CHAPTER 7

The Nearest Neighbor Rule

Recall the pattern recognition problem that we have discussed so far. We are dealing with objects that belong to one of the two classes 0 and 1. We observe a feature vector $\overline{x}$ of an object and we would like to decide whether the object belongs to class 0 or class 1. We want a decision rule that minimizes the probability of error.

If the prior probabilities, $P(0)$ and $P(1)$, and the conditional densities $p(\overline{x}|0)$ and $p(\overline{x}|1)$ are known, then we can implement Bayes decision rule. If these are unknown, then we assume that we have some "training data" $(\overline{x}_1, y_1), (\overline{x}_2, y_2), \ldots, (\overline{x}_n, y_n)$ and wish to come up with a good classification rule on the basis of this data. This is the problem of learning from examples.

In Chapter 6, we described a brute-force approach—first estimate the unknown prior probabilities and conditional densities and then substitute these estimates into the equations for Bayes decision rule. This approach does not work well for reasons described in the previous chapter. In this and in the following chapters, we discuss more practical approaches that bypass the need to first estimate the unknown distributions.

7.1 THE NEAREST NEIGHBOR RULE

Perhaps the simplest decision rule one might come up with is to find in the training data the feature vector $\overline{x}_i$ that is closest to $\overline{x}$, and then decide that $\overline{x}$ belongs to the same class as given by the label y_i. This decision rule is called the "nearest neighbor rule" (NN rule) and can be illustrated by a figure (see Figure 7.1).

In Figure 7.1, the feature vectors are two-dimensional, and so each $\overline{x}_i$ can be represented by a point in the plane. Associated with each $\overline{x}_i$ is a region (called the Voronoi region) consisting of all points that are closer to $\overline{x}_i$ than to any other $\overline{x}_j$. That is, the points in the region associated with $\overline{x}_i$ have $\overline{x}_i$ as their nearest neighbor among the set $\overline{x}_1, \ldots, \overline{x}_n$.

An Elementary Introduction to Statistical Learning Theory, First Edition.
Sanjeev Kulkarni and Gilbert Harman.

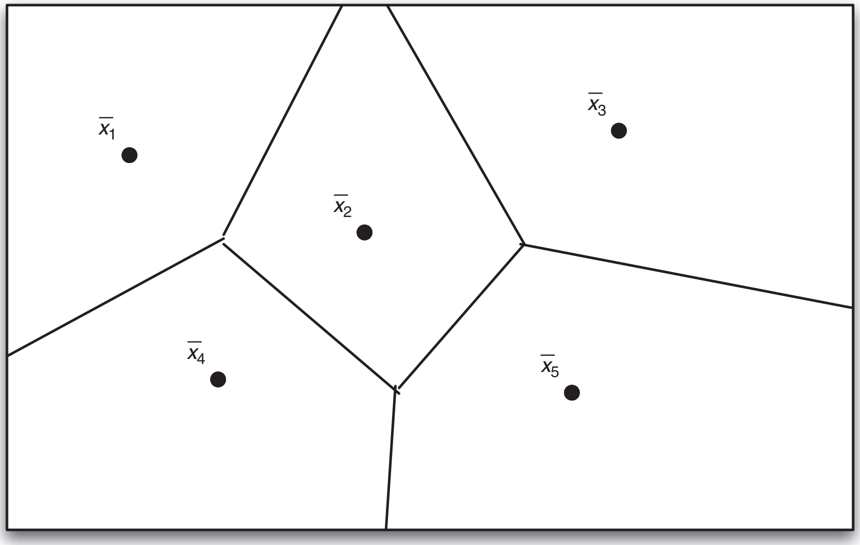

Figure 7.1 Voronoi regions.

Recall that associated with each $\overline{x}_i$ is a label y_i that is either 0 or 1. The NN rule simply classifies a feature vector $\overline{x}$ according to the label associated with the region to which $\overline{x}$ belongs. Alternatively, we can think of the $\overline{x}_i$ as "prototypes." The NN rule classifies a given $\overline{x}$ by assigning it to the same class as the closest prototype.

Given the simplicity of the NN rule, some questions immediately come to mind. For example, is this really a reasonable classification rule? How well does it perform? How should we measure its performance?

7.2 PERFORMANCE OF THE NEAREST NEIGHBOR RULE

As we mentioned above, a natural benchmark for measuring the performance of the NN rule (or any decision rule, for that matter) is the Bayes error rate. Recall that we denote the Bayes error rate by R^*, the "*" indicating optimality (since no rule can have error rate less than R^*).

To discuss the error rate of the NN rule, we need to clarify what we mean. Since the NN rule depends on the data $(\overline{x}_1, y_1), (\overline{x}_2, y_2), \ldots, (\overline{x}_n, y_n)$, the performance of the NN rule will depend on the specific training examples we have seen. Since these training examples are randomly drawn, we cannot expect to do well for every case—there is a chance that we will be unlucky and see particularly poor data. To avoid these problems, we consider the expected performance of the NN rule. The expectation is with respect to the new instance to be classified (as usual) as well as the training examples.

Of course, the expected performance also depends on how much training data have been seen. Let R_n denote the expected error rate of the NN rule after n training examples, and let R_∞ denote the limit of R_n as n tends to infinity. R_∞ is the *asymptotic* expected error rate of the NN rule. That is, it measures the performance of the NN rule as the training sample size tends to infinity. We will focus on R_∞.

What can we say about R_∞? More specifically, can we get bounds on R_∞ in terms of R^*?

A lower bound on R_∞ is easy. Certainly, we must have $R^* \leq R_\infty$, since the Bayes decision rule is optimal and so the nearest neighbor rule (even with infinitely many training examples) can do no better.

The more interesting bounds are upper bounds on R_∞ in terms of R^*. The basic result is that the asymptotic error rate of the NN rule is no worse than twice the Bayes error rate, that is, $R_\infty \leq 2R^*$. Combining the lower and upper bounds, we have

$$R^* \leq R_\infty \leq 2R^*. \tag{7.1}$$

It turns out that one can obtain the following more refined bound:

$$R^* \leq R_\infty \leq 2R^*(1 - R^*). \tag{7.2}$$

Since error rates (and so R^* in particular) must be between 0 and 1, we see that $1 - R^* \leq 1$. Hence, $2R^*(1 - R^*) \leq 2R^*$, so that the simpler bound $2R^*$ follows easily from the more refined upper bound.

The factor of 2 in the upper bound may not seem so good, but for small R^*, even $2R^*$ will be small. Also, in general, we cannot say anything stronger than the bounds given in the sense that there are examples that achieve these bounds. That is, there are choices for the underlying probability distributions for which the performance of the NN rule achieves either the upper or the lower bound.

The upper bounds of Equation (7.1) and (7.2) might seem somewhat surprising. Using only random labeled examples but knowing nothing about the underlying distributions, we can (in the limit) achieve an error rate no worse than twice the error rate that could be achieved by knowing everything about the probability distributions. Moreover, we can do this with an extremely simple rule that bases its decision on just the nearest neighbor to the feature vector we wish to classify. This result is sometimes interpreted informally by saying that half the information is contained in the nearest neighbor.

7.3 INTUITION AND PROOF SKETCH OF PERFORMANCE*

Here we give some intuition and a rough proof sketch on how to obtain the upper bound on the asymptotic error rate.

As we get more and more data, the distance between $\overline{x}$ and its nearest neighbor goes to zero. In computing R_∞, it is as though the distance *is* zero. When the distance between $\overline{x}$ and its nearest neighbor (say $\overline{x}_i$) is zero, it is like guessing the outcome of a coin flip (the label y corresponding to $\overline{x}$) on the basis of another independent coin flip (the label y_i corresponding to $\overline{x}_i$).

Let $R^*(\overline{x})$ be the conditional Bayes error rate, given $\overline{x}$. Then R^* is the average of $R^*(\overline{x})$ over all possible feature vectors $\overline{x}$.

Given $\overline{x}$ (and that the nearest neighbor is distance 0), the probability that the NN rule makes a correct decision is just the probability that both labels are 0 or both labels are 1. That is,

$$\begin{aligned} P(\text{correct}|\overline{x}) &= P(0|\overline{x})P(0|\overline{x}) + P(1|\overline{x})P(1|\overline{x}) \\ &= P(0|\overline{x})^2 + P(1|\overline{x})^2. \end{aligned}$$

Now, recall that the Bayes decision rule simply decides the class that has the larger posterior probability, and so makes an error with the probability that is the smaller of the posterior probabilities. Or, conditioned on $\overline{x}$,

$$R^*(\overline{x}) = \min\{P(0|\overline{x}), P(1|\overline{x})\}.$$

Hence, $R^*(\overline{x})$ is equal to one of $P(0|\overline{x})$ or $P(1|\overline{x})$, and $1 - R^*(\overline{x})$ is equal to the other.

In either case, substituting into the expression for $P(\text{correct}|\overline{x})$, we get

$$P(\text{correct}|\overline{x}) = R^*(\overline{x})^2 + (1 - R^*(\overline{x}))^2 \tag{7.3}$$

$$= 1 - 2R^*(\overline{x}) + 2(R^*(\overline{x}))^2. \tag{7.4}$$

Simple Bound

To obtain the simple upper bound that $R_\infty \leq 2R^*$, note that in Equation (7.4) the term $2(R^*(\overline{x}))^2$ is greater than or equal to zero, so that

$$P(\text{correct}|\overline{x}) \geq 1 - 2R^*(\overline{x}).$$

Therefore,

$$P(\text{error}|\overline{x}) = 1 - P(\text{correct}|\overline{x}) \leq 2R^*(\overline{x})$$

and averaging over $\overline{x}$ we get $R_\infty = P(\text{error}) \leq 2R^*$.

Tighter Bound

To get the tighter upper bound of Equation (7.2), we need to be more careful. From Equation (7.4), we get

$$P(\text{error}|\overline{x}) = 1 - P(\text{correct}|\overline{x}) = 2R^*(\overline{x}) - 2(R^*(\overline{x}))^2.$$

Then

$$\begin{aligned} R_\infty = P(\text{error}) &= E[2R^*(\overline{x}) - 2(R^*(\overline{x}))^2] \\ &= 2R^* - 2\,E[(R^*(\overline{x}))^2] \\ &\le 2R^* - 2(R^*)^2, \end{aligned}$$

since the average of squares is greater than or equal to the square of the average. Thus, $R_\infty \le 2R^*(1 - R^*)$.

7.4 USING MORE NEIGHBORS

Despite its simplicity, the NN rule has impressive performance. Yet, it is natural to ask whether we can do any better. For example, why not use several neighbors, rather than just the *nearest* neighbor?

This is a reasonable suggestion and in fact leads to useful extensions of the (single) NN rule. For example, consider the k-NN rule, in which we use the k nearest neighbors, for some fixed number k. With this rule, given an observed feature vector $\overline{x}$, we use the k nearest neighbors of $\overline{x}$ from among $\overline{x}_1, \ldots, \overline{x}_n$, and take a majority vote of the labels corresponding to these k nearest neighbors. Let R_∞^k be the error rate of the k-NN rule in the limit of infinite data. We might expect that R_∞^k improves (gets smaller) for larger k. This is often the case, but not always. For example, under certain conditions it can be shown that

$$R^* \le R_\infty^k \le \left(1 + \frac{1}{k}\right) R^*.$$

However, it can also be shown that there are some distributions for which the 1-NN rule outperforms the k-NN rule for any fixed $k > 1$.

Another very useful idea is to let the number of neighbors used grow with n (the amount of data we have). That is, we can let k be a function of n, so that we use a k_n-NN rule. If we do this, how should we choose k_n?

We need $k_n \to \infty$ so that we use more and more neighbors as the amount of training data increases. But we should make sure that $\frac{k_n}{n} \to 0$, so that asymptotically the number of neighbors we use is a negligible fraction of the total amount of data. This will ensure that we use neighbors that get closer and closer to the observed feature vector $\overline{x}$. For example, we might let $k_n = \sqrt{n}$ to satisfy both conditions.

It turns out that with any such k_n (such that $k_n \to \infty$ and $k_n/n \to 0$ are satisfied), we get $R_n^{k_n} \to R^*$ as $n \to \infty$. That is, in the limit as the amount of training data grows, the performance of the k_n-NN rule approaches that of the optimal Bayes decision rule! What is surprising about this result is that by observing data but without knowing anything about the underlying distributions, asymptotically we can do as well as if we knew the underlying distributions completely. And, this works without assuming that the underlying distributions take on any particular form or satisfy any stringent conditions. In this sense, the k_n-NN rule is called *universally consistent*, and is truly *nonparametric* learning in that the underlying distributions can be arbitrary and we need no knowledge of their form. It was not known until the early 1970s whether universally consistent rules existed, and it was quite surprising when the k_n-NN rules along with some others were shown to be universally consistent. A number of such decision rules are known today.

However, universal consistency is not the end of the story. This is just an asymptotic property (in the limit of infinite data), and "in the long run, we're all dead" (Keynes). A critical issue is that of convergence rates. Many results on the convergence rates of the NN rule and other rules are known. A fairly generic problem is that for most methods the rate of convergence is very slow in high-dimensional spaces. This is a facet of the so-called "curse of dimensionality" mentioned in Section 6.5. As we discussed, in many real applications the dimension can be extremely large, which bodes ill for many methods. Furthermore, it can be shown that there are no "universal" rates of convergence. That is, for any method, one can always find distributions for which the convergence rate is arbitrarily slow. Thus, as we also mentioned in Section 6.5 in the context of "No Free Lunch Theorems," there is no one method that can universally beat out all other methods. These results make the field continue to be exciting, and makes the design of good learning algorithms and the understanding of their performance an important science and art. In the coming chapters, we discuss some other methods useful in practice, as well as some results on what can be said with a finite amount of training data (see Chapters 11, 12, and 13).

7.5 SUMMARY

In this chapter, we discussed a simple learning method that uses training data to come up with a classification rule. The NN rule classifies a new example in the same class as the nearest feature vector in the training data. Surprisingly, this simple rule has a performance no worse than twice that of the optimal Bayes decision rule. We gave a proof sketch for this result as well as slightly more refined result. We then discussed the possibility of using more neighbors. By letting the number of neighbors k_n that we use grow with the data (so that $k_n \to \infty$), but such that $k_n/n \to 0$, we can actually do as well as Bayes decision rule in the limit as $n \to \infty$. No knowledge or assumptions about the underlying distributions are needed, so that this k_n NN rule is an example of what are called universally consistent rules.

7.6 APPENDIX: WHEN PEOPLE USE NEAREST NEIGHBOR REASONING

7.6.1 Who Is a Bachelor?

How do we decide whether someone is a bachelor? One theory is that we have learned a definition or rule: a bachelor is an unmarried adult male person.

But what about the Pope? He is an unmarried adult male person, but is he a bachelor? Many people would say he is not. What about an unmarried adult male who has been living with a woman “as man and wife,” although they are not legally married. Many people would not consider that man a bachelor.

Furthermore, what about a man who is married but in the process of getting a divorce, who has not lived with his legal wife for several years and has been dating regularly. Many people say that he is a bachelor even though he is still legally married.

Do people use nearest neighbor reasoning to decide what they think about these cases?

7.6.2 Legal Reasoning

A court wants to decide whether a particular case falls under a legal rule, for example the rule, “No vehicles are allowed in Princeton Park.” Some cases are easy to classify, as when someone drives the family sedan into the park. That is clearly forbidden by the rule, but other cases are less clear. Suppose a teenager rides her bicycle into the park. Do bicycles count as *vehicles* under the rule? Do wheelchairs? Rollerblades?

Or consider the legal rule that, if a will clearly specifies that the estate is left to a particular person and that person is living and competent, then the person in question is to inherit the estate. In most cases it is quite clear whether this rule applies. But suppose that the person thus specified in the will has murdered the deceased? Must that person inherit under this rule? Courts have ruled that the person may not inherit.

Or consider the rule that, if one person’s wrongful act causes damage to another, then the first person must reimburse the victim for the damage, except to the extent that the victim was responsible for the damage. How does this rule apply to a case in which a driver causes an accident that kills a woman’s husband and she has a heart attack on learning of her husband’s death? Under this rule must the driver reimburse the wife for her medical expenses and other damage she suffers because of her heart attack?

Various factors enter into legal decisions about such hard cases, especially including *precedent*. A court tries to give the fairest decision, given the facts of the case and given prior decisions that courts have made in similar cases. Is the present case more similar to previously decided case 1, in which case the plaintiff should prevail, or is it more similar to previously decided case 2, in which case the defendant should prevail? Reasoning from precedent is a version of a nearest neighbor strategy for deciding new cases.

7.6.3 Moral Reasoning

In moral reasoning we try to find general principles that apply to particular cases and we also test our general principles against our settled judgments about other cases. So here too is a use of a nearest neighbor strategy.

"Is it morally OK to raise animals for food?" Jack asks Jill. "Of course," she replies. "But would it be morally OK to raise people for food?" "Of course not," she says. Then what is the difference? Perhaps Jill thinks that people are different from (other) animals in morally significant ways. Perhaps it is because people are rational in ways that animals are not rational. But why does that matter? And what about chimpanzees, which seem pretty smart, smarter than young human children? And would it be morally okay to raise brain damaged people for food as long as they will not ever be rational?

In moral reflection of this sort, you try to adjust some of your views in the light of other "nearby" opinions you hold. Perhaps you try to reach a *reflective equilibrium* (Rawls, 1971) between your principles and your particular views about examples so that there is no conflict between principles.

7.7 QUESTIONS

1. (a) What is the NN rule and how does the expected error from the use of this rule compare with the Bayes error rate?

(b) What conditions on k_n are required for the k_n-NN rule to have an asymptotic error rate equal to the Bayes error rate?

2. Recall that for the 1-NN rule, the region associated with a feature vector $\overline{x}_i$ is the set of all points that are closer to x_i than to any of the other feature vectors $\overline{x}_j$ for $j \neq i$. These are the Voronoi regions. Sketch the Voronoi regions for feature vectors $\overline{x}_1 = (0, 0)$, $\overline{x}_2 = (0, 2)$, $\overline{x}_3 = (2, 0)$, and $\overline{x}_4 = (1, 1)$.

3. Come up with a case (i.e., give the prior probabilities and conditional densities) in which the error rate of the NN rule equals the Bayes error rate, and briefly explain why this happens in the case you give.

4. Describe as precisely as you can the tradeoffs of having a small k_n versus a large k_n in the k_n-nearest neighbor classifier. What happens in the extreme cases when $k_n = 1$ and when $k_n = n$?

5. What conditions are required on k_n for the k_n-NN rule to be universally consistent?

6. If we use a k_n-NN rule with $k_n = n$, what would be the resulting error rate in terms of $P(0)$, $P(1)$, $P(\overline{x}|0)$, and $P(\overline{x}|1)$?

7. Briefly discuss the following position. Under appropriate conditions, the k_n-NN rule is universally consistent, so the choice of features does not matter.

7.8 REFERENCES

Nearest neighbor methods were introduced by Fix and Hodges (1951, 1952) in the early 1950s. Cover and Hart (1967) obtained the now classic result that $R_\infty \leq 2R^*$. A great deal of work on nearest neighbor methods has been done since then. Most books or review papers on statistical pattern recognition discuss nearest neighbor methods. For example, see Chapter 4 of Duda *et al*. (2001) and Section 2 of Kulkarni *et al*. (1998).

Devroye *et al*. (1996) has several in-depth chapters on nearest neighbor methods and their performance as well an extensive bibliography with pointers to original research results. Dasarathy (1991) contains a broad, but less recent, collection of work on nearest-neighbor methods. There is also a useful account in Mitchell (1997).

There is a discussion of "bachelor" in Winograd and Flores (1986), p. 112. For legal reasoning, see e.g. Dworkin (1986), Chapters 1–2. Stich (1993) discusses moral reasoning.

Cover TM, Hart PE. Nearest neighbor pattern classification. IEEE Trans Inf Theory 1967;13(1):21–27.

Dasarathy BV, editor. Nearest neighbor (NN) norms: NN pattern classification techniques. Washington (DC): IEEE Computer Society; 1991.

Devroye L, Györfi L, Lugosi G. A probabilistic theory of pattern recognition. New York: Springer Verlag; 1996.

Duda RO, Hart, PE, Stork, DG. Pattern classification. 2nd ed. New York: Wiley; 2001.

Dworkin R. Law's empire. Cambridge (MA): Harvard UP; 1986.

Fix E, Hodges JL. Discriminatory analysis: nonparametric discrimination: consistency properties. USAF Sch Aviat Med 1951;4:261–279.

Fix E, Hodges JL. Discriminatory analysis: nonparametric discrimination: small sample performance. USAF Sch Aviat Med 1952;11:280–322.

Kulkarni SR, Lugosi G, Venkatesh S. Learning pattern classification—A survey. IEEE Trans Inf Theory 1998;44(6):2178–2206.

Mitchell TM. Instance-based learning. Machine learning. Boston (MA): McGraw-Hill; 1997. pp. 226–229, Chapter 8.

Rawls J. A theory of justice. Cambridge (MA): Harvard UP; 1971.

Stich SP. Moral philosophy and mental representation. In: Hechter M, Nadel L, Michod R, editors. The origin of values. Hawthorne (NY): Aldine de Gruyter; 1993. pp. 215–228.

Winograd T, Flores F. Understanding computers and cognition. Norwood (NJ): Ablex; 1986.