19 Nearest Neighbor

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).

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.

In this chapter we describe Nearest Neighbor methods for classification and regression problems. We analyze their performance for the simple case of binary classification and discuss the efficiency of implementing these methods.

19.1 k Nearest Neighbors

Throughout the entire chapter we assume that our instance domain, $\mathcal{X}$, is endowed with a metric function ρ. That is, $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a function that returns the distance between any two elements of $\mathcal{X}$. For example, if $\mathcal{X} = \mathbb{R}^d$ then ρ can be the Euclidean distance, $\rho(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\| = \sqrt{\sum_{i=1}^{d} (x_i - x'_i)^2}$.

Let $S = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ be a sequence of training examples. For each $\mathbf{x} \in \mathcal{X}$, let $\pi_1(\mathbf{x}), \ldots, \pi_m(\mathbf{x})$ be a reordering of $\{1, \ldots, m\}$ according to their distance to $\mathbf{x}$, $\rho(\mathbf{x}, \mathbf{x}_i)$. That is, for all $i < m$,

$$\rho(\mathbf{x}, \mathbf{x}_{\pi_i(\mathbf{x})}) \leq \rho(\mathbf{x}, \mathbf{x}_{\pi_{i+1}(\mathbf{x})}).$$

For a number k, the k-NN rule for binary classification is defined as follows:

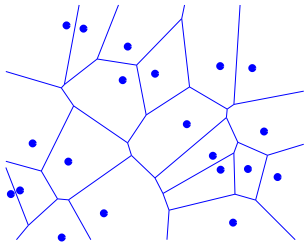


Figure 19.1 An illustration of the decision boundaries of the 1-NN rule. The points depicted are the sample points, and the predicted label of any new point will be the label of the sample point in the center of the cell it belongs to. These cells are called a Voronoi Tessellation of the space.

k-NN

input: a training sample $S = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$
output: for every point $\mathbf{x} \in \mathcal{X}$,
return the majority label among $\{y_{\pi_i(\mathbf{x})} : i \leq k\}$

When $k = 1$, we have the 1-NN rule:

$$h_S(\mathbf{x}) \;=\; y_{\pi_1(\mathbf{x})}.$$

A geometric illustration of the 1-NN rule is given in Figure 19.1.

For regression problems, namely, $\mathcal{Y} = \mathbb{R}$, one can define the prediction to be the average target of the k nearest neighbors. That is, $h_S(\mathbf{x}) = \frac{1}{k}\sum_{i=1}^{k} y_{\pi_i(\mathbf{x})}$. More generally, for some function $\phi : (\mathcal{X} \times \mathcal{Y})^k \to \mathcal{Y}$, the k-NN rule with respect to ϕ is:

$$h_S(\mathbf{x}) \;=\; \phi\left((\mathbf{x}_{\pi_1(\mathbf{x})}, y_{\pi_1(\mathbf{x})}), \ldots, (\mathbf{x}_{\pi_k(\mathbf{x})}, y_{\pi_k(\mathbf{x})})\right). \tag{19.1}$$

It is easy to verify that we can cast the prediction by majority of labels (for classification) or by the averaged target (for regression) as in Equation (19.1) by an appropriate choice of ϕ. The generality can lead to other rules; for example, if $\mathcal{Y} = \mathbb{R}$, we can take a weighted average of the targets according to the distance from $\mathbf{x}$:

$$h_S(\mathbf{x}) \;=\; \sum_{i=1}^{k} \frac{\rho(\mathbf{x}, \mathbf{x}_{\pi_i(\mathbf{x})})}{\sum_{j=1}^{k} \rho(\mathbf{x}, \mathbf{x}_{\pi_j(\mathbf{x})})}\, y_{\pi_i(\mathbf{x})}.$$

19.2 Analysis

Since the NN rules are such natural learning methods, their generalization properties have been extensively studied. Most previous results are asymptotic consistency results, analyzing the performance of NN rules when the sample size, m,

goes to infinity, and the rate of convergence depends on the underlying distribution. As we have argued in Section 7.4, this type of analysis is not satisfactory. One would like to learn from finite training samples and to understand the generalization performance as a function of the size of such finite training sets and clear prior assumptions on the data distribution. We therefore provide a finite-sample analysis of the 1-NN rule, showing how the error decreases as a function of m and how it depends on properties of the distribution. We will also explain how the analysis can be generalized to k-NN rules for arbitrary values of k. In particular, the analysis specifies the number of examples required to achieve a true error of $2L_{\mathcal{D}}(h^\star) + \epsilon$, where $h^\star$ is the Bayes optimal hypothesis, assuming that the labeling rule is "well behaved" (in a sense we will define later).

19.2.1 A Generalization Bound for the 1-NN Rule

We now analyze the true error of the 1-NN rule for binary classification with the 0-1 loss, namely, $\mathcal{Y} = \{0, 1\}$ and $\ell(h, (\mathbf{x}, y)) = \mathbb{1}_{[h(\mathbf{x}) \neq y]}$. We also assume throughout the analysis that $\mathcal{X} = [0, 1]^d$ and ρ is the Euclidean distance.

We start by introducing some notation. Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$. Let $\mathcal{D}_{\mathcal{X}}$ denote the induced marginal distribution over $\mathcal{X}$ and let $\eta : \mathbb{R}^d \to \mathbb{R}$ be the conditional probability[1] over the labels, that is,

$$\eta(\mathbf{x}) = \mathbb{P}[y = 1 | \mathbf{x}].$$

Recall that the Bayes optimal rule (that is, the hypothesis that minimizes $L_{\mathcal{D}}(h)$ over all functions) is

$$h^\star(\mathbf{x}) = \mathbb{1}_{[\eta(\mathbf{x}) > 1/2]}.$$

We assume that the conditional probability function η is c-Lipschitz for some $c > 0$: Namely, for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, $|\eta(\mathbf{x}) - \eta(\mathbf{x}')| \leq c \, \|\mathbf{x} - \mathbf{x}'\|$. In other words, this assumption means that if two vectors are close to each other then their labels are likely to be the same.

The following lemma applies the Lipschitzness of the conditional probability function to upper bound the true error of the 1-NN rule as a function of the expected distance between each test instance and its nearest neighbor in the training set.

LEMMA 19.1 *Let $\mathcal{X} = [0, 1]^d, \mathcal{Y} = \{0, 1\}$, and $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$ for which the conditional probability function, η, is a c-Lipschitz function. Let $S = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ be an i.i.d. sample and let h_S be its corresponding 1-NN hypothesis. Let $h^\star$ be the Bayes optimal rule for η. Then,*

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}^m}[L_{\mathcal{D}}(h_S)] \leq 2 \, L_{\mathcal{D}}(h^\star) + c \mathop{\mathbb{E}}_{S \sim \mathcal{D}^m, \mathbf{x} \sim \mathcal{D}}[\|\mathbf{x} - \mathbf{x}_{\pi_1(\mathbf{x})}\|].$$

[1] Formally, $\mathbb{P}[y = 1 | \mathbf{x}] = \lim_{\delta \to 0} \frac{\mathcal{D}(\{(\mathbf{x}', 1) : \mathbf{x}' \in B(\mathbf{x}, \delta)\})}{\mathcal{D}(\{(\mathbf{x}', y) : \mathbf{x}' \in B(\mathbf{x}, \delta), y \in \mathcal{Y}\})}$, where $B(\mathbf{x}, \delta)$ is a ball of radius δ centered around $\mathbf{x}$.

Proof Since $L_{\mathcal{D}}(h_S) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbb{1}_{[h_S(\mathbf{x})\neq y]}]$, we obtain that $\mathbb{E}_S[L_{\mathcal{D}}(h_S)]$ is the probability to sample a training set S and an additional example $(\mathbf{x}, y)$, such that the label of $\pi_1(\mathbf{x})$ is different from y. In other words, we can first sample m unlabeled examples, $S_x = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$, according to $\mathcal{D}_{\mathcal{X}}$, and an additional unlabeled example, $\mathbf{x} \sim \mathcal{D}_{\mathcal{X}}$, then find $\pi_1(\mathbf{x})$ to be the nearest neighbor of $\mathbf{x}$ in S_x, and finally sample $y \sim \eta(\mathbf{x})$ and $y_{\pi_1(\mathbf{x})} \sim \eta(\pi_1(\mathbf{x}))$. It follows that

$$\begin{aligned}\mathop{\mathbb{E}}_{S}[L_{\mathcal{D}}(h_S)] &= \mathop{\mathbb{E}}_{S_x\sim\mathcal{D}_{\mathcal{X}}^m,\mathbf{x}\sim\mathcal{D}_{\mathcal{X}},y\sim\eta(\mathbf{x}),y'\sim\eta(\pi_1(\mathbf{x}))}[\mathbb{1}_{[y\neq y']}] \\ &= \mathop{\mathbb{E}}_{S_x\sim\mathcal{D}_{\mathcal{X}}^m,\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}}\left[\mathop{\mathbb{P}}_{y\sim\eta(\mathbf{x}),y'\sim\eta(\pi_1(\mathbf{x}))}[y\neq y']\right]. \qquad (19.2)\end{aligned}$$

We next upper bound $\mathbb{P}_{y\sim\eta(\mathbf{x}),y'\sim\eta(\mathbf{x}')}[y \neq y']$ for any two domain points $\mathbf{x}, \mathbf{x}'$:

$$\begin{aligned}\mathop{\mathbb{P}}_{y\sim\eta(\mathbf{x}),y'\sim\eta(\mathbf{x}')}[y\neq y'] &= \eta(\mathbf{x}')(1-\eta(\mathbf{x})) + (1-\eta(\mathbf{x}'))\eta(\mathbf{x}) \\ &= (\eta(\mathbf{x}) - \eta(\mathbf{x}) + \eta(\mathbf{x}'))(1-\eta(\mathbf{x})) \\ &\quad + (1-\eta(\mathbf{x}) + \eta(\mathbf{x}) - \eta(\mathbf{x}'))\eta(\mathbf{x}) \\ &= 2\eta(\mathbf{x})(1-\eta(\mathbf{x})) + (\eta(\mathbf{x}) - \eta(\mathbf{x}'))(2\eta(\mathbf{x}) - 1).\end{aligned}$$

Using $|2\eta(\mathbf{x}) - 1| \leq 1$ and the assumption that η is c-Lipschitz, we obtain that the probability is at most:

$$\mathop{\mathbb{P}}_{y\sim\eta(\mathbf{x}),y'\sim\eta(\mathbf{x}')}[y\neq y'] \leq 2\eta(\mathbf{x})(1-\eta(\mathbf{x})) + c\,\|\mathbf{x}-\mathbf{x}'\|.$$

Plugging this into Equation (19.2) we conclude that

$$\mathop{\mathbb{E}}_{S}[L_{\mathcal{D}}(h_S)] \;\leq\; \mathop{\mathbb{E}}_{\mathbf{x}}[2\eta(\mathbf{x})(1-\eta(\mathbf{x}))] + c\mathop{\mathbb{E}}_{S,\mathbf{x}}[\|\mathbf{x}-\mathbf{x}_{\pi_1(\mathbf{x})}\|].$$

Finally, the error of the Bayes optimal classifier is

$$L_{\mathcal{D}}(h^\star) \;=\; \mathop{\mathbb{E}}_{\mathbf{x}}[\min\{\eta(\mathbf{x}), 1-\eta(\mathbf{x})\}] \geq \mathop{\mathbb{E}}_{\mathbf{x}}[\eta(\mathbf{x})(1-\eta(\mathbf{x}))].$$

Combining the preceding two inequalities concludes our proof. □

The next step is to bound the expected distance between a random $\mathbf{x}$ and its closest element in S. We first need the following general probability lemma. The lemma bounds the probability weight of subsets that are not hit by a random sample, as a function of the size of that sample.

LEMMA 19.2 *Let $C_1, \ldots, C_r$ be a collection of subsets of some domain set, $\mathcal{X}$. Let S be a sequence of m points sampled i.i.d. according to some probability distribution, $\mathcal{D}$ over $\mathcal{X}$. Then,*

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\sum_{i:C_i\cap S=\emptyset}\mathbb{P}[C_i]\right] \;\leq\; \frac{r}{m\,e}.$$

Proof From the linearity of expectation, we can rewrite:

$$\mathop{\mathbb{E}}_{S}\left[\sum_{i:C_i\cap S=\emptyset}\mathbb{P}[C_i]\right] \;=\; \sum_{i=1}^{r}\mathbb{P}[C_i]\mathop{\mathbb{E}}_{S}\left[\mathbb{1}_{[C_i\cap S=\emptyset]}\right].$$

Next, for each i we have

$$\mathop{\mathbb{E}}_{S}\left[\mathbb{1}_{[C_i\cap S=\emptyset]}\right] = \mathop{\mathbb{P}}_{S}[C_i\cap S=\emptyset] = (1-\mathbb{P}[C_i])^m \le e^{-\mathbb{P}[C_i]\,m}.$$

Combining the preceding two equations we get

$$\mathop{\mathbb{E}}_{S}\left[\sum_{i:C_i\cap S=\emptyset}\mathbb{P}[C_i]\right] \;\le\; \sum_{i=1}^{r}\mathbb{P}[C_i]\,e^{-\mathbb{P}[C_i]\,m} \;\le\; r\,\max_i \mathbb{P}[C_i]\,e^{-\mathbb{P}[C_i]\,m}.$$

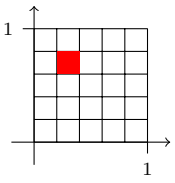
Finally, by a standard calculus, $\max_a ae^{-ma} \le \frac{1}{me}$ and this concludes the proof. □

Equipped with the preceding lemmas we are now ready to state and prove the main result of this section – an upper bound on the expected error of the 1-NN learning rule.

THEOREM 19.3 *Let $\mathcal{X}=[0,1]^d, \mathcal{Y}=\{0,1\}$, and $\mathcal{D}$ be a distribution over $\mathcal{X}\times\mathcal{Y}$ for which the conditional probability function, η, is a c-Lipschitz function. Let h_S denote the result of applying the 1-NN rule to a sample $S\sim\mathcal{D}^m$. Then,*

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(h_S)] \le 2\,L_{\mathcal{D}}(h^\star) + 4\,c\sqrt{d}\,m^{-\frac{1}{d+1}}.$$

Proof Fix some $\epsilon = 1/T$, for some integer T, let $r=T^d$ and let $C_1,\ldots,C_r$ be the cover of the set $\mathcal{X}$ using boxes of length ϵ: Namely, for every $(\alpha_1,\ldots,\alpha_d)\in[T]^d$, there exists a set C_i of the form $\{\mathbf{x}:\forall j, x_j\in[(\alpha_j-1)/T,\alpha_j/T]\}$. An illustration for $d=2, T=5$ and the set corresponding to $\alpha=(2,4)$ is given in the following.



For each $\mathbf{x},\mathbf{x}'$ in the same box we have $\|\mathbf{x}-\mathbf{x}'\|\le\sqrt{d}\,\epsilon$. Otherwise, $\|\mathbf{x}-\mathbf{x}'\|\le\sqrt{d}$. Therefore,

$$\mathop{\mathbb{E}}_{\mathbf{x},S}[\|\mathbf{x}-\mathbf{x}_{\pi_1(\mathbf{x})}\|] \le \mathop{\mathbb{E}}_{S}\left[\mathbb{P}\left[\bigcup_{i:C_i\cap S=\emptyset}C_i\right]\sqrt{d} + \mathbb{P}\left[\bigcup_{i:C_i\cap S\neq\emptyset}C_i\right]\epsilon\sqrt{d}\right],$$

and by combining Lemma 19.2 with the trivial bound $\mathbb{P}[\bigcup_{i:C_i\cap S\neq\emptyset}C_i]\le 1$ we get that

$$\mathop{\mathbb{E}}_{\mathbf{x},S}[\|\mathbf{x}-\mathbf{x}_{\pi_1(\mathbf{x})}\|] \le \sqrt{d}\left(\tfrac{r}{me}+\epsilon\right).$$

Since the number of boxes is $r = (1/\epsilon)^d$ we get that

$$\mathop{\mathbb{E}}_{S,\mathbf{x}}\left[\|\mathbf{x} - \mathbf{x}_{\pi_1(\mathbf{x})}\|\right] \leq \sqrt{d}\left(\frac{2^d\,\epsilon^{-d}}{m\,e} + \epsilon\right).$$

Combining the preceding with Lemma 19.1 we obtain that

$$\mathop{\mathbb{E}}_{S}[L_{\mathcal{D}}(h_S)] \leq 2\,L_{\mathcal{D}}(h^\star) + c\sqrt{d}\left(\frac{2^d\,\epsilon^{-d}}{m\,e} + \epsilon\right).$$

Finally, setting $\epsilon = 2\,m^{-1/(d+1)}$ and noting that

$$\begin{aligned}\frac{2^d\,\epsilon^{-d}}{m\,e} + \epsilon &= \frac{2^d\,2^{-d}\,m^{d/(d+1)}}{m\,e} + 2\,m^{-1/(d+1)} \\ &= m^{-1/(d+1)}(1/e + 2) \leq 4m^{-1/(d+1)}\end{aligned}$$

we conclude our proof. □

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 $(1 + \sqrt{8/k})$ times the error of the Bayes classifier. This is formalized in Theorem 19.5, whose proof is left as a guided exercise.

19.2.2 The "Curse of Dimensionality"

The upper bound given in Theorem 19.3 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 in Theorem 19.3 to be smaller than ϵ is that $m \geq (4\,c\sqrt{d}/\epsilon)^{d+1}$. That is, the size of the training set should increase exponentially with the dimension. The following theorem tells us that this is not just an artifact of our upper bound, but, for some distributions, this amount of examples is indeed necessary for learning with the NN rule.

THEOREM 19.4 *For any $c > 1$, and every learning rule, L, there exists a distribution over $[0,1]^d \times \{0,1\}$, such that $\eta(\mathbf{x})$ 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 $1/4$.*

Proof Fix any values of c and d. Let G_c^d be the grid on $[0,1]^d$ with distance of $1/c$ between points on the grid. That is, each point on the grid is of the form $(a_1/c, \ldots, a_d/c)$ where a_i is in $\{0, \ldots, c-1, c\}$. Note that, since any two distinct points on this grid are at least $1/c$ apart, any function $\eta : G_C^D \to [0,1]$ is a c-Lipschitz function. It follows that the set of all c-Lipschitz functions over G_c^d contains the set of *all* binary valued functions over that domain. We can therefore invoke the No-Free-Lunch result (Theorem 5.1) to obtain a lower bound on the needed sample sizes for learning that class. The number of points on the grid is $(c+1)^d$; hence, if $m < (c+1)^d/2$, Theorem 5.1 implies the lower bound we are after. □

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 – its success depends on the assumption that the dimension and the Lipschitz constant of the underlying distribution, η, are not too high.

19.3 Efficient Implementation*

Nearest Neighbor is a learning-by-memorization type of rule. It requires the entire training data set to be stored, and at test time, we need to scan the entire data set in order to find the neighbors. The time of applying the NN rule is therefore $\Theta(d\, m)$. This leads to expensive computation at test time.

When d is small, several results from the field of computational geometry have proposed data structures that enable to apply the NN rule in time $o(d^{O(1)} \log(m))$. However, the space required by these data structures is roughly $m^{O(d)}$, which makes these methods impractical for larger values of d.

To overcome this problem, it was suggested to improve the search method by allowing an *approximate* search. Formally, an r-approximate search procedure is guaranteed to retrieve a point within distance of at most r times the distance to the nearest neighbor. Three popular approximate algorithms for NN are the kd-tree, balltrees, and locality-sensitive hashing (LSH). We refer the reader, for example, to (Shakhnarovich, Darrell & Indyk 2006).

19.4 Summary

The k-NN rule is a very simple learning algorithm that relies on the assumption that "things that look alike must be alike." We formalized this intuition using the Lipschitzness of the conditional probability. We have shown that with a sufficiently large training set, the risk of the 1-NN is upper bounded by twice the risk of the Bayes optimal rule. We have also derived a lower bound that shows the "curse of dimensionality" – the required sample size might increase exponentially with the dimension. As a result, NN is usually performed in practice after a dimensionality reduction preprocessing step. We discuss dimensionality reduction techniques later on in Chapter 23.

19.5 Bibliographic Remarks

Cover & Hart (1967) gave the first analysis of 1-NN, showing that its risk converges to twice the Bayes optimal error under mild conditions. Following a lemma due to Stone (1977), Devroye & Györfi (1985) have shown that the k-NN rule

is consistent (with respect to the hypothesis class of all functions from $\mathbb{R}^d$ to $\{0,1\}$). A good presentation of the analysis is given in the book of Devroye et al. (1996). Here, we give a finite sample guarantee that explicitly underscores the prior assumption on the distribution. See Section 7.4 for a discussion on consistency results. Finally, Gottlieb, Kontorovich & Krauthgamer (2010) derived another finite sample bound for NN that is more similar to VC bounds.

19.6 Exercises

In this exercise we will prove the following theorem for the **k-NN** rule.

THEOREM 19.5 *Let $\mathcal{X} = [0,1]^d, \mathcal{Y} = \{0,1\}$, and $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$ for which the conditional probability function, η, is a c-Lipschitz function. Let h_S denote the result of applying the k-NN rule to a sample $S \sim \mathcal{D}^m$, where $k \geq 10$. Let $h^\star$ be the Bayes optimal hypothesis. Then,*

$$\mathop{\mathbb{E}}_{S}[L_{\mathcal{D}}(h_S)] \leq \left(1 + \sqrt{\frac{8}{k}}\right) L_{\mathcal{D}}(h^\star) + \left(6\, c\sqrt{d} + k\right) m^{-1/(d+1)}.$$

1. Prove the following lemma.

   LEMMA 19.6 *Let $C_1, \ldots, C_r$ be a collection of subsets of some domain set, $\mathcal{X}$. Let S be a sequence of m points sampled i.i.d. according to some probability distribution, $\mathcal{D}$ over $\mathcal{X}$. Then, for every $k \geq 2$,*

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}^m}\left[\sum_{i:|C_i \cap S|<k} \mathbb{P}[C_i]\right] \leq \frac{2rk}{m}.$$

   Hints:

   - Show that

$$\mathop{\mathbb{E}}_{S}\left[\sum_{i:|C_i \cap S|<k} \mathbb{P}[C_i]\right] = \sum_{i=1}^{r} \mathbb{P}[C_i] \mathop{\mathbb{P}}_{S}\left[|C_i \cap S| < k\right].$$

   - Fix some i and suppose that $k < \mathbb{P}[C_i]\, m/2$. Use Chernoff's bound to show that

$$\mathop{\mathbb{P}}_{S}\left[|C_i \cap S| < k\right] \leq \mathop{\mathbb{P}}_{S}\left[|C_i \cap S| < \mathbb{P}[C_i]m/2\right] \leq e^{-\mathbb{P}[C_i]\, m/8}.$$

   - Use the inequality $\max_a ae^{-ma} \leq \frac{1}{me}$ to show that for such i we have

$$\mathbb{P}[C_i] \mathop{\mathbb{P}}_{S}\left[|C_i \cap S| < k\right] \leq \mathbb{P}[C_i] e^{-\mathbb{P}[C_i]\, m/8} \leq \frac{8}{me}.$$

   - Conclude the proof by using the fact that for the case $k \geq \mathbb{P}[C_i]\, m/2$ we clearly have:

$$\mathbb{P}[C_i] \mathop{\mathbb{P}}_{S}\left[|C_i \cap S| < k\right] \leq \mathbb{P}[C_i] \leq \frac{2k}{m}.$$

2. We use the notation $y \sim p$ as a shorthand for "y is a Bernoulli random variable with expected value p." Prove the following lemma:

   LEMMA 19.7 *Let $k \geq 10$ and let $Z_1, \ldots, Z_k$ be independent Bernoulli random variables with $\mathbb{P}[Z_i = 1] = p_i$. Denote $p = \frac{1}{k}\sum_i p_i$ and $p' = \frac{1}{k}\sum_{i=1}^{k} Z_i$. Show that*

   $$\mathop{\mathbb{E}}_{Z_1,\ldots,Z_k} \mathop{\mathbb{P}}_{y\sim p}[y \neq \mathbb{1}_{[p'>1/2]}] \leq \left(1 + \sqrt{\frac{8}{k}}\right) \mathop{\mathbb{P}}_{y\sim p}[y \neq \mathbb{1}_{[p>1/2]}].$$

   Hints:
   W.l.o.g. assume that $p \leq 1/2$. Then, $\mathbb{P}_{y\sim p}[y \neq \mathbb{1}_{[p>1/2]}] = p$. Let $y' = \mathbb{1}_{[p'>1/2]}$.
   - Show that

     $$\mathop{\mathbb{E}}_{Z_1,\ldots,Z_k} \mathop{\mathbb{P}}_{y\sim p}[y \neq y'] - p = \mathop{\mathbb{P}}_{Z_1,\ldots,Z_k}[p' > 1/2](1-2p).$$

   - Use Chernoff's bound (Lemma B.3) to show that

     $$\mathbb{P}[p' > 1/2] \leq e^{-k\,p\,h\left(\frac{1}{2p}-1\right)},$$

     where

     $$h(a) = (1+a)\log(1+a) - a.$$

   - To conclude the proof of the lemma, you can rely on the following inequality (without proving it): For every $p \in [0, 1/2]$ and $k \geq 10$:

     $$(1-2p)\, e^{-k\,p + \frac{k}{2}(\log(2p)+1)} \;\leq\; \sqrt{\frac{8}{k}}\, p.$$

3. Fix some $p, p' \in [0,1]$ and $y' \in \{0,1\}$. Show that

   $$\mathop{\mathbb{P}}_{y\sim p}[y \neq y'] \leq \mathop{\mathbb{P}}_{y\sim p'}[y \neq y'] + |p - p'|.$$

4. Conclude the proof of the theorem according to the following steps:
   - As in the proof of Theorem 19.3, six some $\epsilon > 0$ and let $C_1, \ldots, C_r$ be the cover of the set $\mathcal{X}$ using boxes of length ϵ. For each $\mathbf{x}, \mathbf{x}'$ in the same box we have $\|\mathbf{x} - \mathbf{x}'\| \leq \sqrt{d}\,\epsilon$. Otherwise, $\|\mathbf{x} - \mathbf{x}'\| \leq 2\sqrt{d}$. Show that

     $$\mathop{\mathbb{E}}_{S}[L_{\mathcal{D}}(h_S)] \leq \mathop{\mathbb{E}}_{S}\left[\sum_{i:|C_i \cap S|<k} \mathbb{P}[C_i]\right] + \max_i \mathop{\mathbb{P}}_{S,(\mathbf{x},y)}\left[h_S(\mathbf{x}) \neq y \mid \forall j \in [k],\ \|\mathbf{x} - \mathbf{x}_{\pi_j(\mathbf{x})}\| \leq \epsilon\sqrt{d}\right]. \quad (19.3)$$

   - Bound the first summand using Lemma 19.6.
   - To bound the second summand, let us fix $S|_x$ and $\mathbf{x}$ such that all the k neighbors of $\mathbf{x}$ in $S|_x$ are at distance of at most $\epsilon\sqrt{d}$ from $\mathbf{x}$. W.l.o.g assume that the k NN are $\mathbf{x}_1, \ldots, \mathbf{x}_k$. Denote $p_i = \eta(\mathbf{x}_i)$ and let $p = \frac{1}{k}\sum_i p_i$. Use Exercise 3 to show that

     $$\mathop{\mathbb{E}}_{y_1,\ldots,y_j} \mathop{\mathbb{P}}_{y\sim\eta(\mathbf{x})}[h_S(\mathbf{x}) \neq y] \leq \mathop{\mathbb{E}}_{y_1,\ldots,y_j} \mathop{\mathbb{P}}_{y\sim p}[h_S(\mathbf{x}) \neq y] + |p - \eta(\mathbf{x})|.$$

W.l.o.g. assume that $p \leq 1/2$. Now use Lemma 19.7 to show that

$$\mathop{\mathbb{P}}_{y_1,\ldots,y_j} \mathop{\mathbb{P}}_{y\sim p}[h_S(\mathbf{x}) \neq y] \leq \left(1 + \sqrt{\frac{8}{k}}\right) \mathop{\mathbb{P}}_{y\sim p}[\mathbb{1}_{[p>1/2]} \neq y].$$

- Show that

$$\mathop{\mathbb{P}}_{y\sim p}[\mathbb{1}_{[p>1/2]} \neq y] = p = \min\{p, 1-p\} \leq \min\{\eta(\mathbf{x}), 1-\eta(\mathbf{x})\} + |p-\eta(\mathbf{x})|.$$

- Combine all the preceding to obtain that the second summand in Equation (19.3) is bounded by

$$\left(1 + \sqrt{\frac{8}{k}}\right) L_{\mathcal{D}}(h^\star) + 3\,c\,\epsilon\,\sqrt{d}.$$

- Use $r = (2/\epsilon)^d$ to obtain that:

$$\mathop{\mathbb{E}}_{S}[L_{\mathcal{D}}(h_S)] \leq \left(1 + \sqrt{\frac{8}{k}}\right) L_{\mathcal{D}}(h^\star) + 3\,c\,\epsilon\,\sqrt{d} + \frac{2(2/\epsilon)^d\,k}{m}.$$

  Set $\epsilon = 2m^{-1/(d+1)}$ and use

$$6\,c\,m^{-1/(d+1)}\,\sqrt{d} + \frac{2k}{e} m^{-1/(d+1)} \leq \left(6c\sqrt{d} + k\right) m^{-1/(d+1)}$$

  to conclude the proof.