Math 547: Mathematical Foundations of Statistical Learning Theory Fall 2017

Lecture 4 — August 28

Instructor: Stas Minsker *Scribe: Mose Wintner*

4.1 Nadaraya-Watson Estimator

$$
\begin{aligned}
&X \sim \Pi \text{ a distribution on } [0,1]^d \\
&p \text{ the density of } \Pi \\
\eta(x) &= \mathbb{E}(Y|X=x) = \int y \frac{p(x,y)}{p(x)}\, dy \\
\widehat{p}_n(x,y) &:= \frac{1}{n}\sum_{j=1}^{n} K_h(x-X_j)K_h(y-Y_j). \\
\widehat{\eta}_n(x) &:= \int y \frac{\widehat{p}(x,y)}{p(x)}\, dy \\
&= \int y \frac{1}{np(x)} \sum_{j=1}^{n} K_h(x-X_j)K_h(y-Y_j)\, dy \\
&= \frac{1}{n}\sum_{j=1}^{n} \int yK_h(y-Y_j)\frac{K_h(x-X_j)}{p(x)}\, dy \\
&= \frac{1}{n}\sum_{j=1}^{n} Y_j \frac{K_h(x-X_j)}{p(x)}.
\end{aligned}
$$

using the kernel property $\mathbb{E}K(y)=0$. To assess the estimator $\widehat{\eta}_n(x)$, we want to bound the maximum MSE

$$\sup_{x\in[0,1]^d} \mathbb{E}(\widehat{\eta}_n(x)-\eta(x))^2.$$

Theorem 1. Assume that

1. $0 < \underline{c} \le p(x) \le \overline{C} < \infty$
2. $\eta(x)p(x)$ is Lipschitz continuous with Lipschitz constant L.

Then there exists $b>0$ such that for all $x \in [0,1]^d$,

$$\mathbb{E}(\widehat{\eta}_n(x)-\eta(x))^2 \le b\left(h^2 + \frac{1}{nh^d}\right).$$

The two terms on the right hand side come from the bias and variance of the estimator, respectively. The "optimal" value of h makes the two terms equal, i.e.

$$\widehat{h} = n^{-\frac{1}{2+d}},$$

giving a convergence rate of $n^{-\frac{2}{2+d}}$. If η is β times differentiable, we can improve the convergence rate to $n^{-\frac{2\beta}{2\beta+d}}$.

$$\begin{aligned}
\widehat{\eta}_n(x) - \eta(x) &= \widehat{\eta}_n(x) - \mathbb{E}\widehat{\eta}_n(x) + \mathbb{E}\widehat{\eta}_n(x) - \eta(x) \\
\mathbb{E}\widehat{\eta}_n(x) - \eta(x) &= \mathbb{E}\frac{1}{n}\sum_{j=1}^{n} Y_j \frac{K_h(x-X_j)}{p(x)} - \eta(x) \\
&= \mathbb{E}[\mathbb{E}Y_1 \frac{K_h(x-X_1)}{p(x)}|X_1] - \eta(x) = \mathbb{E}[\eta(X_1)\frac{K_h(x-X_1)}{p(x)} - \eta(x)] \\
&= \int \eta(y)\frac{K_h(x-y)}{p(x)} p(y)\,dy - \eta(x) \\
&= \frac{\int(\eta(y)p(y) - \eta(x)p(x))K_h(x-y)\,dy}{p(x)}
\end{aligned}$$

which is bounded in magnitude by

$$\frac{1}{p(x)}\int L||x-y||_2 K_h(x-y)\,dy \leq \frac{1}{\underline{c}} b(K) h$$

where $b(K)$ is a constant depending on K.

Let $Z_j = Y_j \frac{K_h(x-X_j)}{p(x)}$. Then

$$\begin{aligned}
\widehat{\eta}_n(x) - \mathbb{E}\widehat{\eta}_n(x) &= \frac{1}{n}\sum_{j=1}^{n}\left(Y_j\frac{K_h(x-X_j)}{p(x)} - \mathbb{E}\widehat{\eta}_n(x)\right) \\
\mathbb{E}(\widehat{\eta}_n(x) - \mathbb{E}\widehat{\eta}_n(x))^2 &= \mathrm{Var}(\frac{1}{n}\sum_{j=1}^{n} Z_j) = \frac{1}{n}\mathrm{Var} Z_1 \leq \frac{1}{n}\mathbb{E}Z_1^2. \\
\mathbb{E}Z_1^2 &= \mathbb{E}Y_1 \frac{K_h(x-X_j)}{p^2(x)} \\
&\leq \frac{1}{\underline{c}^2}\mathbb{E}\frac{1}{h^{2d}}K^2((x-X_1)/h) \\
&= \frac{1}{\underline{c}^2 h^d}\mathbb{E}\frac{1}{h^d}K^2((x-X_1)/h)
\end{aligned}$$

Goal: Find a good prediction rule $\widehat{g}$ such that $P(Y \neq \widehat{g}(X))$ is small.
We are given $(X_1, Y_1), \ldots (X_n, Y_n)$ iid from P.

$$P(Y \neq g(X)) = \mathbb{E}I\{Y \neq g(X)\} \approx \frac{1}{n}\sum_{j=1}^{n} I\{Y_j \neq g(X_j)\}.$$

Approach: minimize the right hand expression over all g. Of course, we can just take this $\tilde{g}$ to send each X_j to Y_j and everything else to 0 to make this expression equal to 0. Then $P(Y \neq \tilde{g}(X)) = 1$ for any nontrivial distribution. Of course, this overfits. Instead of minimizing the risk over all measurable g, choose some "base class" G of functions and find the best $g \in G$.