Big Data Analytics: Probability Refresher

M. Ndaoud

Statistical paradigm

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

$$\mathtt{x}_1, \ldots, \mathtt{x}_n$$

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, $X_1(\omega) = \mathtt{x}_1, \ldots, X_n(\omega) = \mathtt{x}_n$)

- 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)

所有ML目标是找到这个joint distribution

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

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 ?

Logistic Regression是一个classification的问题
（svm也是）

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

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 ?
      P (X=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 ?
  - 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?

Fundamental Theorems

The strong law of large numbers (LLN)

Theorem

Let (X_n) be a sequence of i.i.d. random variables such that $\mathbb{E}\,|X_1| < \infty$. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mathbb{E}\,X_1$$

random to the left, deterministic to the right

Central Limit Theorem (CLT)

Theorem

Let (X_n) be a sequence of i.i.d. random variables such that $\mathbb{E} X_1^2 < \infty$. Then

$$\frac{\sqrt{n}}{\sigma}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E} X_1\right) \xrightarrow{d} \mathcal{N}(0,1)$$

- CLT : "speed" of convergence in the LLN.
- Interpretation of CLT :

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \mu + \frac{\sigma}{\sqrt{n}}\xi^{(n)}, \quad \xi^{(n)} \overset{d}{\approx} \mathcal{N}(0,1).$$

Empirical mean 的分布服从高斯分布

- The type of convergence is a convergence in distribution. (weak convergence).

Slutsky Lemma

- The vector $(X_n, Y_n) \xrightarrow{d} (X, Y)$ if

$$\mathbb{E}\left[\varphi(X_n, Y_n)\right] \to \mathbb{E}\left[\varphi(X, Y)\right],$$

  for any continuous bounded function φ.
- Warning ! If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$, but not necessarily $(X_n, Y_n) \xrightarrow{d} (X, Y)$.
- But (Slutsky Lemma) if $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{\mathbb{P}} c$ (constant), then $(X_n, Y_n) \xrightarrow{d} (X, Y)$.
- Under the Lemma hypotheses, for any continuous function g, we have $g(X_n, Y_n) \xrightarrow{d} g(X, Y)$.

Continuous map theorem

Let $f : \mathbb{R} \mapsto \mathbb{R}$ be a continuous function and (X_n) a sequence of r.v.

1. If (X_n) converges in distribution to X then $f(X_n)$ converges also in distribution to $f(X)$
2. If (X_n) converges in probability to X then $f(X_n)$ converges also in probability to $f(X)$
3. If (X_n) converges a.s. to X then $f(X_n)$ converges also a.s. to $f(X)$

Graphical Statistics

- Quantiles
- Covariance and correlation

Cumulative distribution function (cdf)

Population cdf :

$$F(x) = \mathbb{P}\left(X \leq x\right), \quad x \in \mathbb{R}$$

Cumulative distribution function (cdf)

Population cdf :

$$F(x) = \mathbb{P}\left(X \leq x\right), \quad x \in \mathbb{R}$$

Empirical cdf :

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\left(X_i \leq x\right), \quad x \in \mathbb{R}$$

Some asymptotic properties:

$$\widehat{F}_n(x) \xrightarrow{a.s.} F(x), \quad \left\|\widehat{F}_n - F\right\|_\infty \xrightarrow{a.s.} 0$$

Quantiles

Quantiles

Definition

Let X be a r.v. (of cdf F) and $0 < p < 1$. We call quantile of order p of X (resp. F) :

$$q_p(F) = \inf\{x \in \mathbb{R} : F(x) \geq p\}$$

- When F is continuous and strictly increasing the **quantile of order** p of F is the unique solution to

$$F(q_p) = p \qquad (\text{ i.e. } \; q_p = F^{-1}(p) \text{ }).$$

- the **median** $= \mathrm{med}(F) = q_{1/2}(F)$
- the **quartiles** $= \{q_{1/4}(F), \mathrm{med}(F), q_{3/4}(F)\}$

Population and empirical quantiles

The "population" quantile of order p:

$$T(F) = q_p(F) = \inf\{x \in \mathbb{R} : F(x) \geq p\}$$

The "empirical" quantile of order p:

$$T(\widehat{F}_n) = \widehat{q}_{n,p} = \inf\{x \in \mathbb{R} : \widehat{F}_n(x) \geq q\}$$

Definition

Let $X_1, \dots, X_n$ be a sample of size n of r.v. We call order statistics the n statistics $X_{(1)}, \dots, X_{(n)}$ such that

$$X_{(1)} \leq \cdots \leq X_{(n)}$$

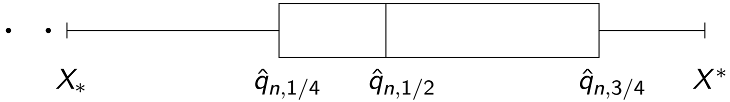
1. For the quantile of order $0 < p < 1$:

$$\boxed{\widehat{q}_{n,p} = X_{(k)} = X_{(\lceil np \rceil)} \text{ when } \frac{k-1}{n} < p \leq \frac{k}{n}}$$

2. In particular, the empirical median satisfies :

$$\boxed{\widehat{q}_{n,1/2} = \mathrm{med}(\widehat{F}_n) = X_{(\lceil n/2 \rceil)}} \text{ where } \lceil t \rceil = \min(n \in \mathbb{N} : n \geq t)$$

The boxplot : synthetic representation of the dispersion of real data



end of the whiskers :

$$X_* = \min\{X_i : |X_i - \hat{q}_{n,1/4}| \leq 1,5\mathcal{I}_n\},$$

$$X^* = \max\{X_i : |X_i - \hat{q}_{n,3/4}| \leq 1,5\mathcal{I}_n\}.$$

Interquartile range:

$$\mathcal{I}_n = \hat{q}_{n,3/4} - \hat{q}_{n,1/4}.$$

Samples beyond the whiskers are considered as *outliers*.

The qq-plot : fit test to some distribution

Given a sample of size n $X_1, \ldots, X_n$ and a cdf F_{ref}, we want to test if the following hypothesis is true :

$$(H_0) \qquad \text{"The } X_i \text{ are distributed according to } F_{ref}\text{"}$$

To "accept or reject visually" this hypothesis, we can draw the qq-plot : it is a scatter plot

$$\boxed{\left(q_{i/n}(F_{ref}), \widehat{q}_{n,i/n}\right)_{i=1}^{n} = \left(q_{i/n}(F_{ref}), X_{(i)}\right)_{i=1}^{n}}$$

1. If the scatter plot is "approximately" aligned with the line $y = x$ then we accept the hypothesis (we also draw the line $y = x$ on the qq-plot)
2. If the scatter plot is "approximately" aligned with a line then the hypothesis is true after centering and scaling (generally, we normalize data first)

Examples of Q-Q plots

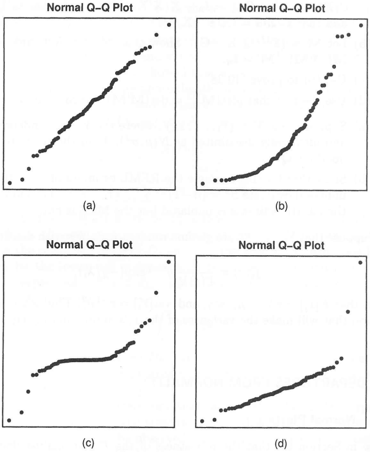


Fig. 10.5 Normal plots of residuals: (a) No indication of non-normality. (b) Skewed errors.(c) Heavy-tailed errors. (d) Outliers.

Covariance et correlation

Dependence between two random variables

- In order to measure the dependence between X and Y, it is relevant to quantify the variation of one variable with respect to the other one.

- If X increases, for example, does Y increase too? If so, what is the level of this dependence?

- In order to address the above questions we introduce the notion of **covariance/correlation**.

The notion of covariance

Definition

Let X and Y be two random variables with finite variances. We define the covariance between X and Y such that

$$\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y)$$

- If X and Y are two independent variables then $\mathrm{Cov}(X, Y) = 0$.
- The reverse **is not always true**.
- For applications, it is desirable to have a dimension-free measure of dependence.

The correlation coefficient

Definition

Let X and Y be two r.v. The correlation coefficient between X and Y, that we denote $\mathrm{Cor}(X, Y)$, *is given by*

$$\mathrm{Cor}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} = \mathrm{Cov}(X_*, Y_*)$$

where $X_* = (X - \mathbb{E}(X))/\sigma_X$ *and* $Y_* = (Y - \mathbb{E}(Y))/\sigma_Y$.

Proposition

For each pair of random variables (X, Y) *we have*

- $|\mathrm{Cor}(X, Y)| \leq 1$.
- $|\mathrm{Cor}(X, Y)| = 1$ **if and only if** $Y = aX + b$ *for constants* a *and* b *in* $\mathbb{R}$.

Estimation of the correlation coefficient

- Suppose that the correlation coefficient between X and Y is not known, but that we observe n i.i.d. copies of $(X_1, Y_1), \ldots, (X_n, Y_n)$. How can we estimate $\mathrm{Cor}(X, Y)$?

- Recall that

$$\boxed{\mathrm{Cor}(X, Y) = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}}$$

Estimation of the correlation coefficient

- Suppose that the correlation coefficient between X and Y is not known, but that we observe n i.i.d. copies of $(X_1, Y_1), \ldots, (X_n, Y_n)$. How can we estimate $\mathrm{Cor}(X, Y)$?

- Recall that

$$\boxed{\mathrm{Cor}(X, Y) = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}}$$

- It seems relevant to replace moments by their estimates here.

Estimation of the correlation coefficient

We get then what we call the empirical correlation coefficient given by

$$\boxed{\hat{\rho}_n(X,Y) := \frac{n\sum_{i=1}^n X_i Y_i - \left(\sum_{i=1}^n X_i\right)\left(\sum_{i=1}^n Y_i\right)}{\sqrt{n\sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i\right)^2}\sqrt{n\sum_{i=1}^n Y_i^2 - \left(\sum_{i=1}^n Y_i\right)^2}}}$$

Empirical correlation matrix

We observe n i.i.d. copies of a random vector $(X_1, \ldots, X_d)$ and define the empirical correlation matrix $\rho^n \in \mathbb{R}^{d \times d}$ such that

$$\text{For all } i, j \quad \rho^n_{ij} = \hat{\rho}_n(X_i, X_j)$$

- The matrix ρ^n is semi-definite positive.
- We can simply compute the lower triangular part of this matrix.

Visualization of the empirical correlation matrix

In order to graphically visualize the different correlations, we use the tool heatmap in Python.