Math 547: Mathematical Foundations of Statistical Learning Theory

Lectures — Week 2

Instructor: Stas Minsker *Scribe: Stas Minsker*

1.1 Nadaraya-Watson estimator and the curse of dimensionality.

The intuition behind the "curse of dimensionality" can be explained as follows: assume that you want to estimate a Lipschitz continuous function f on a d-dimensional unit cube. If nothing is known about this function, then, to get an estimate that is uniformly δ-close to f (up to a multiplicative constant), it is necessary to know the values of f on a δ-net inside the cube (i.e., if you partition the unit cube into smaller cubes with side length δ, you need at least one point in every cube). The cardinality of such a δ-net grows as δ^{-d}.

We will now make this intuition more formal when applied to our problem. Assume that $X \in \mathbb{R}^d$ is a random vector, and assume that Π, the distribution of X, is absolutely continuous with respect to Lebesgue measure with density $p(x)$. Suppose that $X_1, \dots, X_n$ be an iid sample from Π ("training data"), and we want to estimate p based on it. Let $K : \mathbb{R}^d \mapsto \mathbb{R}$ be a compactly supported kernel that satisfies

1. $\int_{\mathbb{R}^d} K(x)dx = 1$,
2. $\int_{\mathbb{R}^d} xK(x)dx = 0$,
3. $\int_{\mathbb{R}^d} \|x\|_2^2 K(x)dx < \infty$.

For example, we can take K to be the characteristic function of a unit cube:

$$K(x) = I\{\|x\|_\infty \le 1/2\}.$$

Next, define the transformation $K_h(x) = \frac{1}{h^d}K(x/h)$ for $h > 0$, and note that $K_h(x)$ is also a kernel satisfying our assumptions. One can estimate p via the *kernel density estimator* defined as

$$\hat{p}_n(x) := \frac{1}{n}\sum_{j=1}^n K_h(x - X_j) = \frac{1}{nh^d}\sum_{j=1}^n K\left(\frac{x - X_j}{h}\right).$$

Note that

$$\mathbb{E}p_n(x) = (p * K_h)(x) = \frac{1}{h^d}\int_{\mathbb{R}^d} K\left(\frac{x-y}{h}\right)p(y)dy.$$

One can think about this convolution as "local averaging": given a small $h > 0$, we replace $p(x)$ by an average value of p over a h-size "window" (cube) around x. As $h \to 0$,

$$(p * K_h)(x) \to p(x)$$

if p is sufficiently smooth (e.g., Hölder continuous).

Similar approach can be used to define an estimator of the regression function. Indeed, the regression function is just the conditional expectation, in case of binary classification

$$\hat{\eta}(x) = \mathbb{E}(Y\,|X = x) = 1 \cdot \Pr(Y = 1|X = x) + (-1) \cdot \Pr(Y = -1|X = x).$$

Assuming that $p(x)$ is known, one can estimate $\Pr(Y = y|X = x)$ by $\frac{1}{nh^d}\sum_{j=1}^n \frac{K\left(\frac{x-X_j}{h}\right)I(Y_j=y)}{p(x)}$. Then the estimator of $\eta(x)$ is

$$\begin{aligned}\hat{\eta}_n(x) &= \sum_{i=1}^2 y_i \frac{1}{nh^d}\sum_{j=1}^n \frac{K\left(\frac{x-X_j}{h}\right)I(Y_j = y_i)}{p(x)} \\ &= \frac{1}{nh^d}\sum_{j=1}^n \frac{K\left(\frac{x-X_j}{h}\right)}{p(x)} \sum_{i=1}^2 y_i I(Y_j = y_i) \\ &= \frac{1}{nh^d}\sum_{j=1}^n Y_j \frac{K\left(\frac{x-X_j}{h}\right)}{p(x)}.\end{aligned}$$

Estimator $\hat{\eta}(x)$ of this form is called the *Nadaraya-Watson estimator*. When $p(x)$ is unknown, one usually replaces it by the kernel density estimator (**exercise:** show that the Nadaraya-Watson estimator estimator looks exactly the same when (X, Y) has absolutely continuous distribution with a density $p_{X,Y}(x, y)$).

Theorem 1. Assume that Π is supported on the unit cube $[0, 1]^d$ with a given density p such that

$$0 < \delta \le p(x) \le D < \infty$$

for $x \in [0, 1]^d$. [1] Moreover, suppose that $\eta(x)p(x)$ is Lipschitz continuous with a Lipschitz constant L. Then for every $x \in [0, 1]^d$,

$$\mathbb{E}(\hat{\eta}_n(x) - \eta(x))^2 \le C\left(\frac{1}{nh^d} + h^2\right),$$

where C is a constant which depends only on δ, D, L and the kernel K.

Remark 1. Minimizing the right-hand side over all $h > 0$, we see that for $h_n = n^{-\frac{1}{2+d}}$,

$$\mathbb{E}(\hat{\eta}_n(x) - \eta(x))^2 \le Cn^{-\frac{1}{2+d}}.$$

The bound of the theorem is sharp in the worst case.

Remark 2. If η is β times continuously differentiable, one can improve the rate to

$$n^{-\left(\frac{2\beta}{2\beta+d}\right)}.$$

[1] $[0, 1]^d$ can be replaced by an arbitrary compact subset of $\mathbb{R}^d$.

Proof. Note that

$$\hat{\eta}_n(x) - \eta(x) = \hat{\eta}_n(x) - \mathbb{E}\hat{\eta}_n(x) + \mathbb{E}\hat{\eta}_n(x) - \eta(x)$$

and

$$\mathbb{E}(\hat{\eta}_n(x) - \eta(x))^2 = \mathbb{E}(\hat{\eta}_n(x) - \mathbb{E}\hat{\eta}_n(x))^2 + (\mathbb{E}\hat{\eta}_n(x) - \eta(x))^2 = (1) + (2),$$

where (1) corresponds to the variance, and (2) corresponds to the bias.

For (1) we have

$$\begin{aligned}
\operatorname{Var}\hat{\eta}_n(x) &= n \operatorname{Var}\left(\frac{1}{nh^d} Y_1 \frac{K\left(\frac{x-X_1}{h}\right)}{p(x)}\right) = \frac{1}{nh^{2d}} \operatorname{Var}\left(Y_1 \frac{K\left(\frac{x-X_1}{h}\right)}{p(x)}\right) \\
&\leq \frac{1}{nh^{2d}} \mathbb{E}\left(Y_1^2 \frac{K^2\left(\frac{x-X_1}{h}\right)}{p^2(x)}\right) \leq \frac{1}{n\delta^2}\frac{1}{h^{2d}} \underbrace{\mathbb{E}K^2\left(\frac{x-X_1}{h}\right)}_{\int K^2\left(\frac{x-y}{h}\right)p(y)dy \leq h^d \int DK^2\left(\frac{x-y}{h}\right)dy} \\
&\leq \frac{D}{n\delta^2}\frac{1}{h^{2d}}h^d = \frac{D}{n\delta^2}\cdot\frac{1}{h^d}.
\end{aligned}$$

For the second term note that

$$\begin{aligned}
\mathbb{E}\,\hat{\eta}_n(x) &= \frac{1}{nh^d}\mathbb{E}\sum_{j=1}^n Y_j \frac{K\left(\frac{x-X_j}{h}\right)}{p(x)} = \frac{1}{h^d}\mathbb{E}\mathbb{E}\left(Y_1 \frac{K\left(\frac{x-X_1}{h}\right)}{p(x)} | X_1\right) \\
&= \frac{1}{h^d}\mathbb{E}\left(\frac{K\left(\frac{x-X_1}{h}\right)}{p(x)}\eta(X_1)\right).
\end{aligned}$$

Hence,

$$\begin{aligned}
|\mathbb{E}\hat{\eta}_n(x) - \eta(x)| &= \left|\frac{1}{h^d}\int_{[0,1]^d} \frac{K\left(\frac{x-y}{h}\right)\eta(y)p(y)}{p(x)}dy - \eta(x)\cdot \underbrace{\frac{1}{h^d}\int_{[0,1]^d}\frac{K\left(\frac{x-y}{h}\right)p(x)}{p(x)}dy}_{=1}\right| \\
&\leq \frac{1}{h^d}\int_{[0,1]^d}\left|\frac{\eta(y)p(y)-\eta(x)p(x)}{p(x)}\right| \cdot K\left(\frac{x-y}{h}\right)dy \\
&\leq \frac{Lh}{\delta}\int_{[0,1]^d}\left\|\frac{x-y}{h}\right\| K\left(\frac{x-y}{h}\right) d\left(\frac{y}{h}\right) \\
&= \frac{Lh}{\delta}\underbrace{\int_{[0,1]^d}\|z\|K(z)d(-z)}_{<\infty} = C(L,\delta,K)\cdot h.
\end{aligned}$$

Here, we used that $|\eta(y)p(y) - \eta(x)p(x)| \leq L\|x-y\|$, where $\|\cdot\|$ is the Euclidean 2-norm. □