6.2 Nearest Neighbor Classifiers

The nearest neighbor (NN) rules were first proposed by Fix and Hodges (1951). Despite the simplicity of the NN rules, certain versions of the NN rules are consistent with the optimal decision rule and they have shown to be competitive in a wide range of applications. Given the training data $\mathcal{D}_n$, the k-NN rule assigns a new point x to the majority class of its k-nearest neighbors. Formally it is defined as follows.

Definition 7 (k-NN rule)**.** For a fixed $x \in \mathcal{X} = \mathbb{R}^d$, order (x_i, y_i) according to $||x_i - x||$ (the *Euclidean* distance). Let $X_{(k)}(x)$ be the kth nearest neighbor of x, and define $N_k(x) = \{X_{(1)}(x), \ldots, X_{(k)}(x)\}$ to be the set of the k nearest neighbors and the weight assigned to the ith observation to be

$$W_{ni} = \begin{cases} \frac{1}{k} & \text{if } x_i \in N_k(x) \\ 0 & \text{otherwise.} \end{cases}$$

The k-NN rule is defined as

$$f_n(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^n W_{ni}\, I(y_i = 1) > \sum_{i=1}^n W_{ni}\, I(y_i = 0) \\ 0 & \text{otherwise.} \end{cases} \tag{6.1}$$

Let $R^{NN}(n) := E_{\mathcal{D}_n} R(f_n) = P_{(X,Y),\mathcal{D}_n}(f_n(X; \mathcal{D}_n) \neq Y)$ denote the unconditional probability of error of the NN classifier when the sample size is n. $R^{NN}(\infty) := \lim_{n\to\infty} R^{NN}(n)$ indicates its limit as n goes to infinity. Some well-known properties of the NN classifiers are mentioned first. For details, see 'Nearest neighbor pattern classification' by Cover and Hart (1967).

Lemma 6 (Convergence of the kth nearest neighbor)**.** *Let X and $X_1, \ldots, X_n$ be iid random variables in $\mathbb{R}^d$. Then for k_n such that $k_n/n \to 0$, as $n \to \infty$, we have*

$$\|X_{(k_n)}(X) - X\| \to 0 \quad \textit{with probability } 1.$$

Theorem 17 (Cover and Hart (1967))**.** *Under the assumption that $\eta(x) = P(Y = 1|x)$ is a continuous function of x with probability 1, we have*

$$R^* \leq R^{NN}(\infty) \leq 2\, R^*(1 - R^*).$$

Remark 11. The theorem above implies that $R^{NN}(\infty) \leq 2\, R^*$. That is, the error rate of the NN rule is at most twice the Bayes error rate in the limit. Also, if $R^* = 0$ or $1/2$, then $R^{NN}(\infty) = R^*$.

Proof. Note that $P_{(X,Y),\mathcal{D}_n}(f_n(X) \neq Y) = E_{X,X_{(1)}}\{P(Y \neq f_n(X)|X, X_{(1)}(X)\}$. Denote $P(Y \neq f_n(X)|X, X_{(1)}(X))$ by $r(X, X_{(1)})$. Then

$$\begin{aligned}
&r(X, X_{(1)}(X)) \\
&= P(Y \neq f_n(X)|X, X_{(1)}(X)) \\
&= P(Y = 1 \text{ and } Y_{(1)}(X) = 0|X, X_{(1)}(X)) + P(Y = 0 \text{ and } Y_{(1)}(X) = 1|X, X_{(1)}(X)) \\
&= P(Y_{(1)} = 0|X, X_{(1)})\, P(Y = 1|X, X_{(1)}) + P(Y_{(1)} = 1|X, X_{(1)})\, P(Y = 0|X, X_{(1)}) \\
&= P(Y_{(1)} = 0|X_{(1)})\, P(Y = 1|X) + P(Y_{(1)} = 1|X_{(1)})\, P(Y = 0|X) \\
&= (1 - \eta(X_{(1)})\, \eta(X) + (1 - \eta(X))\, \eta(X_{(1)}) \\
&\to 2\, \eta(X)(1 - \eta(X))
\end{aligned}$$

with probability 1 as $n \to \infty$ by the continuity of $\eta(.)$. Let $r(X) := 2\,\eta(X)\,(1-\eta(X))$ and $r^*(X) := \min\{\eta(X), 1-\eta(X)\}$. Note that $r(X) = 2\,r^*(X)(1-r^*(X))$ by the symmetry. Therefore

$$\begin{aligned} R^{NN}(\infty) &= \lim_{n\to\infty} E\{r(X, X_{(1)})\} = E\{\lim_{n\to\infty} r(X, X_{(1)})\} \\ &= E\{r(X)\} = 2\,E\{r^*(X)(1-r^*(X))\} = 2\{E(r^*(X)) - E(r^*(X))^2\} \\ &= 2\{R^* - R^{*2} - \mathrm{Var}(r^*(X))\} \\ &\le 2\,(R^* - R^{*2}) = 2\,R^*(1-R^*). \end{aligned}$$

The equality holds if and only if $\mathrm{Var}(r^*(X)) = 0$, i.e. $r^*(X) = E\{r^*(X)\}(= R^*)$ w.p.1. For the lower bound, which apparently holds, observe that

$$\begin{aligned} R^{NN}(\infty) &= E\{2\,r^*(X)(1-r^*(X))\} = E\{r^*(X)\} + E\{r^*(X)(1-2\,r^*(X))\} \\ &= R^* + E\{r^*(X)(1-2\,r^*(X))\} \ge R^* \end{aligned}$$

since $r^*(X) \le 1/2$ and so $E\{r^*(X)(1-2\,r^*(X))\} \ge 0$. The equality holds if and only if $r^*(X)(1-2\,r^*(X)) = 0$ almost everywhere, i.e. $r^*(X) \in \{0, 1/2\}$ w.p.1. □

More generally, let $R^{kNN}(\infty)$ denote the probability of error of the k-NN classifier for an infinite sample.

Theorem 18 (Cover and Hart for k-NN)**.** *Under the assumption that $\eta(x) = P(Y=1|x)$ is continuous in x w.p.1,*

$$R^* \le \cdots \le R^{(k+1)NN}(\infty) \le R^{kNN}(\infty) \le \cdots \le R^{NN}(\infty) \le 2\,R^*(1-R^*).$$

Proof. For simplicity, this theorem will be proved for odd k only, $R^{(2k'-1)NN}(\infty) \le R^{(2k'+1)NN}(\infty)$. For even k, the proof should be similar but some technical arguments concerning ties will be required. So, let k in the k-NN rule be odd and fixed. Let $X_{(1:k)}(X) := (X_{(1)}(X), \ldots, X_{(k)}(X))$ denote the k nearest neighbors of X.

$$\begin{aligned} &P(Y \ne f_n(X) \big| X, X_{(1:k)}(X)) \\ &= P(f_n(X)=1, Y=0 \big| X, X_{(1:k)}(X)) + P(f_n(X)=1, Y=1 \big| X, X_{(1:k)}(X)) \\ &= P\Big(\sum_{i=1}^{k} Y_{(j)}(X) > k/2, Y=0 \Big| X, X_{(1:k)}(X)\Big) \\ &\qquad\qquad + P\Big(\sum_{i=1}^{k} Y_{(j)}(X) < k/2, Y=1 \Big| X, X_{(1:k)}(X)\Big) \\ &= P(Y=0\big|X) P\Big(\sum_{i=1}^{k} Y_{(j)}(X) > k/2 \big| X_{(1:k)}(X)\Big) \\ &\qquad\qquad + P(Y=1\big|X) P\Big(\sum_{i=1}^{k} Y_{(j)}(X) < k/2 \big| X_{(1:k)}(X)\Big) \\ &\to (1-\eta(X)) P\Big(B(k,\eta(X)) \ge \frac{k+1}{2}\Big|X\Big) + \eta(X) P\Big(B(k,\eta(X)) \le \frac{k-1}{2}\Big|X\Big) \text{ w.p.1.} \end{aligned}$$

Define $r_k(X)$ to be the limit

$$(1-\eta(X))P\Big(B(k,\eta(X)) \geq \frac{k+1}{2}\Big|X\Big) + \eta(X)P\Big(B(k,\eta(X)) \leq \frac{k-1}{2}\Big|X\Big).$$

Since $r_k(x)$ is symmetric in $\eta(x)$ and $1-\eta(x)$, it can be rewritten in terms of $r^*(x) = \min\{\eta(x), 1-\eta(x)\}$.

$$\begin{aligned} r_k(x) &= r^*(x)P\Big(B(k,r^*(x)) \leq \frac{k-1}{2}\Big|x\Big) + (1-r^*(x))P\Big(B(k,r^*(x)) \geq \frac{k+1}{2}\Big|x\Big) \\ &= r^*(x) + (1-2r^*(x))P\Big(B(k,r^*(x)) \geq \frac{k+1}{2}\Big|x\Big). \end{aligned} \tag{6.2}$$

Since $1-2r^*(x) \geq 0$ and $P(B(k,r^*(X)) \geq \frac{k+1}{2}|x)$ is a monotonically decreasing function in k, $r_k(x)$ is monotonically decreasing in k as well. Again, by the dominated convergence theorem, $R^{(k+2)NN}(\infty) \leq R^{kNN}(\infty)$. The lower bound is obtained from

$$\begin{aligned} R^{kNN}(\infty) &= \lim_{n\to\infty} EP(f_n(X) \neq Y\big|X, X_{(1:k)}(X)) = E(r_k(X)) \\ &= R^* + E\left((1-2r^*(X))P\Big(B(k,r^*(X)) \geq \frac{k+1}{2}\Big|X\Big)\right) \\ &\geq R^*. \end{aligned}$$

□

Before we state the major theorem on consistency (Stone, 1977), let us consider an empirical estimate of $\eta(X) = P(Y=1|X)$ given as a weighted average of Y_i, $i = 1, \ldots, n$:

$$\eta_n(X) = \sum_{i=1}^{n} Y_i\, W_{ni}(X) = \sum_{i=1}^{n} I(Y_i = 1)\, W_{ni}(X),$$

where $W_{ni}(X) = W_{ni}(X, X_1, \ldots, X_n)$ are non-negative weights and $\sum_{i=1}^{n} W_{ni}(X) = 1$. $\eta_n(x)$ is roughly a weighted relative frequency of data points with class label 1 among the points in the neighborhood of x. This estimate of η_n defines a plug-in decision rule

$$\begin{aligned} f_n(X) &= I(\eta_n(X) > 1/2) \\ &= I\left(\sum I(Y_i = 1)W_{ni}(X) > 1/2\right) \\ &= I\left(\sum (2I(Y_i = 1) - 1)W_{ni}(X) > 0\right) \\ &= I\left(\sum (I(Y_i = 1) - I(Y_i = 0))W_{ni}(X) > 0\right) \\ &= I\left(\sum I(Y_i = 1)W_{ni}(X) > \sum I(Y_i = 0)W_{ni}(X)\right). \end{aligned} \tag{6.3}$$

Thus the k-NN classifier can be viewed as a plug-in decision rule with a local average estimator of $\eta(x)$.