STAT 425: Introduction to Nonparametric Statistics **Winter 2018**

Lecture 7: Density Estimation: k-Nearest Neighbor and Basis Approach

Instructor: Yen-Chi Chen

Reference: Section 8.4 of *All of Nonparametric Statistics*.

7.1 k-nearest neighbor

k-nearest neighbor (k-NN) is a cool and powerful idea for nonparametric estimation. Today we will talk about its application in density estimation. In the future, we will learn how to use it for regression analysis and classification.

Let $X_1, \cdots, X_n \sim p$ be our random sample. Assume each observation has d different variables; namely, $X_i \in \mathbb{R}^d$. For a given point x, we first rank every observation based on its distance to x. Let $R_k(x)$ denotes the distance from x to its k-th nearest neighbor point.

For a given point x, the kNN density estimator estimates the density by

$$\widehat{p}_{\mathsf{knn}}(x) = \frac{k}{n} \cdot \frac{1}{V_d \cdot R_k^d(x)} = \frac{k}{n} \cdot \frac{1}{\text{Volume of a } d\text{-dimensional ball with radius being } R_k(x)},$$

where $V_d = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}$ is the volume of a unit d-dimensional ball and $\Gamma(x)$ is the Gamma function.

Here are the results when $d = 1, 2$, and 3 –

- $d = 1$, $V_1 = 2$: $\widehat{p}_{\mathsf{knn}}(x) = \frac{k}{n}\frac{1}{2R_k(x)}$.
- $d = 2$, $V_1 = \pi$: $\widehat{p}_{\mathsf{knn}}(x) = \frac{k}{n}\frac{1}{\pi R_k^2(x)}$.
- $d = 3$, $V_1 = \frac{4}{3}\pi$: $\widehat{p}_{\mathsf{knn}}(x) = \frac{k}{n}\frac{3}{4\pi R_k^3(x)}$.

What is the intuition of a kNN density estimator? By the definition of $R_k(x)$, the ball centered at x with radius $R_k(x)$

$$B(x, R_k(x)) = \{y : \|x - y\| \leq R_k(x)\}$$

satisfies the fact that

$$\frac{k}{n} = \frac{1}{n}\sum_{i=1}^{n} I(X_i \in B(x, R_k(x))).$$

Namely, ratio of observations within $B(x, R_k(x))$ is k/n.

Recall from the relation between EDF and CDF, the quantity

$$\frac{1}{n}\sum_{i=1}^{n} I(X_i \in B(x, R_k(x)))$$

can be viewed as an estimator of the quantity

$$P(X_i \in B(x, R_k(x))) \approx \int_{B(x,R_k(x))} p(y)dy.$$

When n is large and k is relatively small compared to n, $R_k(x)$ will be small because the ratio $\frac{k}{n}$ is small. Thus, the density $p(y)$ within the region $B(x, R_k(x))$ will not change too much. Namely, $p(y) \approx p(x)$ for every $y \in B(x, R_k(x))$. Note that $p(x)$ is the center of the ball $B(x, R_k(x))$.

Therefore,

$$P(X_i \in B(x, R_k(x))) \approx \int_{B(x,R_k(x))} p(y)dy \approx p(x) \int_{B(x,R_k(x))} dy = p(x) \cdot V_d \cdot R_k^d(x).$$

This quantity will be the target of the estimator $\frac{1}{n}\sum_{i=1}^n I(X_i \in B(x, R_k(x)))$, which equals to $\frac{k}{n}$. As a result, we can say that

$$p(x) \cdot V_d \cdot R_k^d(x) \approx P(X_i \in B(x, R_k(x))) \approx \frac{1}{n}\sum_{i=1}^n I(X_i \in B(x, R_k(x))) \approx \frac{k}{n},$$

which leads to

$$p(x) \cdot V_d \cdot R_k^d(x) \approx \frac{k}{n} \Rightarrow p(x) \approx \frac{k}{n}\frac{1}{V_d \cdot R_k^d(x)}$$

This motivates us to use

$$\widehat{p}_{\mathsf{knn}}(x) = \frac{k}{n}\frac{1}{V_d \cdot R_k^d(x)}$$

as a density estimator.

Example. We consider a simple example in $d = 1$. Assume our data is $\mathcal{X} = \{1, 2, 6, 11, 13, 14, 20, 33\}$. What is the kNN density estimator at $x = 5$ with $k = 2$? First, we calculate $R_2(5)$. The distance from $x = 5$ to each data point in $\mathcal{X}$ is

$$\{4, 3, 1, 6, 8, 9, 15, 28\}.$$

Thus, $R_2(5) = 3$ and

$$\widehat{p}_{\mathsf{knn}}(5) = \frac{2}{8}\frac{1}{2 \cdot R_2(5)} = \frac{1}{24}.$$

What will the density estimator be when we choose $k = 5$? In this case, $R_5(5) = 8$ so

$$\widehat{p}_{\mathsf{knn}}(5) = \frac{5}{8}\frac{1}{2 \cdot R_5(5)} = \frac{5}{64}.$$

Now we see that different value of k gives a different density estimate even at the same x. How do we choose k? Well, just as the smoothing bandwidth in the KDE, it is a very difficult problem in practice. However, we can do some theoretical analysis to get a rough idea about how k should be changing with respect to the sample size n.

7.1.1 Asymptotic theory

The asymptotic analysis of a k-NN estimator is quiet complicated so here I only stated its result in $d = 1$. The bias of the k-NN estimator is

$$\mathbf{bias}(\widehat{p}_{\mathsf{knn}}(x)) = \mathbb{E}(\widehat{p}_{\mathsf{knn}}(x)) - p(x) = b_1\frac{p''(x)}{p^2(x)}\left(\frac{k}{n}\right)^2 + b_2\frac{p(x)}{k} + o\left(\left(\frac{k}{n}\right)^2 + \frac{1}{k}\right),$$

where b_1 and b_2 are two constants. The variance of the k-NN estimator is

$$\mathsf{Var}(\widehat{p}_{\mathsf{knn}}(x)) = v_0 \cdot \frac{p^2(x)}{k} + o\left(\frac{1}{k}\right),$$

where v_0 is a constant. The quantity k is something we can choose. We need $k \to \infty$ when $n \to \infty$ to make sure both bias and variance converge to 0. However, how k diverges affects the quality of estimation. When k is large, the variance is small while the bias is large. When k is small, the variance is large and the bias tends to be small but it could also be large (the second component in the bias will be large).

To balance the bias and variance, we consider the mean square error, which is at the rate

$$\mathsf{MSE}(\widehat{p}_{\mathsf{knn}}(x)) = O\left(\frac{k^4}{n^4} + \frac{1}{k}\right).$$

This motivates us to choose

$$k = C_0 \cdot n^{\frac{4}{5}}$$

for some constant C_0. This leads to the optimal convergence rate

$$\mathsf{MSE}(\widehat{p}_{\mathsf{knn,opt}}(x)) = O(n^{-\frac{4}{5}})$$

for a k-NN density estimator.

Remark.

- When we consider a d-dimensional data, the bias will be of the rate

  $$\mathbf{bias}(\widehat{p}_{\mathsf{knn}}(x)) = O\left(\left(\frac{k}{n}\right)^{\frac{2}{d}} + \frac{1}{k}\right)$$

  and the variance is still at rate

  $$\mathsf{Var}(\widehat{p}_{\mathsf{knn}}(x)) = O\left(\frac{1}{k}\right).$$

  This shows a very different phenomenon compared to the KDE. In KDE, the rate of variance depends on the dimension whereas the bias remains the same. In kNN, the rate of variance stays the same rate but the rate of bias changes with respect to the dimension. One intuition is that no matter what dimension is, the ball $B(x, R_k(x))$ always contain k observations. Thus, the variability of a kNN estimator is caused by k points, which is independent of the dimension. On the other hand, in KDE, the same h in different dimensions will cover different number of observations so the variability changes with respect to the dimension.

- The kNN approach has an advantage that it can be computed very efficiently using kd-tree algorithm[1]. This is a particularly useful feature when we have a huge amount of data and when the dimension of the data is large. However, a downside of the kNN is that the density often has a 'heavy-tail', which implies it may not work well when $|x|$ is very large. Moreover, when $d = 1$, the density estimator $\widehat{p}_{\mathsf{knn}}(x)$ is not even a density function (the integral is infinite!).

7.2 Basis approach

In this section, we assume that the PDF $p(x)$ is supported on $[0, 1]$. Namely, $p(x) > 0$ only in $[0, 1]$. When the PDF $p(x)$ is smooth (in general, we need p to be squared integrable, i.e., $\int_0^1 p(x)^2 dx < \infty$), we can use an orthonormal basis to approximate this function. This approach has several other names: the basis estimator, projection estimator, and an orthogonal series estimator.

[1] https://en.wikipedia.org/wiki/K-d_tree

Let $\{\phi_1(x), \phi_2(x), \cdots, \phi_m(x), \cdots\}$ be a set of basis functions. Then we have

$$p(x) = \sum_{j=1}^{\infty} \theta_j \phi_j(x).$$

The quantity θ_j is the coefficient of each basis. In signal process, these quantities are refereed to as the signal.

The collection $\{\phi_1(x), \phi_2(x), \cdots, \phi_m(x), \cdots\}$ is called a basis if its elements have the following property:

- Unit 1:
$$\int_0^1 \phi_j^2(x)dx = 1 \tag{7.1}$$
for every $j = 1, \cdots$.
- Orthonormal:
$$\int_0^1 \phi_j(x)\phi_k(x)dx = 0 \tag{7.2}$$
for every $j \neq k = 1, \cdots$.

Here are some concrete examples of the basis:

- Cosine basis:
$$\phi_1(x) = 1, \quad \phi_j(x) = \sqrt{2}\cos(\pi(j-1)x), j = 2, 3, \cdots,$$
- Trigonometric basis:
$$\phi_1(x) = 1, \quad \phi_{2j}(x) = \sqrt{2}\cos(2\pi jx), \quad \phi_{2j+1}(x) = \sqrt{2}\sin(2\pi jx), j = 1, 2, \cdots$$

Often the basis is something we can choose so it is known to us. What is unknown to use is the coefficients $\theta_1, \theta_2, \cdots,$. Thus, the goal is to estimate these coefficients using the random sample $X_1, \cdots, X_n$.

How do we estimate these parameters? We start with some simple analysis. For any basis $\phi_j(x)$, consider the following integral:

$$\begin{aligned}
\mathbb{E}(\phi_j(X_1)) &= \int_0^1 \phi_j(x)dP(x) \\
&= \int_0^1 \phi_j(x)p(x)dx \\
&= \int_0^1 \phi_j(x)\sum_{k=1}^{\infty}\theta_k\phi_k(x)dx \\
&= \sum_{k=1}^{\infty}\theta_k \int_0^1 \underbrace{\phi_j(x)\phi_k(x)dx}_{=0 \text{ except } k=j} \\
&= \sum_{k=1}^{\infty}\theta_k I(k=j) \\
&= \theta_j.
\end{aligned}$$

Namely, the expectation of $\phi_j(X_1)$ is exactly the coefficient θ_j. This motivates us to use the sample average as an estimator:

$$\widehat{\theta}_j = \frac{1}{n}\sum_{i=1}^{n}\phi_j(X_i).$$

By construction, this estimator is unbiased, i.e., $\mathbb{E}(\widehat{\theta}_j) - \theta_j = 0$. The variance of this estimator is

$$\begin{aligned}
\mathsf{Var}(\widehat{\theta}_j) &= \frac{1}{n}\mathsf{Var}(\phi_j(X_1)) \\
&= \frac{1}{n}\left(\mathbb{E}(\phi_j^2(X_1)) - \mathbb{E}^2(\phi_j(X_1))\right) \\
&= \frac{1}{n}(\mathbb{E}(\phi_j^2(X_1)) - \phi_j^2) \\
&= \frac{\sigma_j^2}{n},
\end{aligned}$$

where $\sigma_j^2 = \mathbb{E}(\phi_j^2(X_1)) - \phi_j^2$.

In practice, we cannot use all the basis because there will be infinite number of them being calculated. So we will use only M basis as our estimator. Namely, our estimator is

$$\widehat{p}_{n,M}(x) = \sum_{j=1}^{M}\widehat{\theta}_j\phi_j(x). \tag{7.3}$$

Later in the asymptotic analysis, we will show that we should not choose M to be too large because of the bias-variance tradeoff.

7.2.1 Asymptotic theory

To analyze the quality of our estimator, we use the mean integrated squared error (MISE). Namely, we want to analyze

$$\mathbf{MISE}(\widehat{p}_{n,M}) = \mathbb{E}(\int_0^1 (\widehat{p}_{n,M}(x) - p(x))^2 dx) = \int_0^1 \left[\mathbf{bias}^2(\widehat{p}_{n,M}(x)) + \mathsf{Var}(\widehat{p}_{n,M}(x))\right] dx$$

Analysis of Bias. Because each $\widehat{\theta}_j$ is an unbiased estimator of θ_j, we have

$$\mathbb{E}(\widehat{p}_{n,M}(x)) = \mathbb{E}(\sum_{j=1}^{M}\widehat{\theta}_j\phi_j(x)) = \sum_{j=1}^{M}\mathbb{E}(\widehat{\theta}_j)\phi_j(x) = \sum_{j=1}^{M}\theta_j\phi_j(x).$$

Thus, the bias at point x is

$$\mathbf{bias}(\widehat{p}_{n,M}(x)) = \mathbb{E}(\widehat{p}_{n,M}(x)) - p(x) = \sum_{j=1}^{M}\theta_j\phi_j(x) - \sum_{j=1}^{\infty}\theta_j\phi_j(x) = -\sum_{j=M+1}^{\infty}\theta_j\phi_j(x).$$

Thus, the integrated squared bias is

$$\begin{aligned}\int_0^1 \mathbf{bias}^2(\widehat{p}_{n,M}(x))dx &= \int_0^1 \left(-\sum_{j=M+1}^{\infty}\theta_j\phi_j(x)\right)^2 dx \\ &= \int_0^1 \left(\sum_{j=M+1}^{\infty}\theta_j\phi_j(x)\right)\left(\sum_{k=M+1}^{\infty}\theta_k\phi_k(x)\right)dx \\ &= \sum_{j=M+1}^{\infty}\sum_{k=M+1}^{\infty}\theta_j\theta_k \underbrace{\int_0^1 \phi_j(x)\phi_k(x)dx}_{=I(j=k)} \\ &= \sum_{j=M+1}^{\infty}\theta_j^2.\end{aligned} \tag{7.4}$$

Namely, the bias is determined by the *signal strength* of the ignored basis, which makes sense because the bias should be reflecting the fact that we are not using all the basis and if there are some important basis (the ones with large $|\theta_j|$) being ignored, the bias ought to be large.

We know that in KDE and kNN, the bias is often associated with the smoothness of the density function. How does the smoothness comes into play in this case? It turns out that if the density is smooth, the remaining signals $\sum_{j=M+1}^{\infty}\theta_j^2$ will also be small. To see this, we consider a very simple model by assuming that we are using the cosine basis and

$$\int_0^1 |p''(x)|^2 dx \le L_0 \tag{7.5}$$

for some constant $L_0 > 0$. Namely, the overall curvature of the density function is bounded.

Using the fact that for cosine basis function $\phi_j(x)$,

$$\begin{aligned}\phi_j'(x) &= -\sqrt{2}\pi(j-1)\sin(\pi(j-1)x), \\ \phi_j''(x) &= -\sqrt{2}\pi^2(j-1)^2\cos(\pi(j-1)x) = -\pi^2(j-1)^2\phi_j(x).\end{aligned}$$

Thus, equation (7.5) implies

$$\begin{aligned}L_0 &\ge \int_0^1 |p''(x)|^2 dx \\ &= \int_0^1 |\sum_{j=1}^{\infty}\theta_j\phi_j''(x)|^2 dx \\ &= \int_0^1 |\sum_{j=1}^{\infty}\pi^2(j-1)^2\theta_j\phi_j(x)|^2 dx \\ &= \int_0^1 \left(\sum_{j=1}^{\infty}\pi^2(j-1)^2\theta_j\phi_j(x)\right)\left(\sum_{k=1}^{\infty}\pi^2(k-1)^2\theta_k\phi_k(x)\right)dx \\ &= \pi^4\sum_{j=1}^{\infty}\sum_{k=1}^{\infty}(j-1)^2(k-1)^2\theta_j\theta_k \underbrace{\int_0^1 \phi_j(x)\phi_k(x)dx}_{=I(j=k)} \\ &= \pi^4\sum_{j=1}^{\infty}(j-1)^2\theta_j^2.\end{aligned}$$

Namely, equation (7.5) implies

$$\sum_{j=1}^{\infty}(j-1)^4\theta_j^2 \le \frac{L_0}{\pi^4}. \tag{7.6}$$

This further implies that

$$\sum_{j=M+1}^{\infty}\theta_j^2 = O(M^{-4}). \tag{7.7}$$

An intuitive explanation is as follows. Equation (7.6) implies that when $j \to \infty$, the signal $\theta_j^2 = O(j^{-5})$. The reason is: if $\theta_j^2 \approx \frac{1}{j^{-5}}$, equation (7.6) becomes $\sum_{j=1}^{\infty}(j-1)^4\theta_j^2 \approx \sum_{j=1}^{\infty}(j-1)^4\frac{1}{j^{-5}} \approx \sum_{j=1}^{\infty}\frac{1}{j} \to \infty$! Thus, the *tail* signals have to be shrinking toward 0 at rate $O(j^{-5})$. As a result, $\sum_{j=M+1}^{\infty}\theta_j^2 \approx \sum_{j=M+1}^{\infty}O(j^{-5}) = O(M^{-4})$.

Equation (7.7) and (7.4) together imply that the bias is at rate

$$\int_0^1 \mathbf{bias}^2(\widehat{p}_{n,M}(x))dx = O(M^{-4}).$$

Analysis of Variance. Now we turn to the analysis of variance.

$$\begin{aligned}
\int_0^1 \mathsf{Var}(\widehat{p}_{n,M}(x))dx &= \int_0^1 \mathsf{Var}\left(\sum_{j=1}^{M}\widehat{\theta}_j\phi_j(x)\right)dx \\
&= \int_0^1 \left(\sum_{j=1}^{M}\phi_j^2(x)\mathsf{Var}(\widehat{\theta}_j) + \sum_{j\neq k=1}^{M}\phi_j(x)\phi_k(x)\mathsf{Cov}(\widehat{\theta}_j,\widehat{\theta}_k)\right)dx \\
&= \sum_{j=1}^{M}\mathsf{Var}(\widehat{\theta}_j)\underbrace{\int_0^1\phi_j^2(x)dx}_{=1} + \sum_{j\neq k=1}^{M}\mathsf{Cov}(\widehat{\theta}_j,\widehat{\theta}_k)\underbrace{\int_0^1\phi_j(x)\phi_k(x)dx}_{=0} \\
&= \sum_{j=1}^{M}\mathsf{Var}(\widehat{\theta}_j) \\
&= \sum_{j=1}^{M}\frac{\sigma_j^2}{n} \\
&= O\left(\frac{M}{n}\right).
\end{aligned}$$

Now putting both bias and variance together, we obtain the rate of the MISE

$$\mathbf{MISE}(\widehat{p}_{n,M}) = \int_0^1\left[\mathbf{bias}^2(\widehat{p}_{n,M}(x)) + \mathsf{Var}(\widehat{p}_{n,M}(x))\right]dx = O\left(\frac{1}{M^4}\right) + O\left(\frac{M}{n}\right).$$

Thus, the optimal choice is $M = M^* = C_0 n^{1/5}$ for some positive constant C). And this leads to

$$\mathbf{MISE}(\widehat{p}_{n,M^*}) = O(n^{-4/5}),$$

which is the same rate as the KDE and kNN.

Remark.

- **(Sobolev space)** Again, we see that the bias is dependent on the smoothness of the density function. Here, as long as the density function has an overall curvature (second derivative) being bounded, we have an optimal MISE at the rate of $O(n^{-4/5})$. In fact, if the density function has stronger smoothness, such as the overall third derivatives being bounded, we will have an even faster convergence rate $O(n^{-6/7})$. Now for any positive integer β and a positive number $L > 0$, we define

$$W(\beta, L) = \left\{ p : \int_0^1 |p^{(\beta)}(x)|^2 dx \leq L < \infty \right\}$$

  be a collection of smooth density functions. Then the bias between a basis estimator and any density $p \in W(\beta, L)$ is at the rate $O(M^{-2\beta})$ and the optimal convergence rate is $O(n^{-\frac{2\beta}{2\beta+1}})$. The collection $W(\beta, L)$ is a space of functions and is known as the Sobolev Space.

- **(Tuning parameter)** The number of basis M in a basis estimator, the number of neighborhood k in the kNN approach, and the smoothing bandwidth h in the KDE all play a very similar role in bias-variance tradeoff. These quantities are called *tuning parameters* in statistics and machine learning. Just like what we have seen in the analysis of the KDE, choosing these parameters is often a very difficult task because the optimal choice depends on the actual density function, which is unknown to us. A principle approach to choosing the tuning parameters is based on minimizing an error estimator. For instance, in the basis estimator, we try to find an estimator for the MISE, $\widehat{\mathbf{MISE}}(\widehat{p}_{n,M})$. When we change value of M, this error estimator will also change. We then choose M by minimizing the error, i.e.,

$$\widehat{M}^* = \mathsf{argmin}_M \widehat{\mathbf{MISE}}(\widehat{p}_{n,M}).$$

  Then we pick $\widehat{M}^*$ basis and construct our estimator.