Lecture 5: K-Nearest Neigbhor Rules

Tianpei Xie

Dec. 19th., 2022

Contents

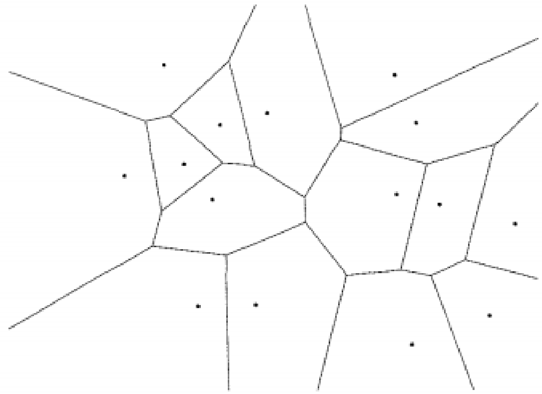

Figure 1: Varona partition of K-NN rules [Devroye et al., 2013].

1 Nearest Neighbor Rules

1.1 The Classification Rule

- **Definition** (***Nearest Neighbor Rules***)
  Formally, we define ***the k-NN rule*** by

$$g_n(x) = \begin{cases} 1 & \sum_{i=1}^{n} w_{n,i} \mathbb{1}\left\{Y_i = 1\right\} > \sum_{i=1}^{n} w_{n,i} \mathbb{1}\left\{Y_i = 0\right\} \\ 0 & \text{o.w.} \end{cases}$$

  where $w_{n,i} = 1/k$ if X_i is among the k ***nearest neighbors*** of x, and $w_{n,i} = 0$ elsewhere.

  X_i is said to be ***the k-th nearest neighbor*** of x if the distance $d(x, X_i)$ is the *k-th smallest* among $d(x, X_1)$, , $\ldots, d(x, X_n)$ In case of a *distance tie*, the candidate with the smaller index is said to be closer to x. The decision is based upon a ***majority vote***. It is convenient to let k be *odd*, to avoid voting ties.

- **Remark** (***Voronoi Partition***)
  At every point the decision is the label ofthe *closest* data point. *The set of points whose nearest neighbor is X_i* is called ***the Voronoi cell*** *of X_i*. The partition induced by *the Voronoi cells* is a ***Voronoi partition***.

- **Remark** (***Ordered Statistic***)
  We fix $x \in \mathbb{R}^d$, and ***reorder*** the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ according to ***increasing values*** of $d(x, X_i)$. The *reordered data sequence* is denoted by

$$\left(X_{(1)}(x), Y_{(1)}(x)\right), \ldots, \left(X_{(n)}(x), Y_{(n)}(x)\right)$$

  where $X_{(k)}(x)$ is the k-th nearest neighbor of x. For short, we write it as $(X_{(k)}, Y_{(k)})$.

2 Consistency

2.1 Asymptotic Consistency

- **Definition** Denote *the probability measure* for X by $\mathcal{P}_X$ and let $B_{x,\epsilon}$ be the ***closed ball*** centered at x of radius $\epsilon > 0$. The collection of all x with $\mathcal{P}_X(B_{x,\epsilon}) > 0$ *for all* $\epsilon > 0$ is called ***the support*** *of* X or $\mathcal{P}_X$.

- **Lemma 2.1** *[Devroye et al., 2013]*
If $x \in support(\mathcal{P}_X)$ and $\lim_{n\to\infty} k/n = 0$, then
$$d(x, X_{(k)}(x)) \to 0, \quad a.s.$$
If X is independent of the data and has probability measure $\mathcal{P}_X$, then
$$d(X, X_{(k)}(x)) \to 0, \quad a.s.$$
whenever $k/n \to 0$.

2.2 Stone's Lemma

- **Lemma 2.2** *(**Stone's Lemma**) [Devroye et al., 2013]*
For any integrable function f, any n, and any $k \le n$:
$$\sum_{i=1}^{k} \mathbb{E}\left[\left|f\left(X_{(i)}(X)\right)\right|\right] \le k\gamma_d \mathbb{E}\left[|f(X)|\right], \tag{1}$$
*where $\gamma_d \le \left(1 + 2/\sqrt{2-\sqrt{3}}\right)^d - 1$ depends upon the **dimension** only.*

- **Lemma 2.3** *(**Approximation with K-NN**) [Devroye et al., 2013]*
For any integrable function f,
$$\frac{1}{k}\sum_{i=1}^{k} \mathbb{E}\left[\left|f(X) - f\left(X_{(i)}(X)\right)\right|\right] \to 0$$
as $n \to \infty$ whenever $k/n \to 0$.

2.3 The Asymptotic Probability of Error

2.4 Stone's Theorem

- **Remark** (***Estimate Posterior Conditional Probability with Weighted Averages***)
Consider a rule based on an estimate of the a ***posteriori probability*** η of the form
$$\eta_n(x) = \sum_{i=1}^{n} \mathbb{1}\left\{Y_i = 1\right\} W_{n,i}(x) = \sum_{i=1}^{n} Y_i\, W_{n,i}(x)$$

where the weights $W_{n,i}(x) = W_{n,i}(x, X_1, \ldots, X_n)$ are nonnegative and sum to one:

$$\sum_{i=1}^{n} W_{n,i}(x) = 1.$$

η_n is a ***weighted average estimator*** of η.

The ***classification rule*** is defined as

$$\begin{aligned} g_n(x) &= \begin{cases} 0 & \sum_{i=1}^n \mathbb{1}\{Y_i = 1\} W_{n,i}(x) \le \sum_{i=1}^n \mathbb{1}\{Y_i = 0\} W_{n,i}(x) \\ 1 & \text{o.w.} \end{cases} \\ &= \begin{cases} 0 & \sum_{i=1}^n Y_i W_{n,i}(x) \le \frac{1}{2} \\ 1 & \text{o.w.} \end{cases} \end{aligned}$$

- **Remark** It is intuitively clear that pairs (X_i, Y_i) such that X_i is *close* to x should provide *more information* about $\eta(x)$ than those far from x. Thus, the weights are typically *much larger in the neighborhood of X*, so η_n is roughly *a **(weighted) relative frequency** of the X_i's* that have label 1 among points in the neighborhood of X. Thus, η_n might be viewed as a ***local average estimator***, and g_n a ***local (weighted) majority vote***.

- **Theorem 2.4** ***(Stone's Theorem, Universal Consistency of Local Average Estimator)*** *[Devroye et al., 2013]*
*Assume that for **any distribution** of X, the **weights** satisfy the following **three conditions**:*

  1. *There is a constant c such that, for every **nonnegative** measurable function f satisfying $\mathbb{E}[f(X)] < \infty$,*

  $$\mathbb{E}\left[\sum_{i=1}^n W_{n,i}(X) f(X_i)\right] \le c\mathbb{E}[f(X)].$$

  2. *For all $a > 0$,*

  $$\lim_{n\to\infty} \mathbb{E}\left[\sum_{i=1}^n W_{n,i}(X) \mathbb{1}\{d(X, X_i) > a\}\right] = 0$$

  3.

  $$\lim_{n\to\infty} \mathbb{E}\left[\max_{1\le i\le n} W_{n,i}(X)\right] = 0.$$

  *Then g_n is **universally consistent**.*

- **Remark** 1. Condition (1) is technical.

  2. Condition (2) requires that ***the overall weight*** of X_i's ***outside*** *of any* ***ball*** *of a fixed radius* ***centered at*** X must go to zero. In other words, *only points in a **shrinking neighborhood** of X should be taken into account in the **averaging**.*

  3. Condition (3) requires that ***no single*** X_i *has **too large** a contribution to the estimate.* Hence, *the **number of points** encountered in the **averaging** must tend to **infinity**.*

References

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.