1 Kernel Density Estimation

Recall from last time that given $X_1, X_2, \ldots, X_n$ i.i.d. $\sim f$, we define the Kernel density estimator as

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

(1)

Sometimes we may also refer to $\hat{f}_n(x)$ as $\hat{f}_h(x)$.

Notice that there are two degrees of freedom: the choice of $h$ and the choice of the kernel $K$.

**Definition 1.** The risk of a density estimator is defined as

$$L(h) = \int (\hat{f}_n - f)^2.$$

(2)

Sometimes the risk is defined as the expected value of $L(h)$, i.e., $\mathbb{E}[L(h)]$.

Note that $L(h)$ is a random variable, and that in order to compute it one needs to know the true distribution $f$. In the case where a genie gives us the true distribution $f$, one could choose $h$ to be

$$h_g = \arg\min_h L(h),$$

(3)

in which case the loss of the genie is given by

$$\min_h L(h).$$

(4)

We sometimes consider instead

$$J(h) = \int \hat{f}_n^2 - 2 \int \hat{f}_n f,$$

(5)

and refer to it (or to $\mathbb{E}[J(h)]$) as the risk, although it differs from the true risk by the constant term $\int f^2(x)dx$.

It is clear that the $h$ that minimizes $L(h)$ is the same as the one that minimizes $J(h)$, i.e.,

$$h_g = \arg\min_h L(h) = \arg\min_h J(h).$$

(6)

It is possible to estimate $J(h)$ from the data (without the need of the true density $f$) as follows

$$\hat{J}(h) = \int \hat{f}_n^2 - \frac{2}{n} \sum_{i=1}^n \hat{f}_{(-i)}(X_i),$$

(7)

where $\hat{f}_{(-i)}$ is the density estimator obtained after removing the $i^{th}$ observation.

**Definition 2.** We refer to $\hat{J}(h)$ given by (7) as the cross-validation estimator of risk.
Figure 1: $J(h)$ and its approximation $\hat{J}(h)$ vs $h$. $h_g$ and $\hat{h}$ are their corresponding minimizes.

We proved that if the kernel is bounded ($||K||_\infty \leq 1$), then

$$P(|\hat{J}(h) - J(h)| \geq \epsilon) \leq n c_1(\epsilon, h) e^{-n \frac{\epsilon^2}{8h^2}} \quad (8)$$

Since in practice the true distribution $f$ is not available, we choose $h$ as

$$\hat{h} = \arg \min_h \hat{J}(h), \quad (9)$$

and hope that it does not differ by much from $h_g$ (see Fig. 1).

We would like to have something like

$$\lim_{n \to \infty} (L(\hat{h}) - L(h_g)) = 0 \quad (10)$$

However, as $n$ becomes large, more data is available, and we know $\hat{f}_n$ can get arbitrarily close to $f$. We proved that

$$\int |\hat{f}_n - f| \to 0 \quad (11)$$

as $nh \to \infty$ and $h \to 0$, which implies converge in L2

$$\int (\hat{f}_n - f)^2 \to 0. \quad (12)$$

Therefore, we know one can get $L(h)$ to converge to zero even data independently. A fortiori $L(h_g)$ goes to zero and we’d like to be able to assert about our data-dependent choice of $h$ something much stronger than that $L(\hat{h})$ goes to zero.

Instead, we therefore look at the ratio between $L(\hat{h})$ and $L(h_g)$. Specifically, we want to show that it goes to one.

**Theorem 3.** Stone’s Theorem (1984): If the true density $f$ satisfies $||f||_\infty \leq \infty$, then

$$\lim_{n \to \infty} \frac{L(\hat{h})}{L(h_g)} = 1 \text{ w.p.1.} \quad (13)$$

**Proof** We present a very rough outline of the proof (see [Stone 1984] for a complete proof):
Exercise 4.

Define $\tilde{K}$

We now specify a way to compute $\hat{J}(h)$.

Computing $c$

The inequality (a) follows because we can find two constants $c_n$ and $C_n$ such that $c_n \leq h \leq C_n$ and $c_n \leq h_y \leq C_n$ with high probability (w.h.p.).

The inequality (b) follows by looking at a very fine collection of points (instead of the whole interval) s.t. the maximum is essentially the same. We need to look closely at $\tilde{J}(h)$ as a function of $h$. It is possible to find a large $M_n$ and a small $\delta_n$ s.t. the inequality follows. Basically, we are dividing the interval $[c_n, C_n]$ into $M_n$ points that are close to each other.

For the last inequality (c), we use the union bound.

Thus we have $L(\hat{h}) - \min_h L(h) \leq c_n \min_h L(h) + \delta_n$ w.h.p. Therefore, we can conclude that

$$0 \leq \frac{L(\hat{h})}{\min_h L(h)} - 1 \leq 2c_n \text{ w.h.p., for large } n.$$  \hfill (22)

The proof can be completed using the Borel-Cantelli Lemma.

For more about Empirical processes, refer to the Uniform Laws of Large Numbers [2] (Chapter 9 and 19).

1.1 Computing $\hat{J}(h)$

We now specify a way to compute $\hat{J}(h)$.

$$\hat{J}(h) = \min h \int \hat{f}_n^2 - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{(-i)}(X_i)$$

$$= \frac{1}{n^2h^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K(\frac{x-X_i}{h})K(\frac{x-X_j}{h})dx - \frac{2}{n} \sum_{i=1}^{n} \frac{1}{(n-1)h} \sum_{i \neq j} K(\frac{X_i - X_j}{h})$$  \hfill (23)

Exercise 4. Define $\bar{J}(h)$ as follows

$$\bar{J}(h) = \frac{1}{nh^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K^*(\frac{X_i - X_j}{h}) + \frac{2}{nh} K(0) + O\left(\frac{1}{n^2}\right).$$  \hfill (25)

where $K^*(x) = K^{(2)}(x) - 2K(x)$, with $K^{(2)}(z) = \int K(z-y)K(y)dy$.

Show that $\bar{J}(h)$ is a very good approximation of $\hat{J}(h)$.  

3
References
