Winter 2021 Math 106
Topics in Applied Mathematics
Data-driven Uncertainty Quantification

Yoonsang Lee (yoonsang.lee@dartmouth.edu)

Lecture 5: Nonparametric Inference

5.1 Empirical Distribution Function

Let $X_1, X_2, ..., X_n$ be an independent, identically distributed (IID) sample from a distribution $P(x)$.

Goal of nonparametric inference: Infer $P(x)$ without assuming any special structure or parametrization for $P(x)$.

The **empirical distribution** $\hat{P}_n$, an estimator of P using the sample $\{X_i\}$ of size n, is the CDF that puts mass $1/n$ at each data point

$$\hat{P}_n(x) = \frac{\sum_i^n I(X_i \leq x)}{n}$$

where

$$I(X_i \leq x) = \begin{cases} 1 & \text{if } X_i \leq x, \\ 0 & \text{if } X_i > x. \end{cases}$$

5.1 Empirical Distribution Function

Let $X_1, X_2, ..., X_n$ be an independent, identically distributed (IID) sample from a distribution $P(x)$.

Goal of nonparametric inference: Infer $P(x)$ without assuming any special structure or parametrization for $P(x)$.

The **empirical distribution** $\hat{P}_n$, an estimator of P using the sample $\{X_i\}$ of size n, is the CDF that puts mass $1/n$ at each data point

$$\hat{P}_n(x) = \frac{\sum_i^n I(X_i \leq x)}{n}$$

where

$$I(X_i \leq x) = \begin{cases} 1 & \text{if } X_i \leq x, \\ 0 & \text{if } X_i > x. \end{cases}$$

Exercise. Show that
$E(\hat{P}_n(x)) = P(x)$ and $Var(\hat{P}_n(x)) = \frac{P(x)(1-P(x))}{n}$.

5.1 Empirical Distribution Function

Theorem. (Glivenko-Cantelli) For each x and $\epsilon > 0$,

$$\mu(|\hat{P}(x) - P(x)| \geq \epsilon) \to 0 \text{ as } n \to \infty.$$

5.2 Curve Estimation (Smoothing)

Goal of curve estimation: Approximate the unknown density from a sample.

An example of curve estimation: Histograms.

Let $g(x)$ be the unknown true density and $\{X_i\}$ be IID of size n from $g(x)$. The estimator of g using $\{X_i\}$ is denoted by

$$\hat{g}(x; \{X_i\})$$

For simplicity, we often use $\hat{g}_n(x)$ for $\hat{g}(x; \{X_i\})$.

5.2 Curve Estimation (Smoothing)

Goal of curve estimation: Approximate the unknown density from a sample.
An example of curve estimation: Histograms.
Let $g(x)$ be the unknown true density and $\{X_i\}$ be IID of size n from $g(x)$. The estimator of g using $\{X_i\}$ is denoted by

$$\hat{g}(x; \{X_i\})$$

For simplicity, we often use $\hat{g}_n(x)$ for $\hat{g}(x; \{X_i\})$.
Integrated squared error

$$L(g, \hat{g}_n) = \int (g(u) - \hat{g}_n(u))^2 du.$$

Risk (or mean integrated squared error)

$$R(g, \hat{g}_n) = E[L(g, \hat{g}_n)].$$

5.2 Curve Estimation (Smoothing)

The risk can be written as

$$R(g, \hat{g}_n) = \int b^2(x)dx + \int v(x)dx$$

where

$$b(x) = E[\hat{g}_n(x)] - g(x)$$

is the bias of $\hat{g}_n(x)$ **at a fixed** x and

$$v(x) = Var(\hat{g}_n(x))$$

is the variance of $\hat{g}_n(x)$ **at a fixed** x.

5.2.1 Histogram

Let $X_1, X_2, ..., X_n$ be IID on $[0, 1]$ with density p. Let m be the number of bins where each bin $B_i, i = 1, 2, ..., m$ is defined by $B_i = [\frac{i-1}{m}, \frac{i}{m})$.
Define the **binwidth** $h = 1/m$ and let ν_j be the number of observations in B_i and $\hat{p}_i = \frac{\nu_j}{n}$.
The **histogram estimator** is defined by

$$\hat{p}_n(x) = \frac{\hat{p}_i}{h} \text{ if } x \in B_i$$

which can be written succinctly as

$$\hat{p}_n(x) = \sum_{i=1}^{n} \frac{\hat{p}_i}{h} I(x \in B_i)$$

where $I(x \in B_i) = 1$ if $x \in B_i$ and 0 otherwise.

5.2.1 Histogram

Theorem. For fixed x, m, let B_j be the bin containing x. Then

$$E[\hat{p}_n(x)] = \frac{p_j}{h}$$

and

$$Var(\hat{p}_n(x)) = \frac{p_j(1-p_j)}{nh^2}.$$

Theorem. Suppose that $\int p'(x)^2 dx < \infty$. Then

$$R(\hat{p}_n, p) \approx \frac{h^2}{12}\int (p'(u))^2 du + \frac{1}{nh}.$$

The value h^* that minimizes this is

$$h^* = \frac{1}{n^{1/3}}\left(\frac{6}{\int (p'(u))^2 du}\right)^{1/3}$$

With this choice of binwidth,

$$R(\hat{p}_n, p) \approx \frac{C}{n^{2/3}}.$$

5.2.2 Kernel Density Estimation

Given a Kernel K and a positive bandwidth h, the kernel density estimator (KDE) is defined to be

$$\hat{p}(x) = \frac{1}{n} \sum_{i}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$$

- KDE is smoother than histograms.
- KDE also converges faster to the true density than histograms.

A **kernel** is defined to be any smooth function K such that

- $K(x) \geq 0$,
- $\int K(x)dx = 1$,
- $\int xK(x)dx = 0$, and
- $\sigma_K^2 = \int x^2 K(x)dx > 0$.

5.2.2 Kernel Density Estimation

Theorem Under some assumptions on p and K,

$$R(p, \hat{p}_n) \approx \frac{1}{4}\sigma_K^4 h^4 \int (p''(x))^2 + \frac{K^2(x)dx}{nh}$$

where $\sigma_K^2 = \int x^2 K(x)dx$. The optimal bandwidth is

$$h^* = \frac{c_1^{-2/5} c_2^{1/5} c_3^{-1/5}}{n^{1/5}}$$

where $c_1 = \int x^2 K(x)dx$, $c_2 = \int K(x)^2 dx$ and $c_3 = \int (p''(x))^2 dx$. With this choice of bandwidth,

$$R(p, \hat{p}_n) \approx \frac{c_4}{n^{4/5}}$$

for some constant $c_4 > 0$.

5.3 Regression

Let we have a sample $(X_1, Y_1), (X_1, Y_1), ..., (X_n, Y_n)$. Most of you are familiar with a regression function as the minimizer $r(x)$ of the residual sums of squares

$$RSS = \sum_i^n (y_i - r(x_i))^2.$$

- Our definition of the **regression function** $r(x)$ is

$$r(x) = E[Y|X = x] = \int yf(y|x)dy.$$

- We approach the regression as a statistical inference problem. That is, we infer the joint density of (X, Y), say $p(x, y)$, to estimate the conditional expected value.
- We will discuss (i) parametric and (ii) nonparametric regression functions.

5.3.1 Parametric Regression

For simplicity, we will consider only linear models.

- We assume that the **conditional density** of Y for a given $X = x$ is a Gaussian with a mean $\alpha_0 + \alpha_1 X$ and a variance σ^2

$$p(y|x) = \phi(y; \alpha_0 + \alpha_1 x, \sigma^2)$$

  where ϕ is a Gaussian density.

- Thus, the density is parametrized by α_0 and α_1,

$$p(y|x; \alpha_0, \alpha_1)$$

  and their joint density is

$$p(x, y) = p(y|x)p(x).$$

5.3.1 Parametric Regression

- The likelihood function is

$$\mathcal{L}_n(\alpha_0, \alpha_1) = \Pi_i^n p(y_i|x_i; \alpha_0, \alpha_1)p(x_i)$$

- Log-likelihood function is

$$l_n(\alpha_0, \alpha_1) = \sum_i^n \ln p(y_i|x_i; \alpha_0, \alpha_1) + \sum_i^n p(x_i)$$

- The last term is independent of the parameters.
- Thus, MLE is the maximizer of the following

$$-\sum_i^n (y_i - \alpha_0 - \alpha_1 x_i)^2,$$

that is, the minimizer of RSS.

5.3.2 Nonparametric Regression

- The definition of the regression function does not change. The regression function $r(x)$ is the conditional expected value of Y

$$r(x) = E[Y|X = x].$$

- Estimate the joint density $p(x, y)$ using a nonparametric method, for example, KDE.
- Use the estimated density for the calculation of the regression function

$$r(x) = E[Y|X = x] = \int yp(y|x)dy = \frac{\int yp(x, y)dy}{\int p(x, y)dy}$$

5.3.2 Nonparametric Regression

The Nadaraya-Watson nonparametric regression.

$$\hat{r}(x) = \sum_{i}^{n} w_i(x) y_i$$

where K is a Kernel and the weights $w_i(x)$ are given by

$$w_i(x) = \frac{K(\frac{x-x_i}{h})}{\sum_{j=1}^{n} K(\frac{x-x_j}{h})}.$$

5.3.2 Nonparametric Regression

The Nadaraya-Watson nonparametric regression.

$$\hat{r}(x) = \sum_{i}^{n} w_i(x) y_i$$

where K is a Kernel and the weights $w_i(x)$ are given by

$$w_i(x) = \frac{K(\frac{x-x_i}{h})}{\sum_{j=1}^{n} K(\frac{x-x_j}{h})}.$$

Exercise. Derive the Nadaraya-Watson nonparametric regression.

5.4 Bootstrap

The **bootstrap** is a method for estimating standard errors **se**, i.e., the standard deviation of an estimator $\hat{T}$.

5.4 Bootstrap

The **bootstrap** is a method for estimating standard errors **se**, i.e., the standard deviation of an estimator $\hat{T}$.

1. Estimate $Var_P(T_n)$ with $Var_{\hat{P}}(T_n)$.
2. Approximate $Var_{\hat{P}}(T_n)$ using simulation.

$Var_P(T_n)$ is the variance of T_n with respect to P.

5.4 Bootstrap

The **bootstrap** is a method for estimating standard errors **se**, i.e., the standard deviation of an estimator $\hat{T}$.

1. Estimate $Var_P(T_n)$ with $Var_{\hat{P}}(T_n)$.
2. Approximate $Var_{\hat{P}}(T_n)$ using simulation.

$Var_P(T_n)$ is the variance of T_n with respect to P.
How do we estimate $Var_{\hat{P}}(T_n)$?

1. Draw $\{X_i^*\}$ from $\hat{P}$.
2. Compute T_n^* using $\{X_i^*\}$.
3. Repeat steps 1 and 2 M times, $T_{n,1}^*, \ldots, T_{n,M}^*$.
4. Estimate $Var_{\hat{P}}(T_n) = \frac{1}{M}\sum_m^M \left(T_{n,m}^* - \frac{1}{M}\sum T_{n,m}^*\right)^2$

Homework

1. Find and learn a KDE library of your choice.
2. Let X be a random variable with a density $\frac{1}{3}\phi(x; 0, 1) + \frac{2}{3}\phi(x; 1, 1)$ where $\phi(x; m, \sigma^2)$ is a Gaussian density with a mean m and a variance σ^2.
3. Generate an IID sample of X.
4. From the sample, $\{X_i\}$, estimate the density using (i) histogram, and (ii) KDE.
5. Compute the relative entropy using the estimated densities.
6. Plot the relative entropy as a function of the sample size n.
7. Let $Y = X^2$. Find the density of Y (numerically and analytically).

   8-9 For a Gaussian distribution $N(1, 1)$, we estimate the mean using the sample mean of a sample $\{X_i\}$

$$\hat{m} = \frac{1}{n} \sum_i X_i.$$

8. Calculate the variance of $\hat{m}$.
9. Estimate the variance of $\hat{m}$ using the bootstrap.