Elements of probability theory. Statistical models, functionals and distances.

Maxim Panov

Skoltech

November, 2021

Skoltech

Outline

Outline

Statistical model

Let the data sample consists of independent identically distributed random variables: $X_1, \ldots, X_n \sim F$.

Our goal is to infer F or some feature of F given a sample.

Definition

A statistical model $\mathfrak{F}$ is a set of distributions (densities, regression functions, etc.).

Example

If we assume that the data come from a Normal distribution, then the model is

$$\mathfrak{F} = \left\{ f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2 \right) \right\}.$$

Skoltech
Skolkovo Institute of Science and Technology

Definition

A parametric model is a set $\mathfrak{F}$ that can be parameterized by a finite number of parameters.

$$\mathfrak{F} = \left\{ f(x;\theta),\ \theta \in \Theta \subseteq \mathbb{R}^p \right\},$$

where $p \in \mathbb{N}$.

Definition

A nonparametric model is a set $\mathfrak{F}$ that cannot be parameterized by a finite number of parameters.

Examples: parametric estimation

Example (One-dimensional Parametric Estimation)

Let $X_1, \dots, X_n$ be independent $\mathrm{Bernoulli}(p)$ observations. The problem is to estimate the parameter p.

Example (Two-dimensional Parametric Estimation)

Suppose that $X_1, \dots, X_n$ are independent and distributed according to Normal distribution with parameters μ and σ. The goal is to estimate the parameters from the data.

Example (Multidimensional Observations)

We observe multidimensional r.v. $\vec{X}_1, \dots, \vec{X}_n$ which are independent Normal with mean $\vec{\mu}$ and covariance matrix Σ. We need to estimate $\vec{\mu}$ and/or Σ based on the data.

Examples: nonparametric estimation

Example (Nonparametric estimation of the CDF)

Le $X_1, \dots, X_n$ be independent observations from a CDF F. The problem is to estimate F assuming only that $F \in \mathfrak{F}_{ALL}$ (class of all CDF's).

Example (Nonparametric estimation of functionals)

Let $X_1, \dots, X_n \sim F$. Suppose we want to estimate $\mu = \mathbb{E}X_1 = \int x dF(x)$ assuming only that μ exists.
We can think of μ as a function of F:

$$\mu = T(F) = \int x dF(x),$$

where $T(F)$ is called **statistical functional**. We can estimate $T(F)$ by substituting F with its estimate.

Examples: nonparametric estimate of density

Example

- Let $X_1, \ldots, X_n$ be independent observations from a CDF F and let $f = F'$ be the corresponding PDF.
- It is not possible to estimate f, assuming only that $f \in \mathfrak{F}_{DENS}$, where $\mathfrak{F}_{DENS}$ is the set of all probability density functions.
- We need to assume some smoothness on f.
- For example, we might assume that $f \in \mathfrak{F}_{DENS} \cap \mathfrak{F}_{SOB}$, where

$$\mathfrak{F}_{SOB} = \left\{ f \;\middle|\; \int \left(f''(x)\right)^2 dx < \infty \right\}.$$

Examples: regression, prediction and classification

Example

- Suppose we observe pairs of data $(X_1, Y_1), \ldots, (X_n, Y_n)$. We need to estimate **regression function**

$$r(x) = \mathbb{E}(Y \mid X = x).$$

- Class of considered regression functions $r \in \mathfrak{F}$ can be either parametric or nonparametric.
- Regression models are sometimes written as

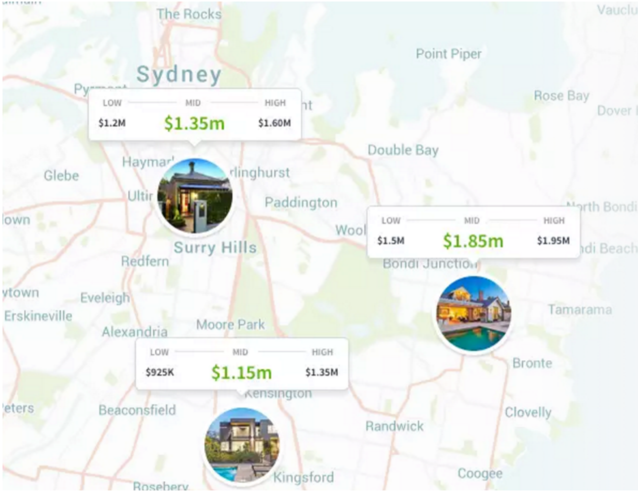
$$Y = r(X) + \epsilon.$$

We can derive that

$$\mathbb{E}\epsilon = \mathbb{E}\mathbb{E}(\epsilon \mid X) = \mathbb{E}\big(\mathbb{E}(Y \mid X) - r(X)\big) = \mathbb{E}\big(r(X) - r(X)\big) = 0.$$

Outline

Concrete Example: Prices of Houses



Question: what is MID?

Estimators of the Mean

Let $X^n = \{X_1, \ldots, X_n\}$ be measurements of some value μ.

Possible estimates of μ:

- mean
$$\bar{X}^n = \arg\min_m \sum_{k=1}^{n} (X_k - m)^2 = \frac{1}{n} \sum_{k=1}^{n} X_k;$$
- median
$$\mathrm{med}(X^n) = \arg\min_m \sum_{k=1}^{n} |X_k - m|;$$
- central point
$$\arg\min_m \Big\{ \max_k |X_k - m| \Big\} = \frac{\min(X^n) + \max(X^n)}{2}.$$

Natural Model for the Estimation of the Mean

Simple and natural model for the estimation of the mean:

$$X_k = \mu + \epsilon_k, \quad k = 1, \ldots, n, \tag{1}$$

where

- $\epsilon_k \in \mathbb{R}$ — i.i.d. random variables with zero mean,
- $\mu \in \mathbb{R}$ — unknown parameter which we aim to estimate based on X^n.

To estimate parameter μ means to construct a function:

$$\hat{\mu}(X_1, \ldots, X_n) \colon \mathbb{R}^n \to \mathbb{R},$$

which should be close to μ is some probabilistic sense.

Outline

Cumulative Distribution Function

From probabilistic point of view the model

$$X_k = \mu + \epsilon_k, \quad k = 1, \dots, n, \tag{2}$$

is equivalent to sample X^n being generated from the CDF

$$F(x_1, \dots, x_n; \mu) = \mathbb{P}\{X_1 \le x_1, \dots, X_n \le x_n\} = \prod_{k=1}^{n} \mathbb{P}\{X_k \le x_k\} = \prod_{k=1}^{n} F(x_k - \mu),$$

where

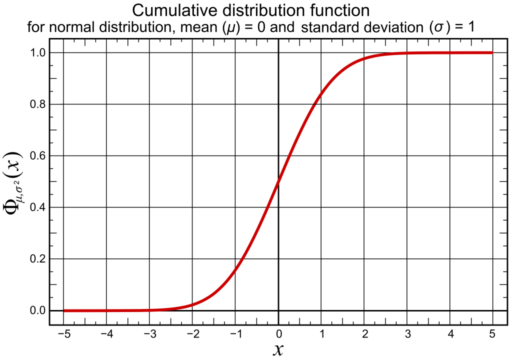
$$F(x) = \mathbb{P}\{\epsilon_k \le x\}$$

is CDF of random variable ϵ_k.

Properties of CDF

Simple properties of CDF:

- $0 \leq F(x) \leq 1$;
- $F(x)$ – nondecreasing function;
- $\lim\limits_{x \to -\infty} F(x) = 0$ and $\lim\limits_{x \to +\infty} F(x) = 1$.



Skoltech
Skolkovo Institute of Science and Technology

Probability Density Function

Derivative of CDF (if exists)

$$f(x) = \frac{dF(x)}{dx}$$

is called *probability density function (PDF)* of random variable ϵ_k.

Simple properties of PDF:

- $f(x) \geq 0$ for all $x \in \mathbb{R}$;
- $\displaystyle\int_{-\infty}^{\infty} f(x)\, dx = 1.$

Physical meaning is very clear:

$$\mathbb{P}\{\epsilon_k \in [x, x+h]\} \approx f(x) \cdot h$$

for small h.

Empirical CDF

- In statistics usually for every probabilistic object an empirical analogue is considered which serves as an estimate of this object.
- For the CDF of random variable ϵ

$$F(x) = F_\epsilon(x) = \mathbb{P}\{\epsilon \le x\}$$

the empirical (i.e. the one based on data sample $\epsilon^n = \{\epsilon_1, \dots, \epsilon_n\}$) analogue is

$$F_n(x) = F(x; \epsilon^n) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\epsilon_i \le x\}.$$

- This object is called *Empirical distribution function*.

Properties of Empirical CDF

- Clearly, Empirical CDF is non-decreasing function.
- It doesn't change value for any permutation of $\{\epsilon_i,\ i = 1, \dots, n\}$.
- Thus, empirical CDF can be written as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{\epsilon_{(i)} \leq x\},$$

  where we have introduced so-called *order statistics*:

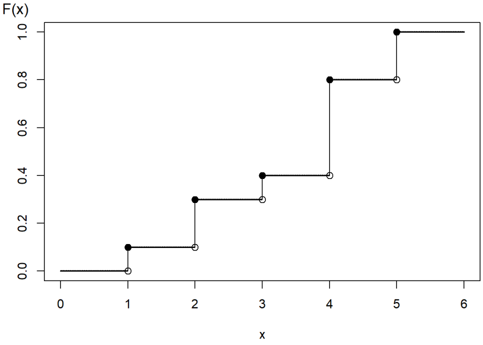
$$\epsilon_{(1)} \leq \epsilon_{(2)} \leq \cdots \leq \epsilon_{(n)}.$$

- Empirical CDF is piecewise constant and has jumps of $1/n$ at points $\epsilon_{(i)}, i = 1, \dots, n$.

Empirical CDF

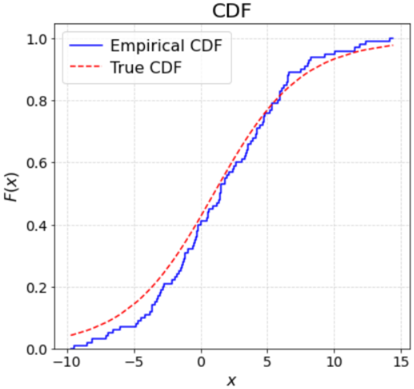
Empirical distribution function based on the data sample $\epsilon^n = \{\epsilon_1, \ldots, \epsilon_n\}$:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{\epsilon_i \leq x\}.$$



Empirical CDF vs CDF

- The only reason to consider empirical analogues is that they should be close to their prototypes.
- Empirical CDF vs CDF



- **Question:** how can we measure closeness between stochastic objects?

Convergence of Random Variables

- In calculus:
  - Consider a sequence of real numbers x_n;
  - x_n converges to real number x if for any $\varepsilon > 0$ it holds

    $$|x_n - x| < \varepsilon$$

    for all $n \geq N = N(\varepsilon)$.
  - Then, the standard notation is $x_n \to x, \; n \to \infty$.
- However, it is not easy to generalize for probabilistic sequences:
  - Consider example of i.i.d. random variables X_i having the distribution $\mathcal{N}(0, 1)$.
  - All the variables have the same distribution so we want to conclude that the sequence converges to X which is also $\mathcal{N}(0, 1)$.
  - However, apparently, $\mathbb{P}(X_n = X) = 0$ for all n.

Convergence in Distribution

The previous issue can be fixed by considering *convergence in distribution*.

Definition

- Let $X_1, X_2, \ldots$ be a sequence of random variables and let X be another random variable.
- Let F_n denote the CDF of X_n and let F denote the CDF of X.
- Then X_n converges to X in distribution, written $X_n \rightsquigarrow X$, if

$$\lim_{n \to \infty} F_n(x) = F(x)$$

at all x for which F is continuous.

Convergence in Probability

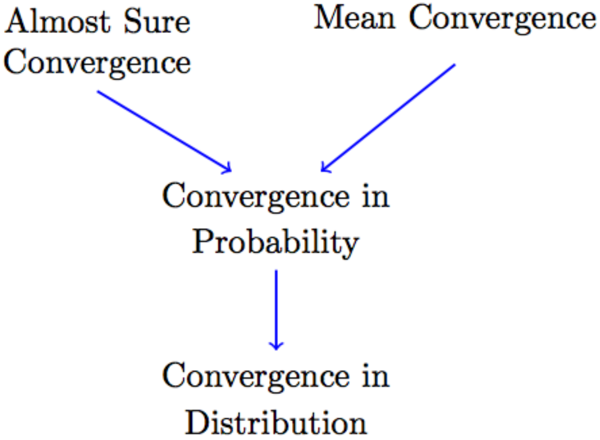
The stronger type of convergence is *convergence in probability*.

Definition

- Let $X_1, X_2, \ldots$ be a sequence of random variables and let X be another random variable.
- Then X_n converges to X in probability, written $X_n \xrightarrow{\mathbb{P}} X$, if for any $\varepsilon > 0$

$$\mathbb{P}(|X_n - X| > \varepsilon) \to 0, \ n \to \infty.$$

Relationship between Types of Convergence



Law of Large Numbers

- Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables.
- Let $\mu = \mathbb{E}(X_1)$ and let $\sigma^2 = \mathbb{V}(X_1)$.
- Recall that the sample mean is defined as

$$\bar{X}^n = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

- Then $\mathbb{E}(\bar{X}^n) = \mu$ and let $\mathbb{V}(\bar{X}^n) = \sigma^2/n$.

Theorem (The Weak Law of Large Numbers)

Under the conditions above

$$\bar{X}^n \xrightarrow{\mathbb{P}} \mu.$$

Central Limit Theorem

Theorem (The Central Limit Theorem (CLT))

- *Let* $X_1, X_2, \ldots$ *be a sequence of i.i.d. random variables.*
- *Let* $\mu = \mathbb{E}(X_1)$ *and let* $\sigma^2 = \mathbb{V}(X_1)$.
- *Let*

$$\bar{X}^n = \frac{1}{n} \sum_{k=1}^{n} X_k.$$

Then

$$Z_n \equiv \frac{\bar{X}^n - \mathbb{E}\bar{X}^n}{\sqrt{\mathbb{V}(\bar{X}^n)}} = \frac{\sqrt{n}(\bar{X}^n - \mu)}{\sigma} \rightsquigarrow Z,$$

where $Z \sim \mathcal{N}(0,1)$. *In other words,*

$$\lim_{n\to\infty} \mathbb{P}\{Z_n \le z\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \mathrm{e}^{-x^2/2}\, dx.$$

Moments of Empirical CDF

- Let us prove the convergence for Empirical CDF assuming that ϵ_i are *independent identically distributed* random variables with CDF $F(x)$.
- Then for a fixed x we have

$$\mathbb{E}F_n(x) =$$

and

$$\mathbb{E}[F_n(x) - F(x)]^2 =$$

Moments of Empirical CDF

- Let us prove the convergence for Empirical CDF assuming that ϵ_i are *independent identically distributed* random variables with CDF $F(x)$.
- Then for a fixed x we have

$$\mathbb{E}F_n(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\mathbf{1}\{\epsilon_i \le x\} = \mathbb{P}\{\epsilon_1 \le x\} = F(x)$$

and

$$\mathbb{E}[F_n(x) - F(x)]^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n}\mathbb{E}\big[\mathbf{1}\{\epsilon_i \le x\} - F(x)\big]\big[\mathbf{1}\{\epsilon_k \le x\} - F(x)\big]$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\big[\mathbf{1}\{\epsilon_i \le x\} - F(x)\big]^2 = \frac{1}{n^2}\sum_{i=1}^{n}\Big[F(x) - 2F^2(x) + F^2(x)\Big] = \frac{F(x)[1-F(x)]}{n}.$$

CLT for Empirical CDF

- By Central Limit Theorem we obtain

Theorem

$$\lim_{n\to\infty} \mathbb{P}\left\{ \frac{\sqrt{n}|F_n(x) - F(x)|}{\sqrt{F(x)[1-F(x)]}} \geq z \right\} = \frac{2}{\sqrt{2\pi}} \int_z^{\infty} \mathrm{e}^{-u^2/2}\, du.$$

- Thus, for any fixed x empirical CDF deviates from CDF for the value of order $\frac{1}{\sqrt{n}}$.

Kolmogorov Distance

- **Question:** How close $F_n(x)$ and $F(x)$ are as functions?
- Kolmogorov distance:

$$\sup_x \left| F_n(x) - F(x) \right| \stackrel{\text{def}}{=} \|F_n - F\|_\infty.$$

- We further focus on the distribution of $\|F_n - F\|_\infty$:

$$\mathbb{P}\Big\{ \|F_n - F\|_\infty > z \Big\}.$$

DKW Inequality

Theorem (Dvoretzky–Kiefer–Wolfowitz)

For any $n \geq 1$ it holds

$$\mathbb{P}\left\{\sqrt{n}\|F_n - F\|_\infty > z\right\} \leq 2\mathrm{e}^{-2z^2}.$$

- The proof of this non-asymptotic result is non-trivial and it was proved relatively recently.
- The important conclusion: the distance between CDF and empirical CDF has the order $1/\sqrt{n}$, i.e. the same as at fixed point.

Outline

δ-method

- For the density
$$f(x) = \frac{dF(x)}{dx}$$
its empirical analogue is not a function in the ordinary sense.
- It is given by
$$f_n(x) = f(x; \epsilon^n) = \frac{dF_n(x)}{dx} = \frac{1}{n}\sum_{i=1}^{n} \delta(x - \epsilon_i),$$
where $\delta(\cdot)$ is Dirac delta function.
- It is generalized function, i.e. a linear functional on the space of functions.
- For some continuous function f an action of the functional on some function φ can be defined as:
$$\langle f, \varphi \rangle = \int_{-\infty}^{\infty} f(x)\varphi(x)\, dx.$$

δ-method

- Dirac delta function is defined as

$$\langle \delta, \varphi \rangle = \varphi(0).$$

- From an intuitive point of view δ function can be defined as a limit for $n \to \infty$ of a sequence

$$\delta_n(x) = \begin{cases} n, & x \in [-1/(2n), 1/(2n)], \\ 0, & x \notin [-1/(2n), 1/(2n)]. \end{cases}$$

- As $\delta_n(x)$ is integrable, then

$$\lim_{n\to\infty} \langle \delta_n, \varphi \rangle = \lim_{n\to\infty} n \int_{-1/(2n)}^{1/(2n)} \varphi(x)\, dx = \varphi(0).$$

δ-method

- In statistics, empirical density plays very important role.
- In many tasks one needs to estimate functionals of density. For example,
  - *mean value* of the random variable
$$\mu(p) = \int_{-\infty}^{\infty} x f(x)\, dx;$$
  - *variance*
$$\sigma^2(p) = \int_{-\infty}^{\infty} \left[x - M(p)\right]^2 f(x)\, dx = \int_{-\infty}^{\infty} x^2 f(x)\, dx - \left[\int_{-\infty}^{\infty} x f(x)\, dx\right]^2;$$
  - *quantile* $t^\alpha(p)$ *of level* α, defined as a root of equation
$$\int_{-\infty}^{t^\alpha(p)} f(x)\, dx = \alpha.$$

δ-method

- In all these examples we may represent the object in question as a functional $\Phi(p)$ of density p.

- *δ-method* is an estimate of the functional $\Phi(p)$ defined as

$$\hat{\Phi}(\epsilon^n) = \Phi\big[p(\cdot;\epsilon^n)\big].$$

- In other words, to estimate functional $\Phi(p)$ we plug-in empirical density instead of unknown density p.

δ-method

For the considered examples we obtain the following empirical analogues:

- empirical mean

$$\hat{\mu}(\epsilon^n) = \frac{1}{n}\sum_{k=1}^{n} \epsilon_k;$$

- empirical variance

$$\hat{\sigma}^2(\epsilon^n) = \frac{1}{n}\sum_{k=1}^{n} \epsilon_k^2 - \left[\frac{1}{n}\sum_{k=1}^{n} \epsilon_k\right]^2;$$

- empirical quantile $\hat{t}^{\alpha}(\epsilon^n)$, which is defined as a root of equation

$$\#\left\{k \colon \epsilon_k \leq \hat{t}^{\alpha}(\epsilon^n)\right\} = \lfloor n\alpha \rfloor,$$

where $\lfloor x \rfloor$ the integer part of number x.

Outline

Back to the Estimation of the Mean

- We consider the data $Y^n = \{Y_1, \ldots, Y_n\}$ being obtained from the model

$$Y_i = \mu + \epsilon_i, \quad i = 1, \ldots, n$$

  for some unknown parameter $\mu \in \mathbb{R}$.
- For simplicity, we assume that ϵ_i are i.i.d. random variables with density $f(x)$.
- The joint PDF of $Y^n = \{Y_1, \ldots, Y_n\}$ is given by

$$f(y_1, \ldots, y_n; \mu) = \prod_{i=1}^{n} f_\epsilon(y_i - \mu), \quad \mu \in \mathbb{R}.$$

- However, the data Y^n in fact has the unknown true density $f_\circ(x)$.
- What we have in our hands is δ-function

$$f(x; Y^n) = \delta(x - Y^n), \quad x \in \mathbb{R}^n.$$

Three densities involved

Thus, we consider 3 objects in the space of all densities

- true density $f_\circ(x)$, which is unknown;
- empirical density $\delta(x - Y^n)$;
- the family of modelling densities $f(x;\mu)$, $\mu \in \mathbb{R}$.

Basic idea: to estimate the parameter μ we aim to find the density from the family $f(x;\mu),\ \mu \in \mathbb{R}$, which is closer to empirical density $\delta(x - Y^n)$.

Question: how do we define closeness?

Distance in Statistics

- We need to define the distance or some measure of closeness between distances.
- If $d(\cdot,\cdot)$ is defined then the optimal value is given by

$$\mu^*(f_\circ) = \arg\min_{\mu} d\big[f_\circ, f(\cdot;\mu)\big].$$

- $\mu^*(f_\circ)$ is not an estimate as it depend on unknown density $f_\circ(\cdot)$.
- To estimate $\mu^*(f_\circ)$ from observations the only available choice is δ-method:

$$\bar{\mu}(Y^n) = \mu^*\big[\delta(\cdot - Y^n)\big].$$

Total Variation Distance

- The most natural distance between probability distributions is so-called *total variation distance*.
- For random variables ξ, η with densities $f_\xi(\cdot)$, $f_\eta(\cdot)$ it is defined as:

$$D(f_\xi, f_\eta) = \sup_A \left| \int_A f_\xi(x)\, dx - \int_A f_\eta(x)\, dx \right|,$$

  where $\sup$ is computed over all measurables sets.
- This definition seems to be not convenient as $\sup$ is computed over all sets A.
- However, the following important fact helps:

Theorem (Scheffe (1947))

If densities $f_\xi(x),\ f_\eta(x),\ x \in \mathbb{R}^n$ exist for random variables $\xi,\ \eta$, then

$$D(f_\xi, f_\eta) = \frac{1}{2} \int_{\mathbb{R}^n} |f_\xi(x) - f_\eta(x)|\, dx.$$

Kullback-Leibler Divergence

- However, total variation distance is hard to use for construction of estimates due to complexity of computation:

$$\mu^*(f_\circ) = \arg\min_{\mu} D\big[f_\circ, f(\cdot;\mu)\big].$$

- A realizable approach to the problem of approximation of unknown distribution is based on Kullback-Leibler Divergence:

$$K(f_\xi, f_\eta) = \int_{\mathbb{R}^n} f_\xi(x) \log \frac{f_\xi(x)}{f_\eta(x)}\, dx.$$

Kullback-Leibler Divergence

Kullback-Leibler (KL) Divergence:

$$K(f_\xi, f_\eta) = \int_{\mathbb{R}^n} f_\xi(x) \log \frac{f_\xi(x)}{f_\eta(x)}\, dx.$$

Properties of KL divergence:

1. $K(f_\xi, f_\eta) \geq 0$.
2. $K(f_\xi, f_\eta) = 0 \Leftrightarrow f_\xi = f_\eta$ almost everywhere.
3. $K(f_\xi, f_\eta)$ is <u>not</u> a distance as $K(f_\xi, f_\eta) \neq K(f_\eta, f_\xi)$.
4. KL divergence is additive:

Theorem

Let $\xi^n = \{\xi_1, \dots, \xi_n\}$ and $\eta^n = \{\eta_1, \dots, \eta_n\}$ be random vectors consisting of i.i.d. random variables. Then

$$K(f_{\xi^n}, f_{\eta^n}) = nK(f_{\xi_1}, f_{\eta_1}).$$

Pinsker Inequality

It is important to understand the relation between KL divergence and total variation distance.

Theorem (Pinsker Inequality)

If random variables ξ, η *have densities* $f_\xi(x)$, $f_\eta(x)$, $x \in \mathbb{R}^n$, *then*

$$D(f_\xi, f_\eta) \leq \sqrt{\frac{K(f_\xi, f_\eta)}{2}}.$$

Why KL Divergence is Useful?

- KL divergence is very convenient to use as it allows for efficient numerical optimization over the parameters of distributions.

- As we will see at the next lecture it gives basis for some of the most prominent statistical methods.

- Moreover, KL divergence is widely used in Information Theory, Machine Learning and many other sciences.

Thank you for your attention![1]

[1]Slides are partially based on lecture notes "Introduction to Mathematical Statistics" by Yury Golubev (in Russian).