CMSC 35900 (Spring 2009) Large Scale Learning **Lecture: 5**

Nearest Neighbor Rules

Instructors: Sham Kakade and Greg Shakhnarovich

In this lecture we will go over some basic asymptotic properties of the nearest neighbor rules for classification and regression.

1 Nearest neighbor properties

We consider a set of labeled data points $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ drawn i.i.d. from a joint distribution $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y}|\mathbf{x})$ over $\mathcal{X} \times \mathcal{Y}$. We will for now assume that $\mathcal{X} = \mathbb{R}^d$, and that $\mathcal{Y} = \{1, \ldots, M\}$ which means simply M-class classification.

We will denote by $\mathbf{x}_{(i)}$ the i-th nearest neighbor (i-NN) of $\mathbf{x}$ among $\mathbf{x}_1, \ldots, \mathbf{x}_n$, and by $y_{(i)}$ the label of that i-NN.

A basic property of the NN is its convergence to the "query" point (the point in neighbors of which we are interested), as $n \to \infty$. We start with the result from [3]:

Lemma 1.1. *(convergence of the nearest neighbor) Let $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_n$ be an i.i.d. sequence of random variables in $\mathbb{R}^d$.*[1] *Then, $\mathbf{x}_{(1)} \to \mathbf{x}_0$ with probability 1.*

Proof. Let $B_r(\mathbf{x})$ be the (closed) ball of radius r centered at $\mathbf{x}$

$$B_r(\mathbf{x}) \triangleq \{\mathbf{z} \in \mathbb{R}^d : D(\mathbf{z}, \mathbf{x}) \leq r\},$$

for some metric D defined on $\mathbf{R}^d$. We will first consider a point $\mathbf{x}$ such that for any $r > 0$,

$$P(B_r(\mathbf{x})) \triangleq \Pr[\mathbf{z} \in B_r(\mathbf{x})] = \int_{B_r(\mathbf{x})} p(\mathbf{z})d\mathbf{z} > 0. \tag{1}$$

Then, for any $\delta > 0$, we have

$$\Pr\left[\min_{i=1,\ldots,n}\{D(\mathbf{x}_i, \mathbf{x})\} \geq \delta\right] = [1 - P(B_r(\mathbf{x}))]^n \to 0.$$

What about points that do not satisfy (1)? Let $\bar{X}$ denote the set of all such points. Consider a point $\bar{\mathbf{x}} \in \bar{X}$, that is, for some $\bar{r}$, we have $P(B_{\bar{r}}(\bar{\mathbf{x}})) = 0$. There exists (by the denseness of rationals in $\mathbb{R}$) a rational point $\mathbf{a}_{\bar{\mathbf{x}}}$ such that $\mathbf{a}_{\bar{\mathbf{x}}} \in B_{\bar{r}/3}(\bar{\mathbf{x}})$. Consequently, there exists a small sphere $B_{\bar{r}/2}(\mathbf{a}_{\bar{\mathbf{x}}})$ such that

$$\begin{aligned} B_{\bar{r}/2}(\mathbf{a}_{\bar{\mathbf{x}}}) &\subset B_{\bar{r}}(\bar{\mathbf{x}}) \\ &\Rightarrow P(B_{\bar{r}/2}(\mathbf{a}_{\bar{\mathbf{x}}})) = 0. \end{aligned} \tag{2}$$

[1] In the original paper, the results is proven more generally for $\mathcal{X}$ being a separable metric space

Also, $\bar{\mathbf{x}} \in B_{\bar{r}/2}(\mathbf{a}_{\bar{\mathbf{x}}})$. Since $\mathbf{a}_{\bar{\mathbf{x}}}$ is rational, there is at most a countable set of such spheres that contain the entire $\bar{X}$; therefore,

$$\bar{X} \subseteq \bigcup_{\bar{\mathbf{x}} \in \bar{X}} B_{\bar{r}/2}(\mathbf{a}_{\bar{\mathbf{x}}}),$$

and from (2) this means $P(\bar{X}) = 0$.

To summarize: we show that for a random choice of $\mathbf{x}_0$, the NN $\mathbf{x}_{(1)}$ converges to $\mathbf{x}_0$ with probability 1. □

A more general result can be obtained using the following lemma:

Lemma 1.2. *(Stone) For any integrable function f, any n, and any $k \leq n$,*

$$\sum_{i=1}^{k} E_{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_n \sim p}\left[\left|f(\mathbf{x}_{(i)})\right|\right] \leq k\gamma_d E_{\mathbf{x}_0}\left[|f(\mathbf{x}_0)|\right], \tag{3}$$

where the constant $\gamma_d \leq \left(1 + 2/\sqrt{3 - \sqrt{3}}\right)^d$ only depends on the dimension d.

The proof (rather technical) can be found in [4]. We can apply the lemma as follows:

Lemma 1.3. *For any integrable function f,*

$$\frac{1}{k}\sum_{i=1}^{k} E_{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_n \sim p}\left[\left|f(\mathbf{x}_0) - f(\mathbf{x}_{(i)}\right|\right] \to 0, \tag{4}$$

as $n \to \infty$, as long as $k/n \to 0$.

That is, asymptotically, the nearest neigbor of $\mathbf{x}_0$ will have the same properties as $\mathbf{x}_0$.

2 NN classification

In the classification setup, we have a function $C : \mathcal{X} \to \mathcal{Y}$. We will for now assume that the objective in predicting the label $C(\mathbf{x}) = \hat{y}$ is to minimize the 0-1 classification loss

$$L(\hat{y}, y_0) = \begin{cases} 0 & \text{if } \hat{y} = y_0, \\ 1 & \text{if } \hat{y} \neq y_0. \end{cases}.$$

The conditional risk of a classifier based on the given n training examples is defined as $R^n(\mathbf{x}_0) \triangleq E_{y_0}[L(\hat{y}, y_0)]$. The expected risk with the n-sample training set is then

$$R^n \triangleq E_{\mathbf{x}_0}[R(\mathbf{x}_0, n)]$$

In the case of NN classification, the classifier finds the NN of the input $\mathbf{x}_0$ and outputs the label of that NN.

Before we discuss the classification guarantees of a NN classifier, we review the optimality baseline given by the Bayes error. According to a well known result in decision theory, the optimal decision strategy is to predict, for any $\mathbf{x}$, the label $y^*(\mathbf{x})$ such that

$$y^*(\mathbf{x}) = \text{argmax}_{c \in \mathcal{Y}} p(c|\mathbf{x})$$

that is, the risk R^* attained by such a classifier is minimial over all classifiers.

We first deal with the simplest case, in which $M = 2$ (binary classification). We will denote $\eta(\mathbf{x}) \triangleq p(y = 1|\mathbf{x})$. Note that in this case, the conditional Bayes risk for a given $\mathbf{x}_0$ is given by

$$R^*(\mathbf{x}_0) \;=\; \min\{\eta(\mathbf{x}_0),\, 1 - \eta(\mathbf{x}_0)\}. \tag{5}$$

Let us consider the following scenario: first draw $\mathbf{x}_0$ and $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from $p(\mathbf{x})$, then draw the labels $y_0, y_1, \ldots, y_n$ from η. Now, we look at the conditional risk

$$\begin{aligned} r(\mathbf{x}_0, \mathbf{x}_{(1)}) \;&=\; E_{y_0, y_{(1)}}\left[L(y_0, y_{(1)})\right] \\ &=\; \Pr\left[y_0 \neq y_{(1)} \mid \mathbf{x}_0, \mathbf{x}_{(1)}\right] \\ &=\; \Pr\left[y_0 = 1,\, y_{(1)} = 0 \mid \mathbf{x}_0, \mathbf{x}_{(1)}\right] \\ &\qquad +\; \Pr\left[y_0 = 0,\, y_{(1)} = 1 \mid \mathbf{x}_0, \mathbf{x}_{(1)}\right] \\ &=\; \Pr\left[y_0 = 1 \mid \mathbf{x}_0\right] \Pr\left[y_{(1)} = 0 \mid \mathbf{x}_{(1)}\right] \\ &\qquad +\; \Pr\left[y_0 = 0 \mid \mathbf{x}_0\right] \Pr\left[y_{(1)} = 1 \mid \mathbf{x}_{(1)}\right] \end{aligned} \tag{6}$$

(we decompose the joint probability to a product using the conditional independence of the labels). Therefore,

$$r(\mathbf{x}_0, \mathbf{x}_{(1)}) \;=\; \eta(\mathbf{x}_0)(1 - \eta(\mathbf{x}_{(1)})) + (1 - \eta(\mathbf{x}_0))\eta(\mathbf{x}_{(1)}) \tag{7}$$

Now we will make an assumption about the class-conditional densities $p_c(\mathbf{x}) \triangleq p(\mathbf{x}|y = c)$: we assume that with probability one, $\mathbf{x}$ is either a continuity point of p_1 and p_2, or a point with non-zero probability mass. In the latter case, suppose that the probability mass $P(\mathbf{x}_0) = \nu > 0$. Then,

$$\Pr\left[\mathbf{x}_{(1)} \neq \mathbf{x}_0\right] \;=\; (1 - \nu)^n \;\to\; 0.$$

Once the sequence of NN converges to $\mathbf{x}_0$ is stays there (having attained the lowest possible value of 0 for the distance). Therefore, from (7), we have

$$r(\mathbf{x}_0, \mathbf{x}_{(1)}) \;\to\; 2\eta(\mathbf{x}_0)(1 - \eta(\mathbf{x}_0)). \tag{8}$$

In the other case, namely,

$$\Pr\left[p_1 \text{ and } p_2 \text{ are continuous in } \mathbf{x}_0 \mid P(\mathbf{x}_0) = 0\right] \;=\; 1.$$

it follows that η is also continuous in $\mathbf{x}_0$ with probability one. Applying Lemma 1.1 we get that with probability one,

$$\eta(\mathbf{x}_{(1)}) \;\to\; \eta(\mathbf{x}_0), \tag{9}$$

and from (7), with probability one

$$r(\mathbf{x}_0, \mathbf{x}_{(1)}) \;\to\; 2\eta(\mathbf{x}_0)(1 - \eta(\mathbf{x}_0)). \tag{10}$$

Thus, under our assumptions on p_1 and p_2, with probability one

$$R(\mathbf{x}_0) \;\triangleq\; \lim_{n\to\infty} r(\mathbf{x}_0, \mathbf{x}_{(1)}) \;=\; 2\eta(\mathbf{x}_0)(1 - \eta(\mathbf{x}_0)). \tag{11}$$

Combining (5) and (11), we get

$$R(\mathbf{x}_0) \;=\; 2R^*(\mathbf{x}_0)(1 - R^*(\mathbf{x}_0)). \tag{12}$$

That is, as the size of the labeled training set goes to ininity, the probability of a randomly chosen point to be misclassified by the NN classifier approaches $2R^*(\mathbf{x}_0)(1 - R^*(\mathbf{x}_0)$ with probability one.

Now taking the expectation over $\mathbf{x}_0$ we can look at the total risk of the NN classifier

$$R \triangleq \lim_{n\to\infty} E_{\mathbf{x}_0}\left[r(\mathbf{x}_0, \mathbf{x}_{(1)})\right].$$

Applying the dominated convergence theorem, we can switch the order of the limit and expectation, and get

$$\begin{aligned} R &= E_{\mathbf{x}_0}\left[\lim_{n\to\infty} r(\mathbf{x}_0, \mathbf{x}_{(1)})\right] \\ &= E_{\mathbf{x}_0}\left[2R^*(\mathbf{x}_0)(1 - R^*(\mathbf{x}_0))\right] \\ &= 2E_{\mathbf{x}_0}\left[R^*(\mathbf{x}_0)\right] - 2E_{\mathbf{x}_0}\left[(R^*(\mathbf{x}_0))^2\right] \\ &= 2R^* - 2(R^*)^2 - 2\,\mathrm{var}\; R^*(\mathbf{x}_0) \\ &= 2R^*(1 - R^*) - 2\,\mathrm{var}\; R^*(\mathbf{x}_0) \\ &\leq 2R^*(1 - R^*). \end{aligned} \tag{13}$$

On the other hand, a similar manipulation yields

$$\begin{aligned} R &= E_{\mathbf{x}_0}\left[R^*(\mathbf{x}_0) + R^*(\mathbf{x}_0)(1 - 2R^*(\mathbf{x}_0))\right] \\ &= R^* + E_{\mathbf{x}_0}\left[R^*(\mathbf{x}_0)(1 - 2R^*(\mathbf{x}_0))\right] \\ &\geq R^*, \end{aligned} \tag{14}$$

using the fact that $R^* \leq 1/2$. Note that this inequality also follows directly from the optimality properties of Bayes classifier.

We have thus proven the famous result from [3]:

Theorem 2.1. *(Cover-Hart inequality) Let p_1, p_2 be the class-conditional probability densities over $\mathbb{R}^d$ such that with probability one, $\mathbf{x}$ is either (a) a continuity point of p_1 and p_2, or (b) a point of nonzero probability mass. Then, the asymptotic risk R of the NN classifier is bounded by*

$$R^* \leq R \leq 2R^*(1 - R^*). \tag{15}$$

These bounds are tight, in the sense that there exist distributions p_1 and p_2 for which the limits are attained exactly.

It is interesting to consider the cases in which the bounds are attained. In particular, the necessary and sufficient condition for the upper bound to hold with equality is that $\mathrm{var}\; R^*(\mathbf{x}_0) = 0$ which holds if and only if $R^*(\mathbf{x}_0) = R^*$ with probability one. That in turn happens if and only if

$$\frac{\eta(\mathbf{x}_0)}{1 - \eta(\mathbf{x}_0)} = \frac{R^*}{1 - R^*} \quad \text{or} \quad \frac{\eta(\mathbf{x}_0)}{1 - \eta(\mathbf{x}_0)} = \frac{1 - R^*}{R^*}$$

for almost every $\mathbf{x}_0$ (i.e., all $\mathbf{x}_0$ except a set with zero probability mass). Similarly, we find that the lower bound holds with equality if and only if $R^* = 0$ or $R^* = 1/2$ almost everywhere.

The case of $M > 2$ requires a little additional work. The proof is given in the Appendix, here we just state the result: under conditions similar to those in Theorem 2.1, the risk of the NN classifier for $M > 2$ classes is subject to the tight bounds

$$R^* \leq R \leq R^*\left(2 - \frac{M}{M-1}R^*\right).$$

3 NN regression

In the regression setup, the Bayes estimator is the estimator $C^* : \mathcal{X} \to \mathcal{Y}$ that minimizes the expected risk

$$R^*(\mathbf{x}_0) \;=\; E_{y_0}\left[L(y_0, C^*(\mathbf{x}_0)) \,|\, \mathbf{x}_0\right]$$

A number of results were derived in [1], under various assumptions on the properties of the loss function.

3.1 Metric loss

Theorem 3.1. *Let L be a metric loss such that for every y_0, $E_y\left[L(y, y_0) \,|\, \mathbf{x}_0\right]$ is a continuous function of $\mathbf{x}_0$ with probability one. Then,*

$$R^*(\mathbf{x}_0) \;\leq\; R(\mathbf{x}_0) \;\leq\; 2R^*(\mathbf{x}_0),$$

with probability one.

Corollary 3.2. *If L is as in Theorem 3.1 and bounded, then*

$$R^* \;\leq\; R \;\leq\; 2R^*.$$

3.2 Squared loss

Since squared loss $L(\hat{y}, y_0) = (\hat{y} - y_0)^2$ is not a metric, the results in the previous section do not apply. To derive the bounds in this case, we will need the conditional moments of the label y given $\mathbf{x}_0$:

$$\mu_1(\mathbf{x}_0) \;\triangleq\; E_{y_0}\left[y_0 \,|\, \mathbf{x}_0\right], \tag{16}$$

$$\mu_2(\mathbf{x}_0) \;\triangleq\; E_{y_0}\left[y_0^2 \,|\, \mathbf{x}_0\right], \tag{17}$$

and the conditional variance

$$\sigma^2(\mathbf{x}_0) \;\triangleq\; \mu_2(\mathbf{x}_0) - \mu_1^2(\mathbf{x}_0). \tag{18}$$

Since under the squared loss the Bayes estimator is the conditional mean μ_1, the conditional Bayes risk is given by $R^*(\mathbf{x}_0) = \sigma^2(\mathbf{x}_0)$, yielding the total risk

$$R^* \;=\; E_{\mathbf{x}_0}\left[\sigma^2(\mathbf{x}_0)\right].$$

Theorem 3.3. *If $\mu_1(\mathbf{x}_0)$ and $\mu_2(\mathbf{x}_0)$ are continuous with probability one, then*

$$R(\mathbf{x}_0) \;=\; 2R^*(\mathbf{x}_0),$$

with probability one.

The result in Theorem 3.3 can not be extended to the total risk R without some additional conditions on the behavior of y. Here is an example showing how lack of such conditions can affect the risk. Suppose $p(x)$ is a Gaussian on 1D, and let $p(y|x)$ be such that $\mu_1(x) = 1/x$ (this still makes μ_1 continuous with probability one!) and $\sigma^2(x) \equiv \sigma^2 < \infty$. Then, for any x in the limit, $R(x) = 2R^*(x)$ according to the theorem. However, for any n the total risk of the NN estimate is infinite; we can show it to be arbitrarily large by considering the combination of x and $x_{(1)}(x)$ that are very close to origin but with opposite signs.

Corollary 3.4. *Suppose* $E_{\mathbf{x},\mathbf{x}'}\left[D^2(\mathbf{x},\mathbf{x}')\right] < \infty$*, and let there exist constants* A, B *such that*

$$|\mu_1(\mathbf{x}_1) - \mu_1(\mathbf{x}_2)|^2 \;\leq\; AD^2(\mathbf{x}_1,\mathbf{x}_2)$$

and

$$|\sigma^2(\mathbf{x}_1) - \sigma^2(\mathbf{x}_2)| \;\leq\; BD^2(\mathbf{x}_1,\mathbf{x}_2)$$

for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$*. Then, under squared loss,*

$$R \;=\; 2R^*.$$

4 k nearest neighbor rules

A natural extension of the NN rule is to consider k nearest neighbors of $\mathbf{x}_0$, and use their labels to infer the unknown label y_0. We will denote the conditional and total risks of the k-NN classifier as $R_k(\mathbf{x}_0)$ and R_k, respectively.

For classification, the resulting rule is to predict the label given by the majority vote among the k neighbors. It is easy to show that $R_k \;\leq\; 2R^*(1 - R^*)$, however stronger bounds have been obtained. In particular,

Theorem 4.1. *(Devroye, 1981) For all distributions and any odd* $k \geq 3$*,*

$$R_k \;\leq\; R^*\left(1 + \frac{\gamma}{\sqrt{k}}\left(1 + O(k^{-1/6})\right)\right)$$

where $\gamma = 0.33994241...$ *is a constant, and the O notation refers to the limit as* $k \to \infty$*.*

For regression, the simple k-NN rule makes prediction according to

$$\hat{y} \;=\; \frac{1}{k}\sum_{i=1}^{k} y_{(i)}.$$

Its risk under squared loss is characterized by the following corollary to Theorem 3.3:

Corollary 4.2. *Under the assumptions of Theorem 3.3, the* k*-NN conditional risk is*

$$R_k(\mathbf{x}_0) \;=\; (1 + 1/k)\, R^*(\mathbf{x}_0)$$

and under the additional assumptions of Corollary 3.4the total risk is

$$R_k \;=\; (1 + 1/k)\, R^*,$$

all with probability one.

5 Finite sample behavior

We have established the asymptotic behavior of the NN estimator, but how fast does the finite sample risk approach the limit? Cover has shown in [2] that this convergence can be arbitrarily slow; in the paper, he uses the following example. Let $\mathcal{X}$ be the set of positive integers 1,2,3... and let the density of x be defined by $\pi_i \triangleq p(x = i) = c/i^{1+\delta}$, where $\delta > 0$ and $c(\delta)$ is set so that the density sums to 1 over $\mathcal{X}$. Next, let the labels $y_i \in \{1, 2\}$ be drawn i.i.d. from the "fair coin" Bernoulli distribution, that is, for any i, $\Pr[y_i = 1] = 1/2$. Once we have drawn y_i, we define $p(y = y_i|x = i) = 1$. Obviously, the Bayes risk for any such distribution is zero (knowing x we know y with no uncertainty). Therefore, the asymptotic NN risk R also converges to zero!

Now consider a NN rule using a finite sample of n pairs drawn from $p(x, y)$. If we are lucky and $x_{(1)} = x_0$, we don't make a mistake. But if $x_{(1)} \neq x_0$, we have only 1/2 chance of getting y_0 right, since effectively, the distribution of $y_{(1)}$ has no bearing on that of y_0. Thus, the probability of error of the NN rule is

$$R^n \;=\; \frac{1}{2}\sum_{i=1}^{\infty} \pi_i(1-\pi_i)^n \;\geq\; c(\delta)n^{-\delta/(1+\delta)}$$

By changing δ we can make the convergence of R^n to R^∞ arbitrarily slow.

Of course, this example is a bit pathological, and we can get a reasonable convergence rate guarantees if we assume the underlying density is a bit more well-behaved.

Theorem 5.1. *(Cover, 1968) Let the class-conditional densities p_1, p_2 have uniformly bounded third derivatives, and be bounded away from zero almost everywhere (i.e., with probability one over their support). Then,*

$$R^n \;=\; R^\infty + O(1/n^2).$$

Appendix

A Proof of the Cover-Hart inequality for $M > 2$

First of all, we introduce the notation

$$\eta_i(\mathbf{x}) \;\triangleq\; p(y = i|\mathbf{x})$$

and

$$\vec{\eta}(\mathbf{x}) \;\triangleq\; [p(y = 1|\mathbf{x}), \ldots, p(y = M|\mathbf{x})]^T.$$

The extension of the above result to $M > 2$ gives the following

Theorem A.1. *Let $p_1, \ldots, p_M$ be (class-conditional) probability densities over $\mathbb{R}^d$ such that with probability one, $\mathbf{x}$ is either (a) a continuity point of $p_1, \ldots, p_M$ or (b) a point of nonzero probability mass. Then, the asymptotic risk R of the NN classifier is bounded by*

$$R^* \;\leq\; R \;\leq\; R^*\left(2 - \frac{M}{M-1}R^*\right), \tag{19}$$

with the bounds being tight.

Proof. We already have seen that under the hypothesized conditions, $\mathbf{x}_{(1)}\mathbf{x}_0) \rightarrow \mathbf{x}_0$ with probability one, and therefore $\vec{\eta}(\mathbf{x}_{(1)}) \rightarrow \vec{\eta}(\mathbf{x}_0)$ with probability one. Now, the conditional NN risk is

$$\begin{aligned} r(\mathbf{x}_0, \mathbf{x}_{(1)}) &= E_{y_0, y_{(1)}}\left[L(y_0, y_{(1)}) \,|\, \mathbf{x}_0, \mathbf{x}_{(1)}\right] = \sum_{i \neq j} \eta_i(\mathbf{x}_0)\eta_j(\mathbf{x}_{(1)}) \\ &= 1 - \sum_{i=1}^{M} \eta_i(\mathbf{x}_0)\eta_i(\mathbf{x}_{(1)}) \end{aligned}$$

(summing the probabilities of all the "bad" events of label mismatch), which converges with probability one to

$$R(\mathbf{x}_0) = \lim_{n \rightarrow \infty} r(\mathbf{x}_0, \mathbf{x}_{(1)}) = 1 - \sum_{i=1}^{M} \eta_i^2(\mathbf{x}_0). \tag{20}$$

Suppose that $c = \text{argmax}_i \eta_i(\mathbf{x}_0)$; then the conditional Bayes risk is given by

$$R^*(\mathbf{x}_0) = 1 - \eta_c(\mathbf{x}_0).$$

We now use the Cauchy-Schwarz inequality (for vectors $\mathbf{1}_{M-1}$ and $[\eta_1, \ldots, \eta_{k-1}, \eta_{k+1}, \ldots, \eta_M]^T$).

$$(M-1)\sum_{i \neq c}^{M} \eta_i^2(\mathbf{x}_0) \geq \left(\sum_{i \neq c} \eta_i(\mathbf{x}_0)\right)^2 = (1 - \eta_k(\mathbf{x}_0))^2 = (R^*(\mathbf{x}_0))^2 \tag{21}$$

Adding $(M-1)\eta_k^2(\mathbf{x}_0)$ to each side, we get

$$\begin{aligned} (M-1)\sum_{i=1}^{M} \eta_i^2(\mathbf{x}_0) &\geq (R^*(\mathbf{x}_0))^2 + (M-1)\eta_k^2(\mathbf{x}_0) \\ &= (R^*(\mathbf{x}_0))^2 + (M-1)\left(1 - R^*(\mathbf{x}_0)\right)^2 \\ &= 1 - 2R^*(\mathbf{x}_0) + \frac{M}{M-1}\left(R^*(\mathbf{x}_0)\right)^2 \end{aligned} \tag{22}$$

from which we get

$$\sum_{i=1}^{M} \eta_i^2(\mathbf{x}_0) \geq \frac{(R^*(\mathbf{x}_0))^2}{M-1} + (1 - R^*(\mathbf{x}_0))^2 \tag{23}$$

Now subsituting this inequality into (20), we have

$$R(\mathbf{x}_0) \leq 2R^*(\mathbf{x}_0) - \frac{M}{M-1}\left(1 - R^*(\mathbf{x}_0)\right)^2 \tag{24}$$

Taking expectation with respect to $\mathbf{x}_0$, and using the dominated convergence theorem as before, we have

$$\begin{aligned} R &= 2R^* - \frac{M}{M-1}(R^*)^2 - \frac{M}{M-1}\,\text{var}\; R^*(\mathbf{x}_0) \\ &\leq R^*\left(2 - \frac{M}{M-1}R^*\right). \end{aligned} \tag{25}$$

□

References

[1] T. M. Cover. Estimation by the nearest neighbor rule. *IEEE Transactions on Information Theory*, 14:21–27, January 1968.

[2] T. M. Cover. Rates of Convergence for Nearest Neighbor Procedures. In *Proc. 1st Ann. Hawaii Conf. Systems Theory*, pages 413–415, January 1968.

[3] T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, January 1967.

[4] L. Devroye, L. Gyöfri, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.