18.650. Statistics for Applications
Fall 2016. Problem Set 9

Due Friday, Nov. 18 at 12 noon

Problem 1 Nonparametric regression with fixed design **(60 points)**

Consider a fixed and regular design on the interval $[0, 1]$:

$$x_i = \frac{i}{n}, \quad i = 0, \dots, n.$$

For each i, let $Y_i = f(x_i) + \varepsilon_i$, where

- f is an unknown function on $[0, 1]$;
- $\varepsilon_0, \dots, \varepsilon_n \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$,

for some $\sigma^2 > 0$.
Assume that the unknown function f is differentiable, and

$$|f'(x)| \le L, \quad \forall x \in [0, 1],$$

for some positive number L. The aim of this exercise is to estimate the regression function f.

Let $k \le n$ be some positive integer. For $i = 0, \dots, n$, let $I_i = \{j = 0, \dots, n : |j - i| \le k\}$.

1. Compute the size $|I_i|$ of I_i, for each $i \in \{0, \dots, n\}$ and show that

$$k + 1 \le |I_i| \le 2k + 1.$$

2. For $i = 0, \dots, n$ we estimate $f(x_i)$ by

$$\hat{f}_i = \frac{1}{|I_i|} \sum_{j \in I_i} Y_j.$$

a) Compute the variance of $\hat{f}_i$ and show that

$$Var(\hat{f}_i) \le \frac{\sigma^2}{k}.$$

b) Compute $\mathbb{E}[\hat{f}_i]$ and prove that the bias b_i of $\hat{f}_i$ satisfies

$$|b_i| \leq \frac{1}{|I_i|} \mathop{}_{j \in I_i} |f(x_j) - f(x_i)|.$$

c) Using the assumptions on f, conclude that

$$b_i^2 \leq \frac{L^2 k^2}{n^2}.$$

d) **Using the previous questions**, prove that the quadratic risk of $\hat{f}_i$ satisfies:

$$\mathbb{E}\left[(\hat{f}_i - f(x_i))^2\right] \leq \frac{L^2 k^2}{n^2} + \frac{\sigma^2}{k}.$$

e) What is the optimal choice of k, i.e., the one that minimizes the previous upper bound on the quadratic risk ?

f) For this choice of k, what is the speed of convergence of the quadratic risk of $\hat{f}_i$ to zero ?

g) Prove that if L and σ^2 are unknown, there still is a choice of k that does not depend on L and σ^2 for which the quadratic risk of $\hat{f}_i$ is still of order $n^{-2/3}$.

3. *(Optional question)* Define the estimator $\hat{f}$ of f as the piecewise linear function $\hat{f}$ on $[0,1]$ such that $\hat{f}(x_i) = \hat{f}_i$, $i = 0, \ldots, n$ (where k is again any integer between 1 and n). We define the integrated quadratic risk of $\hat{f}$ as:

$$R(\hat{f}, f) = \mathbb{E}\left[\int_0^1 \left(\hat{f}(x) - f(x)\right)^2 \mathrm{d}x\right].$$

Prove that

$$R(\hat{f}, f) \leq \frac{c_1 k^2}{n^2} + \frac{c_2}{k},$$

where c_1 and c_2 are positive constants that depend on L and σ^2 only. Conclude that there is a choice of k that leads to convergence to zero at the speed $n^{-2/3}$.

Problem 2 Nonparametric estimation of a density **(40 points)**

Let $X_1, \ldots, X_n$ be i.i.d. random variables in the interval $[0,1]$ with some unknown density f. Throughout this exercise, we will assume that f is differentiable and satisfies $|f(x)| \leq L$, $|f'(x)| \leq L$, $\forall x \in [0,1]$, where L is a fixed positive number.

Let $h > 0$. For $x \in [0,1]$, we define the estimator of $f(x)$ as

$$\hat{f}(x) = \frac{1}{2nh} \sum_{i=1}^{n} \mathbb{1}_{|X_i - x| \leq h}.$$

Let $x \in (h, 1-h)$.

1. What is the distribution of each random variable $\mathbb{1}_{|X_i - x| \leq h}$, $i = 1, \ldots, n$?
2. Denote by $b(x)$ the bias of $\hat{f}(x)$ and by $\sigma^2(x)$ its variance. Recall the relationship between its quadratic risk, $b(x)$ and $\sigma^2(x)$.
3. Prove that
$$b(x) = \frac{F(x+h) - F(x-h) - 2hf(x)}{2h},$$
where F is the cdf of X_1.
4. Using a Taylor formula, conclude that
$$|b(x)| \leq \frac{Lh}{2}.$$
5. Show that
$$\sigma^2(x) \leq \frac{L}{2nh}.$$
6. Using the previous questions, give an upper bound for the quadratic risk of $\hat{f}(x)$.
7. Show that if L is unknown, the window size h can be taken such that the quadratic risk of $\hat{f}(x)$ is bounded from above by $Cn^{-2/3}$, where C is a positive constant that depends on L.

MIT OpenCourseWare
https://ocw.mit.edu

18.650 / 18.6501 Statistics for Applications
Fall 2016

For information about citing these materials or our Terms of Use, visit: https://ocw.mit.edu/terms.