ that maximise such probability Maximum Likelihood Estimator (MLE)

Example: exponential distribution

Let $X_1, \dots, X_n \sim exp(\lambda)$

$$f(x|\lambda) = \lambda e^{-\lambda x} \qquad \text{(pdf)}$$

Thus the likelihood is,

$$\mathcal{L}_n(\lambda) = \prod_{i=1}^{n} f(X_i|\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda X_i}$$

$$\ell_n(\lambda) = \sum_{i=1}^{n} (\ln(\lambda) + (-\lambda X_i)) = n \ln(\lambda) - \lambda \sum_{i=1}^{n} X_i$$

Example: Bernoulli distribution

Let $X_1, \ldots, X_n \sim Bernoulli(p)$

$$f(x|p) = p^x (1-p)^{1-x} \qquad \text{(pmf)}$$

Hence the likelihood is,

$$\mathcal{L}_n(p) = \prod_{i=1}^{n} f(X_i|p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i}$$

$$\ell_n(p) = \sum_{i=1}^{n} \ln\left(p^{X_i}\right) + \ln\left((1-p)^{1-X_i}\right) =$$

$$\ln(p)\sum_{i=1}^{n} X_i + \ln(1-p)\sum_{i=1}^{n}(1-X_i)$$

The maximum likelihood estimator

The maximum likelihood estimator (MLE), is $\hat{\theta}$ the value of the parameter θ that maximizes $\mathcal{L}_n(\theta)$

- The maximum of $\mathcal{L}_n(\theta)$ occurs at the same place as the maximum of $\ell_n(\theta)$ (the log-likelihood). Often it is easier to work with the log-likelihood.
- If we multiply $\mathcal{L}_n(\theta)$ by any positive constant c (not depending on θ) then this will not change the MLE. Hence we can drop multiplicative constants in $\mathcal{L}_n(\theta)$ or equivalently additive constants in $\ell_n(\theta)$