Lecture 1: k-Nearest Neigbhor Rules

Tianpei Xie

Dec. 19th., 2022

Contents

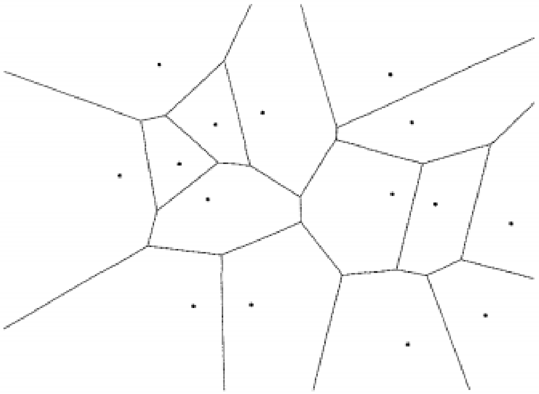

Figure 1: Varona partition of K-NN rules [Devroye et al., 2013].

1 Nearest Neighbor Rules

1.1 The Classification Rule

- **Remark** (***Memorization of Training Set and Learning by Similarity Search***)
  Nearest Neighbor algorithms are among the simplest of all machine learning algorithms. *The* ***idea*** is to ***memorize*** *the training set* and then to ***predict*** the *label* of any new instance on the basis of the labels of *its* ***closest neighbors*** *in the training set.*

  The ***rationale*** behind such a method is based on the assumption that the *features that are used to describe the domain points are relevant to their labelings in a way that makes* ***close-by points likely to have the same label***. Furthermore, in some situations, even when the training set is immense, *finding a nearest neighbor can be done extremely fast* (for example, when the training set is the entire Web and distances are based on links).

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

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

  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

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

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

- **Remark** (***Efficient Learning Without Hypothesis Class***) [Shalev-Shwartz and Ben-David, 2014]
  Note that, in contrast with the algorithmic paradigms that we have discussed so far, like *ERM*, *SRM*, *MDL*, or *RLM*, that are determined by some hypothesis class, $\mathcal{H}$, *the Nearest Neighbor* method figures out a label *on any test point* ***without searching for a predictor within some predefined class of functions***.

2 Asymptotic Analysis

2.1 Consistency of k-Nearest Neighbor Statistics

- **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]*
  Let $x \in support(\mathcal{P}_X)$ and let $X_{(k)}(x)$ be the k-th nearest neigbhor of x among n i.i.d. samples $\mathcal{D}_n = \{X_i\}_{i=1}^n$ drawn according to $\mathcal{P}_X^n$. If $n \to \infty$ 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 $\mathcal{D}_n$ 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 and Function of k-Nearest Neighbor

- **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 Stone's Theorem and Universal Consistency of k-NN Rules

- **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

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

- **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 $\mathbb{E}\left[f(X)\right] < \infty$,*

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

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

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

  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****.*

3 Non-Asymptotic Analysis

3.1 A Generalization Bound for the k-NN Rule

- **Lemma 3.1** ***(Lipschitz Bayes Classifier Case)*** *[Shalev-Shwartz and Ben-David, 2014]* *Let* $\mathcal{X} = [0,1]^d$*,* $\mathcal{Y} = \{0,1\}$*, and* $\mathcal{P}$ *be a distribution over* $\mathcal{X} \times \mathcal{Y}$ *for which* ***the conditional probability function****,* η*, is a* ***c-Lipschitz function****. Let* $\mathcal{D}_m = \{(X_1, Y_1), \ldots, (X_m, Y_m)\}$ *be an i.i.d. sample and let* g_m *be its corresponding* 1*-NN hypothesis. Let* g^* *be the* ***Bayes optimal rule*** *for* η*. Then,*

$$\mathbb{E}_{\mathcal{D}_m}\left[L(g_m)\right] \le 2L(g^*) + c\,\mathbb{E}_{X,\mathcal{D}_m}\left[\left\|X - X_{(1)}(X)\right\|\right]$$

- **Lemma 3.2** ***(Nearest Neighbor Distance Bound)*** *[Shalev-Shwartz and Ben-David, 2014]* *Let* $C_1, \ldots, C_r$ *be a collection of subsets of some domain set,* $\mathcal{X}$ *. Let* $\mathcal{D}$ *be a sequence of* m *points sampled i.i.d. according to some probability distribution,* $\mathcal{P}$ *over* $\mathcal{X}$*. Then,*

$$\mathbb{E}_{\mathcal{D}_m}\left[\sum_{i: C_i \cap \mathcal{D}_m = \emptyset} \mathcal{P}\left\{C_i\right\}\right] \le \frac{r}{e\,m}$$

- **Proposition 3.3** ***(Generalization Bounds for 1-NN Rule)*** *[Shalev-Shwartz and Ben-David, 2014]* *Let* $\mathcal{X} = [0,1]^d$*,* $\mathcal{Y} = \{0,1\}$*, and* $\mathcal{P}$ *be a distribution over* $\mathcal{X} \times \mathcal{Y}$ *for which* ***the conditional probability function****,* η*, is a* ***c-Lipschitz function****. Let* g_m *denote the result of applying the* 1*-NN rule to a sample* $\mathcal{D}_m \sim \mathcal{P}^m$*. Then,*

$$\mathbb{E}_{\mathcal{D}_m}\left[L(g_m)\right] \le 2L(g^*) + 4\,c\,\sqrt{d}\,m^{-\frac{1}{d+1}}$$

- **Proposition 3.4** ***(Generalization Bounds for*** k***-NN Rule)*** *[Shalev-Shwartz and Ben-David, 2014]* *Let* $\mathcal{X} = [0,1]^d$*,* $\mathcal{Y} = \{0,1\}$*, and* $\mathcal{P}$ *be a distribution over* $\mathcal{X} \times \mathcal{Y}$ *for which* ***the conditional probability function****,* η*, is a* ***c-Lipschitz function****. Let* g_m *denote the result of applying the* k*-NN rule to a sample* $\mathcal{D}_m \sim \mathcal{P}^m$ *where* $k \ge 10$*. Let* g^* *be the* ***Bayes optimal rule*** *for* η*.*

$$\mathbb{E}_{\mathcal{D}_m}\left[L(g_m)\right] \le \left(1 + \sqrt{\frac{8}{k}}\right) L(g^*) + (6\,c\,\sqrt{d} + k)\,m^{-\frac{1}{d+1}}.$$

- **Remark** The theorem implies that if we first fix the data-generating distribution and then *let m go to infinity*, then the error of the 1-NN rule converges to *twice* the *Bayes error.* The analysis can be generalized to larger values of k, showing that the expected error of the k-NN rule converges to $\left(1+\sqrt{8/k}\right)$ times the error of the Bayes classifier. So when $m \to \infty$ and $k \to \infty$ with $k/m \to 0$, we have universal consistency result.

3.2 The "Curse of Dimensionality"

- **Remark** (***Sample Complexity Exponentially Growth with Dimensionality***)
  The upper bound given above grows with c (the Lipschitz coefficient of η) and with d, *the Euclidean dimension of the domain set $\mathcal{X}$.* In fact, it is easy to see that a necessary condition for the last term to be smaller than ϵ is that
  $$m \geq \left(\frac{4\,c\sqrt{d}}{\epsilon}\right)^{d+1}.$$
  That is, *the* ***size*** *of the training set* ***should increase exponentially with the dimension***.

- **Proposition 3.5** *[Shalev-Shwartz and Ben-David, 2014]*
  For any $c > 1$, and every learning rule, L, there exists a distribution over $[0,1]^d \times \{0,1\}$, such that $\eta(x)$ is c-Lipschitz, the Bayes error of the distribution is 0, but for sample sizes $m \leq (c+1)^d/2$, the true error of the rule L is greater than $1/4$.

- **Remark** (***The Curse of Dimensionality***)
  The ***exponential dependence on the dimension*** is known as ***the curse of dimensionality***.

  As we saw, the 1-NN rule might fail if the number of examples is smaller than $\Omega((c+1)^d)$. Therefore, while the 1-NN rule does not restrict itself to a predefined set of hypotheses, *it still relies on some* ***prior knowledge*** since its success depends on *the* ***assumption*** *that the* ***dimension*** *and* ***the Lipschitz constant*** *of the underlying distribution, η, are* ***not too high***.

References

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

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.