Nearest Neighbor Classifiers

1 The 1 Nearest-Neighbor (1-N-N) Classifier

The 1-N-N classifier is one of the oldest methods known. The idea is extremely simple: to classify $\mathbf{X}$ find its closest neighbor among the training points (call it $\mathbf{X}^{'}$,) and assign to $\mathbf{X}$ the label of $\mathbf{X}^{'}$.

1.1 Questions

What is good about this method?

- It is conceptually simple.
- It does not require learning (*term: memory-base*).
- It can be used even with few examples.
- Even for moderate k: wonderful performer.
- It works very well in low dimensions for complex decision surfaces.
- For $R^* = 0$: consistent.

What is bad about this method?

- For fixed k it is asymptotically suboptimal, but not by much.
- Classification is slow.
- It suffers A LOT from the curse-of-dimensionality.

1.2 Distance Functions and Metric Spaces

Def. If $\mathcal{X}$ is a set, $D(\cdot,\cdot)$ is a function $\mathcal{X} \times \mathcal{X} \stackrel{D(\cdot,\cdot)}{\rightarrow} \mathbb{R}$ is a *distance* if, for all $\mathbf{X},\ \mathbf{Z},\ \mathbf{V} \in \mathcal{X}$ the following three properties hold:

- $D(\mathbf{X},\mathbf{Z}) \geq 0$ and $D(\mathbf{X},\mathbf{X}) = 0$ (nonnegativity),
- $D(\mathbf{X},\mathbf{Z}) = D(\mathbf{Z},\mathbf{X})$ (symmetry),

- $D(\mathbf{X}, \mathbf{Z}) \leq D(\mathbf{X}, \mathbf{V}) + D(\mathbf{V}, \mathbf{Z})$ (triangular inequality).

Def. A *metric space* is a pair $(\mathcal{X}, D(\cdot, \cdot))$, where $\mathcal{X}$ is a set, and $D(\cdot, \cdot)$ is a distance (or metric) on $\mathcal{X}$.

Def. A metric space $\mathcal{X}$ is *separable* if there is a countable dense subset A of $\mathcal{X}$. (E.g.: real line, countable dense subset: rationals)

We will use the concept of separability to prove a fundamental theorem in the next section.

2 Asymptotics for 1-N-N

We first prove that the nearest neighbor of $\mathbf{X}$ converges almost surely to $\mathbf{X}$ as the training size grows to infinity. This means that, for a set of training set sequences having probability 1[1] the following property holds: For every ϵ there is a n_ϵ (dependent on the specific sequence of training set) such that for all $n > n_\epsilon$ the distance between the (random!) sample $\mathbf{X}$ to be classified and its nearest neighbor is less than ϵ.

Theorem Convergence of Nearest Neighbor *If* $\mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n$ *are i.i.d. in a separable metric space* $\mathcal{X}$, $\mathbf{X}'_n(\mathbf{X}) \stackrel{\triangle}{=} \mathbf{X}'_n$ *is the closest of the* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *to* $\mathbf{X}$ *in a metric* $D(\cdot, \cdot)$*, then*

$$\mathbf{X}'_n \to \mathbf{X} \quad a.s.$$

Proof To prove the theorem we show that the probability that $\mathbf{X}$ does not satisfy the desired property converges to zero exponentially fast with the training set size.

Recall now that $\mathbf{X}$ is a random variable, namely, it is a point of $\mathcal{X}$ selected according to a probability measure. Let's divide the points of $\mathcal{X}$ in "good" points and "bad" points. "Good" points satisfy the desired property. "Bad" points do not. We will show that "bad" points form a set of probability zero, while "good" points do not satisfy the desired property with probability exponentially small in the training set size.

We start by denoting by $S_{\mathbf{x}}(r)$ a sphere centered at $\mathbf{x}$, of radius r.

A "good" point for us is a point $\mathbf{x}$ such that, for every $\delta > 0$, the law of $\mathbf{X}$ assigns positive probability to a sphere centered at $\mathbf{x}$ and having radius δ; formally: $\forall \delta > 0 \;\; \Pr\{S_{\mathbf{x}}(\delta)\} > 0$.

The probability that the nearest neighbor of $\mathbf{x}$ does not fall into $S_{\mathbf{x}}(\delta)$ is the probability that no point in the training set falls within such sphere.

[1] It is the set of sequences that has probability 1: we are not talking of a set containing sequences having probability 1.

Since the training points are independent, the probability that all training points lie outside $S_{\mathbf{x}}(\delta)$ is the product of the probability that each individual point lies outside $S_{\mathbf{x}}(\delta)$, which is the nth power of the probability that any of those points lies outside $S_{\mathbf{x}}(\delta)$ because the points are identically distributed according to the same law as $\mathbf{X}$.

Formally:

$$\Pr\left\{d(\mathbf{X}_n^{'}(\mathbf{x}),\mathbf{x}) > \delta\right\} = \Pr\left\{\mathbf{X}_n^{'}(\mathbf{x}) \notin S_{\mathbf{x}}(r)\right\} = (1 - \Pr\{S_{\mathbf{x}}(\delta)\})^n \to 0.$$

Now, recall the (first) Borel-Cantelli Lemma: if the events $A_1, A_2, \ldots$ of an infinite sequence have probabilities that sum to a finite number ($\sum_i \Pr\{A_i\} = S < \infty$), then the probability that an outcome $\mathbf{X}$ occurs that belong to infinitely many such events is equal to zero.

In our case, $\mathbf{X}^{'} \in S_{\mathbf{x}}(\delta)$ eventually. Since δ was arbitrary, we conclude that

$$\Pr\left\{\lim_{n\to\infty} d(\mathbf{X}_n^{'}(\mathbf{x}),\mathbf{x}) <= \delta\right\} = 1,$$

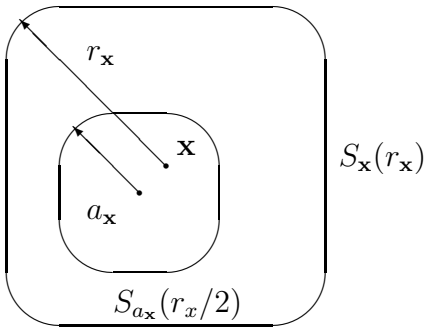
for all $\delta > 0$, namely that

$$\Pr\left\{\lim_{n\to\infty} d(\mathbf{X}_n^{'},\mathbf{x}) = 0\right\} = 1.$$

Now, all we have to show is that the "good" $\mathbf{x}$'s form a set of probability 1, namely, that the "bad" $\mathbf{x}$'s form a set of probability zero.[2].

Call N the set of $\mathbf{x}$ where $\exists r_{\mathbf{x}}$ s.t. $\forall r < r_{\mathbf{x}}, \Pr\{\ S_{\mathbf{x}}(r)\} = 0$ (the bad set). We want to show that $\Pr\{N\} = 0$.

Now we use separability: every $\mathbf{x}$ can be approximated as we wish by a point in A. More specifically, $\forall \mathbf{x} \in N\ \exists a_{\mathbf{x}} \in A$ for which $a_{\mathbf{x}} \in S_{\mathbf{x}}(r_{\mathbf{x}}/3)$. Pick such $a_{\mathbf{x}}$ and draw a sphere around it, all contained in $S_{\mathbf{x}}(r_b x)$, and containing $\mathbf{x}$. For example, we can choose the radius of such sphere to be equal to $r_{\mathbf{x}}/2$: call this sphere $S_{a_{\mathbf{x}}}(r_{\mathbf{x}}/2)$. Since $S_{a_{\mathbf{x}}}(r_{\mathbf{x}}/2) \subset S_{\mathbf{x}}(r_{\mathbf{x}})$, we can conclude that $\Pr\{S_{a_{\mathbf{x}}}(r_{\mathbf{x}}/2)\} = 0$. The following picture illustrates the construction (sorry for the L^p spheres that are not round...)



[2]We could, as suggested in class, assume that this is true. However, every time we wanted to use the theorem we would have to show that the "bad" set has probability zero

Repeat the same construction for every $\mathbf{x} \in N$. If the same $a_{\mathbf{x}}$ is selected multiple times, so that multiple spheres are constructed, use the largest such sphere, or the closure of the supremum. Note that each point of N is contained in at least one of the above defined spheres, so N is contained in the union of these spheres. Note, additionally, that we have at most a countable number of such spheres, one per element a of the countable set A. Hence, since probability is monotonic, the probability of N (that is, the probability that $\mathbf{X} \in N$) is less than or equal to the probability of the union of the spheres, which is less than or equal to the sum of the probabilities of the individual spheres (because we have countably many of them). But each addend is equal to zero, and therefore $\Pr\{N\} = 0$.
QED.

Now we start working on the asymptotic risk of Let $R_n^{(1)}$ be the risk of the $1-$N-N with n training samples. We first claim that the risk of the $1-$N-N converges (that is, it has a limit).

Theorem If $\mathcal{X}$ is a separable metric space,

$$\lim_{n\to\infty} R_n^{(1)} = R^{(1)}$$

Then, we find a functional form of the risk of the $1-$N-N

Theorem *If $\mathcal{X}$ is a separable metric space,*

$$R^{(1)} = E\left[2\eta(\mathbf{X})\left(1-\eta(\mathbf{X})\right)\right].$$

where $\eta(\mathbf{X}) = Pr\{Y = 1|midX\}$ is the conditional probability of observing a sample of class 1 given that the observation is $\mathbf{X}$, and the expectation is taken with respect to the measure of $\mathbf{X}$.

This is NOT an easy theorem to prove (what is in the book does not constitute a proof). The intuition, though, is rather simple. Note that $R^{(1)} = \Pr\left\{Y' = 1, Y = 0\right\} + \Pr\left\{Y' = 0, Y = 1\right\}$, where Y' is the label of $\mathbf{X}'$. Since we showed that $\mathbf{X}'_n \to \mathbf{X}$, with probability 1, we could assume that $\mathbf{X} = \mathbf{X}'$, and that, therefore Y and Y' are independent identically distributed Bernoulli random variables[3] with parameter $\eta(\mathbf{X})$.

Then $\Pr\left\{Y' = 1, Y = 0\right\} = \Pr\left\{Y' = 1\right\}\Pr\left\{Y = 0\right\} = \eta(\mathbf{X})\big(1-\eta(\mathbf{X})\big)$, and $\Pr\left\{Y' = 0, Y = 1\right\} = \Pr\left\{Y' = 0\right\}\Pr\left\{Y = 1\right\} = \big(1-\eta(\mathbf{X})\big)\eta(\mathbf{X})$.

Again, note that **this was NOT a proof.**

We can use the result of the previous theorem in a variety of ways. For example, we can prove the following bound for $R^{(1)}$, due to Cover and Hart, which is very famous:

[3]The Bernoulli distribution describes binary random variables, for example, coin flips; it has one parameter, for example, the probability that the observation is equal to 1, or that the coin flip yields heads.

Theorem *If $\mathcal{X}$ is a separable metric space, $(\mathbf{X}, Y)$ distributed as usual*[4]*, then*

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

where the bounds are tight.

Hence: $R^{(1)}$ is at most twice the Bayes Risk. Note also $R^* = 0 \Leftrightarrow R^{(1)} = 0$, and $R^* = 1/2 \Leftrightarrow R^{(1)} = 1/2$.

Proof

$$\begin{aligned} R^{(1)} &= E\left[2\eta(\mathbf{X})\left(1 - \eta(\mathbf{X})\right)\right] \\ &= 2E\left[\min\left\{\eta(\mathbf{X}), \big(1 - \eta(\mathbf{X})\big)\right\} \max\left\{\eta(\mathbf{X}), \big(1 - \eta(\mathbf{X})\big)\right\}\right] \end{aligned}$$

where the last equality follows trivially from the fact that the product of two numbers is the product of the larger times the smaller. Now, recall that $\min\left\{\eta(\mathbf{X}), \big(1 - \eta(\mathbf{X})\big)\right\}$ is the conditional probability of error given $\mathbf{X}$ of a rule that is optimal at $\mathbf{X}$ (because such rule would choose the class with larger probability of occurrence). Call this loss $r^*(\mathbf{X})$ (because it is the loss that leads to the Bayes Risk.) Note, additionally, that $E\left[r^*(\mathbf{X})\right] = R^*$ by definition of Bayes Risk. Hence, we have

$$\begin{aligned} R^{(1)} &= 2E\left[r^*(\mathbf{X})\big(1 - r^*(\mathbf{X})\big)\right] \\ &= 2E\left[r^*(\mathbf{X}) - {r^*}^2(\mathbf{X})\right] \\ &= 2E\left[r^*(\mathbf{X}) - {r^*}^2(\mathbf{X}) + {R^*}^2 - {R^*}^2 + 2R^*r^*(\mathbf{X}) - 2R^*r^*(\mathbf{X})\right] \\ &= 2E\left[r^*(\mathbf{X}) + {R^*}^2 - 2R^*r^*(\mathbf{x}) - \left(r^*(\mathbf{X}) - R^*\right)^2\right] && (1) \\ &= 2E\left[r^*(\mathbf{X}) + {R^*}^2 - 2R^*r^*(\mathbf{x})\right] - \mathrm{var}\left(r^*(\mathbf{X})\right) && (2) \\ &\leq 2E\left[r^*(\mathbf{X}) + {R^*}^2 - 2R^*r^*(\mathbf{x})\right] && (3) \\ &= 2\Big(E\left[r^*(\mathbf{X})\right] + E\left[{R^*}^2\right] - 2E\left[R^*r^*(\mathbf{x})\right]\Big) && (4) \\ &= 2\big(R^* + {R^*}^2 - 2{R^*}^2\big) && (5) \\ &= 2R^*\big(1 - R^*\big) \end{aligned}$$

where Equation 1 follows from simple reorganization of the terms, Equation 2 follows by noting that R^* is the expected value of $r^*(\mathbf{X})$ and from the definition of variance, Inequality 3 follows from the fact that the variance is nonnegative, Equation 4 follows from the linearity of expectation, Equation 5 follows again by noting that R^* is the expected value of $r^*(\mathbf{X})$.

QED.

The previous theorem deals with the 2-class problem. The following extends the result to the M class problem.

[4]Namely, $\mathbf{X}$ is distributed according to $F(X)$, Y takes values 0 and 1, and the conditional probability of $Y = 1$ given $\mathbf{X}$ is $\eta(\mathbf{X})$.

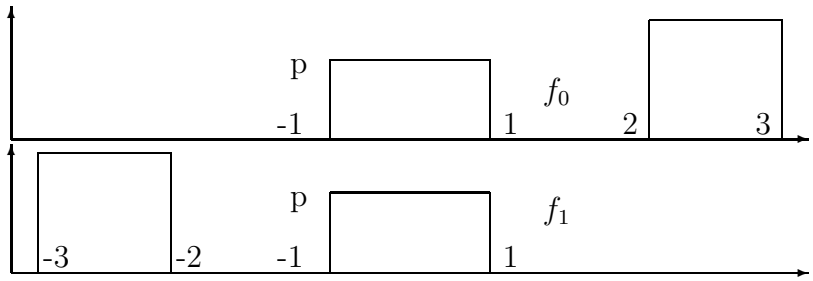
Theorem *If $\mathcal{X}$ is a separable metric space, $(\mathbf{X}, Y)$ distributed as usual, with Y being the label of one of M classes*[5]*, then*

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

From the equation, it is clear that the upper bound is smaller in the 2-class problem and becomes progressively larger as the number of classes increases. However, the value of $R^{(1)}$ is never more than twice the value of the Bayes Risk.

One might be interested in knowing if the bounds of the theorems are tight for every value of R^*. This is indeed the case: for every value of R^* there is at least a joint distribution of $\mathbf{X}$ and Y for which $R^{(1)} = R^*$, and a joint distribution of $\mathbf{X}$ and Y for which $R^{(1)} = 2R^*(1 - R^*)$.

Example



The two distributions have densities, f_0 and f_1. Each density is the mixture of two uniform distributions. The uniform distribution between -1 and 1, where the density has value $p < 1/2$, is a component of both f_0 and f_1. Assume that the prior probabilities of the two classes are equal to $1/2$. The Bayes Risk is therefore $1/2 * 2 * p = p$. If $\mathbf{X}$ falls between -3 and -2, it definitely belong to class 1. With probability 1, the training set contains at least 1 sample of class 1 between -3 and -2 (recall, the training set size is growing to infinity), hence the 1−N-N will classify the sample correctly. An analogous argument applies when $\mathbf{X}$ falls between 2 and 3. However, between -1 and 1, the conditional probability that $Y = 1$ given $\mathbf{X}$ is $1/2$, because the prior are equal, and the class-conditional densities are equal. Hence, no matter how the 1-N-N classifies a sample $\mathbf{X} \in [-1, 1]$, it is wrong with probability $1/2$, just like the Bayes Decision Rule. Hence the 1-N-N has the same conditional probability of error as the Bayes Decision Rule for every $\mathbf{X}$, and therefore its risk equals the Bayes Risk.

---

[5] $\mathbf{X}$ is distributed according to $F(X)$, Y takes M values $(0, \ldots, M-1)$, and the conditional probability that Y equals i is $\eta_i(\mathbf{X})$.

The following example shows that we do not need any smoothness assumption for the asymptotic behavior of the 1-N-N to hold. In particular, we do not even need the existence of densities!

Example (Cover and Hart)

Recall that $R^{(1)} = 0$ when $R^* = 0$. In this example we construct a distribution in (X, Y) (where X is now a scalar), that yields Bayes decision regions that are discontinuous everywhere (i.e., for every 2 points where the BDR decides 1 there is a point between them where the BDR decides 0, and vice versa).

Generate (X, Y) as follows:

- Let $\Pr\{Y = 0\} = \Pr\{Y = 1\} = 1/2$;
- conditional on $Y = 0$, $X \sim U([0, 1])$;
- conditional on $Y = 1$, generate two independent Geometric random variables[6] Z_1, Z_2. Let then $X = \min\{Z_1, Z_2\} / \max\{Z_1, Z_2\}$.

The conditional distribution of X given $Y = 1$ puts probability 1 on the rational numbers of the interval $(0, 1]$. Each rational number has positive probability: for example, the number p/q, where p and q are integers, has probability at least equal to the probability that $Z_1 = p$ and $Z_2 = q$ plus the probability that $Z_1 = q$ and $Z_2 = p$, namely $2 \cdot 2^{-p} \cdot 2^{-q}$. However, under $f_0(X)$ no point has positive probability. The Bayes Decision Rule assigns label 0 to irrational numbers and label 1 to rational numbers. The Bayes Risk is equal to zero: if an irrational number is observed, it is clearly of class 0, since class 1 does not generate irrational numbers. If the number is rational, the BDR makes a mistake if this number is of class 0. The probability that $f_0(X)$ generates a rational number is zero.

But $R^* = 0$ implies $R^{(1)} = 0$, hence, as the training set size grows to infinity, the probability of error of the 1-N-N vanishes. This means that we can decide whether a number is rational or irrational *by looking at whether its nearest neighbor is rational or irrational.* This is rather counterintuitive.

One could be *mislead* by the following argument: since all rationals have positive probability, we will see them all, hence we can classify them all correctly, because the nearest neighbor will be the point itself.
This argument has to main mistakes:
1. what we claimed is that for every ϵ there is a n_ϵ large enough, so that if the training set contains more than n_ϵ points, the risk of the $1-$N-N classifier is less than ϵ. At this point in time we have seen at least n_ϵ points, hence we still have to see an infinite number of rationals !

[6]The Geometric distribution describes the number of independent fair coin tosses up to, and including, the first heads. It is easy to show that $\Pr\{X = n\} = 2^{-n}$.

2. even if we had seen "all" rational numbers, we haven't seen all the irrationals, so the nearest neighbor of an irrational point $\mathbf{X}$ to be classified need not be the point itself.

3 $k-$Nearest Neighbor

Let's consider an extension of the 1-N-N rule. Consider the 2-class problem, let k be an *odd* number. Denote by $X_n^{(i)}$ the ith nearest neighbor of $\mathbf{X}$ in a training set of size n, and let $Y_n^{(i)}$ be the corresponding label.

$k-$N-N Rule

Decide 0 if $\sum_{i=1}^{k} Y_n^{(i)} <= k/2$

Decide 1 if $\sum_{i=1}^{k} Y_n^{(i)} >= k/2$

Hence, the k-N-N rule finds the k nearest neighbors of $\mathbf{X}$, and uses the majority vote of their labels to assign a label to X.

Since we have a majority vote, we ought to use an odd value of k. It is not particularly smart to use an even value of k in the 2-class problem. In fact, consider the possible outcomes for the 2-nearest neighbor

	$Y_n^{(1)}$	$Y_n^{(2)}$
case 1	0	0
case 2	0	1
case 3	1	0
case 4	1	1

Cases 1 and 4 are immediately clear, the rule assigns label 0 and 1 respectively, just like the 1-N-N. In cases 2 and 3 it is unclear what to do. One way to break ties is to assign the label of the nearest point, in which case our overall rule would be identical to the 1-N-N rule. Or we could break ties by always assigning label 0 (which might be a stupid thing to do), or by flipping a coin (again, this looks somewhat stupid).

We can now write the form of the asymptotic risk of the k-N-N **Theorem** *For fixed, odd, k,*

$$\lim_{n\to\infty} E\left[L_n(k-N-N)\right] = L^{(k)} =$$
$$E\left[\sum_{j=0}^{k}\binom{k}{j}\eta^j(\mathbf{X})\left(1-\eta(\mathbf{X})\right)^{k-j}\cdot \left(\eta(\mathbf{X})1_{\{j<k/2\}} + \left(1-\eta(\mathbf{X})\right)1_{\{j>k/2\}}\right)\right]$$

So, just as for the 1−N-N, the risk of the $k-$N-N asymptotically converges to a value that depends only on the distribution of $\mathbf{X}$ and on the conditional probability of Y given $\mathbf{X}$.

This expression is not particularly useful. However, with some manipulations we obtain another expression which is much easier to interpret. First, some notation: let $B(n,p)$ denote the Binomial PMF (probability mass function) with parameters n and p[7].

Theorem *For fixed, odd, k,*

$$\begin{aligned} L^{(k)} &= E\left[\eta(\mathbf{X}) Pr\left\{B(k,\eta(\mathbf{X})) < \frac{k}{2} \;\middle|\; \mathbf{X}\right\}\right] \\ &\quad + E\left[\big(1-\eta(\mathbf{X})\big) Pr\left\{B(k,\eta(\mathbf{X})) > \frac{k}{2} \;\middle|\; \mathbf{X}\right\}\right] \end{aligned} \tag{6}$$

This expression is just a different version of the first one. However, it has a nice interpretation: it emphasizes the voting process. As $n \to \infty$, all the k nearest neighbors converge to $\mathbf{X}$, and therefore the corresponding probability of having class label 1 becomes closer and closer to $\eta(\mathbf{X})$. The k-N-N rule makes a mistake when $\mathbf{X}$ is of class 1 and the nearest neighbors vote for class 0, or when $\mathbf{X}$ is of class 0 and the nearest neighbors vote for class 1. Note that the two events are disjoint, so the overall probability is the sum of the probabilities of the individual events. Hence Equation 6 has two terms. The first term has two components: the probability $\eta(\mathbf{X})$ that $\mathbf{X}$ has class 1, and the probability that the majority vote is class 0, namely, that less than $k/2$ of the k nearest neighbors are of class 1. The components of the second term are analogous: the probability $1-\eta(\mathbf{X})$ that $\mathbf{X}$ is of class 0 times the probability that more than $k/2$ of the k nearest neighbors have label 1.

Equation 6 also has a different form, which is not as useful as the previous one in terms of interpretation, but is the starting point for a collection of upper bounds.

$$\begin{aligned} L^{(k)} &= E\left[\min\left\{\eta(\mathbf{X}), 1-\eta(\mathbf{X})\right\}\right] \\ &\quad + E\Bigg[\Big(1 - 2\min\left\{\eta(\mathbf{X}), 1-\eta(\mathbf{X})\right\}\Big) \\ &\qquad\qquad \Pr\left\{B(k,\eta(\mathbf{X})) > \frac{k}{2} \;\middle|\; \mathbf{X}\right\}\Bigg] \end{aligned}$$

Evaluating Equation 6 is somewhat difficult. However, one can find upper bounds to the asymptotic probability of error. The first is not the strictest one:

[7]The binomial distribution with parameters n and p describes the number of heads in n independent tosses of a coin for which the probability of heads is p. Under the Binomial distribution, $\Pr\{X=k\} = \binom{n}{k} p^k (1-p)^{n-k}$.

Theorem *For all odd k, all distributions*

$$R^{(k)} \leq R^* + \frac{1}{\sqrt{ke}}$$

The following theorem bounds $R^{(k)}$ in terms of $R^{(1)}$.

Theorem (Györfi and Györfi, 1978) *For all odd k, all distributions,*

$$R^{(k)} \leq R^* + \sqrt{\frac{2R^{(1)}}{k}}.$$

Hence, if $R^{(1)} = 0$, $R^{(k)} = 0$ too.

The following is a stricter bound, which is universal in the sense that it holds for all distributions.

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

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

where $\gamma = \sup_{r>0} 2rPr\{N > r\} = .33994241 \simeq 1/3$, N is Normal(0,1), and the O notation refers to $k \to \infty$ (not to $n \to \infty$).

The bound says that the risk of the k-N-N classifier is to the first order equal to R^* times a factor that converges to zero as the reciprocal of the square root of the number of neighbors.

The next natural question is whether increasing the number of neighbors helps. The answer is affirmative in the asymptotic sense, as the following theorem shows:

Theorem

$$R^* \leq \ldots \leq R^{(2h+1)} \leq R^{(2h-1)} \leq \ldots R^{(3)} \leq R^{(1)} \leq 2R^*(1 - R^*)$$

Of course, an asymptotic result does not need to hold for finite sample sizes. If that were the case, we would not bother, say, with the 1−N-N classifier. We will be more specific below, but first we introduce a notion, that of admissibility, that allows us to compare rules (i.e., sequences of classifiers) *within a restricted collection.*

Def. *A rule g_n is* **Inadmissible** *within a class $\mathcal{G}$ if there exist $g_n^{'} \in \mathcal{G}$ such that $R(g_n) \geq R(g_n^{'})$ for all n, all distribution, and there exists at least one distribution and a value of n such that $R(g_n) > R(g_n^{'})$, namely, such that the inequality is strict.*
A rule is **Admissible** *if not inadmissible.*

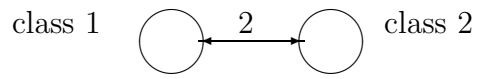
If you were wondering, no universally admissible rule exists, namely, we cannot take $\mathcal{G}$ to be the collection of all possible rules (there are, nevertheless, universally inadmissible rules, such as flipping coins)[8]

Theorem (Cover) *The 1−N-N rule is admissible among all k−N-N rules*

Proof The proof is constructive and relies on the following example:

- Let the feature space be R^3, namely, the 3−dimensional Euclidean Space.
- Let $\Pr\{Y=0\} = \Pr\{Y=1\} = 1/2$;
- Conditional on $\Pr\{Y=0\}$, $\mathbf{X}$ is distributed uniformly on the surface of a sphere of radius 1, centered at $[-200]$.
- Conditional on $\Pr\{Y=1\}$, $\mathbf{X}$ is distributed uniformly on the surface of a sphere of radius 1, centered at $[200]$.

This picture illustrates the distribution.



Consider the 1−N-N classifier. It makes a mistake when the training set contains no points of class 1 (i.e., the sum of the class labels is zero) and $\mathbf{X}$ is of class 1, or when the training set contains no points of class 0 (i.e., the product of the class labels is 1) and $\mathbf{X}$ is of class 0. The probability that the training set contains no points of class 0 is $1/2^n$, and is equal to the probability that the training set contains no points of class 1. Hence

$$\begin{aligned} R^{(1)} &= \Pr\{Y=1\}\Pr\{\text{all } Y_i = 0\} + \Pr\{Y=0\}\Pr\{\text{all } Y_i = 1\} \\ &= \frac{1}{2}\frac{1}{2^n} + \frac{1}{2}\frac{1}{2^n} = \frac{1}{2^n}. \end{aligned}$$

The 3−N-N classifier makes a mistake when $Y=0$ and training set contains zero or one point of class 0, or when the $Y=1$ and the training set contains zero or one point of class 1, namely for a collection of training sets that strictly contains those for which the 1−N-N classifier makes a mistake. Hence, the probability of error of the 3−N-N classifier is strictly greater than that of the 1−N-N for ALL training set sizes for this problem. It is also trivial to show, using the same method, that

$$R_n^{(1)} < R_n^{(3)} < R_n^{(5)} < \ldots < R_n^{(2k+1)} < \ldots$$

[8]Note that the definition of admissibility does not involve specific training sets: that is, the rule g_n could perform better than $g_n^{'}$ on specific training sets, but its risk could still be larger, when the expectation is taken over the training set.

for this particular problem.

The $k-$N-N classifier has historical importance: it gives rise to the first rule to be proven universally consistent.

Theorem If $k \to \infty$, $k/n \to 0$, the $k(n)-$N-N rule is universally consistent, i.e., its asymptotic risk is R^*.

3.1 Large (but finite) sample performance of the Nearest Neighbor Classifier

All the results of the previous section are asymptotic; it would be nice to know how fast does the $k-$N-N risk converges to its limit as the number of training samples increases. Unfortunately, no convergence guarantees exist: in particular, convergence can be arbitrarily slow, namely, for every convergence rate one can always choose a "bad" distribution for which the convergence rate is slower (sigh!).

However, if one is willing to make some assumptions on the smoothness of the distribution, then one can show meaningful results. The following result is due to Venkatesh and ??

Theorem Let the following assumptions hold (call π_0 and π_1 the priors of class 0 and 1 respectively):

H1 Let the class-conditional distributions of $\mathbf{X}$ have densities, $f_0(\cdot)$ and $f_1(\cdot)$.

H2 Let $\pi_0 f_0(\cdot) + \pi_1 f_1(\cdot)$ be bounded away from zero on its probability 1 support S.

H3 Let $f_0(\cdot)$ and $f_1(\cdot)$ have uniformly bounded derivatives up to order $N+1$.

H4 Let one or the other of $f_0(\cdot)$ and $f_1(\cdot)$ be vanishing close to the boundaries of S.

Then, for $j = 2, 3, \ldots, N$ (where $N+1$ is the number of uniformly bounded derivatives) the following result holds

$$R_n^{(k)} = R^{(k)} + \sum_{j=2}^{N} c_j n^{-j/d} + O(n^{-(N+1)/d}),$$

where the constants c_j do not depend on n or on the distributions (check!)

This theorem tells us that under the smoothness assumptions, the risk of the k-N-N classifier converges to its asymptotic value as $O(n^{-2/d})$ (although, for small values of n, the constants in the following terms of the expansion might dominate). Additionally, it shows that the convergence rate is significantly

slowed down by the dimensionality (d appears on the denominator of the exponent!), namely, that the $k-$N-N will not be a great performance in high-dimensional spaces.

4 Variations on the N-N Classifier

4.1 Weighted $k-$N-N

The $k-$N-N classifier tries to estimate the value of $\eta(\mathbf{X})$ by taking the majority vote of the labels of the nearest neighbors of $\mathbf{X}$. The implicit assumption here is that the value of $\eta(\mathbf{X}_n^{(i)})$ for $i = 1, \ldots, k$ is close to the value of $\eta(\mathbf{X})$. However, the farther the neighbor, the "less likely" is that the value of η is close to the value at $\mathbf{X}$ (this is a non-technical statement, not a theorem !

An approach to solving this problem is to assign different weights to different neighbors, namely, to weigh less the vote of farther neighbors than those of close neighbors. From the viewpoint of notation, call $w_1 \geq w_2 \ldots \geq w_k$ the weights of $\mathbf{X}_n^{(1)}, \mathbf{X}_n^{(2)}, \ldots, \mathbf{X}_n^{(k)}$.

Weighted $k-$N-N Rule

Decide 0 if $\sum_{i=1}^k Y_n^{(i)} w_i < \sum_{i=1}^k (1 - Y_n^{(i)}) w_i$

Decide 1 if $\sum_{i=1}^k Y_n^{(i)} w_i > \sum_{i=1}^k (1 - Y_n^{(i)}) w_i$

Note One must be significantly more careful with ties than in the regular nearest neighbor algorithms, but this topic beyond the scope of this course.

We ask ourselves whether using weights is indeed a sensible thing to do. Intuitively, they should not help asymptotically (although, one should be VERY careful with this line of reasoning!) because all the k neighbors are arbitrarily close to $\mathbf{X}$. In fact, asymptotically weights might hurt. The following theorem states just that.

Theorem *For* **fixed** k *odd, let* $w_i = 1/k$*; for* k *even, let* $w_1 = 1/k + \epsilon$*, all the other* $w_i = 1/k - \epsilon/(k-1)$*, call* $R^{(k)}$ *the corresponding asymptotic risk. Call* $R(w_1, \ldots, w_k)$ *the asymptotic risk of the weighted rule with weights* $w_1, \ldots, w_k$*. Then*

$$R^{(k)} \leq R(w_1, \ldots, w_k)$$

where equality holds if $Pr\{\eta(\mathbf{X}) = 1/2\} = 1$*, or if every numerical minority of the* w_i *carries less than 1/2 of the total weight.*

So: asymptotically it is NOT a good idea to use weighted k-N-N for fixed k.

However: in 1966 it was shown that, if k can vary with n, then the weighted $k-$N-N can be advantageous.

Having addressed the asymptotic situation, now we deal with the finite sample case. Here, in general, we cannot say anything meaningful unless we are willing to make severely restrictive assumptions.

Finite training set sizes are the realm of empirical investigation. For example, MacLeod, Luk, and Titterington proposed the following class of weights, which seem to work well in many practical cases

$$w_j = \frac{(d_s - d_j) + \alpha(d_s - d_1)}{(1+\alpha)(d_s - d_1)}$$

for $s > 1$, $s \geq k$. Here, $d_j = \left\| \mathbf{X} - \mathbf{X}_n^{(j)} \right\|$, $\alpha \geq 0$.

Experiments show that with appropriate s and α the weighted k-N-N can be better than the $k-$N-N.

How would one choose α and s? Again, empirically, for example using cross-validation.

5 A variation: $(k-l)$-N-N

What if are allowed to "refuse" to make a decision? Hellman (1970) proposed the $(k-l)$-N-N classifier: if at least $l > k/2$ nearest neighbors belong to the same class, classify; otherwise, refuse.

So: the classifier classifies only if it is "sufficiently" certain. Let $R_n^{(k,l)}$ be the "pseudo"-probability of error, namely the probability of misclassifying a sample that is not rejected. The following theorem describes the form of the asymptotic risk for the $(k-l)$-N-N classifier.

Theorem *Fix k, l*

$$\begin{aligned} \lim_{n\to\infty} R_n^{(k,l)} &= E\Bigg[\eta(\mathbf{X}) Pr\left\{ B(k,\eta(\mathbf{X})) \leq k-l \;\middle|\; \mathbf{X} \right\} \\ &\qquad + \big(1-\eta(\mathbf{X})\big) Pr\left\{ B(k,\eta(\mathbf{X})) \geq l \;\middle|\; \mathbf{X} \right\} \Bigg] \\ &\triangleq R^{(k,l)} \end{aligned}$$

This theorem is somewhat obscure: it is very difficult to figure out how well the classifier performs. The following theorem compares the risk of the $(k-l)$-N-N classifier, and shows that, under simple conditions on l, the pseudo-risk of the $(k-l)$-N-N classifier is smaller than the Bayes Risk.

Theorem *Fix k odd; for all distributions*

$$R^{(k,k)} \leq R^{(k,k-1)} \leq \ldots \leq R^{(k,\lceil k/2 \rceil + 1)} \leq R^* \leq R^{(k,\lceil k/2 \rceil)} = R^{(k)} \qquad (7)$$

and

$$\frac{R^{(k)} + R^{(k,\lceil k/2 \rceil + 1)}}{2} \leq R^* \leq R^{(k)} \tag{8}$$

Equation 7 shows that refusing to label samples is definitely advantageous. For example, the 7−N-N that refuses to label unless 5 of the 7 neighbors belong to the same class has lower risk than the Bayes Risk.

Equation 8 is even more interesting: you can read it as follows $R^{(k)} - R^* < R^* - R^{(k,\lceil k/2 \rceil + 1)}$, namely that the improvement that one makes by going from the BDR to the $(k - \lceil k/2 \rceil + 1)$-N-N classifier (and ignoring the cost of rejecting samples) is larger than the improvement that one makes by going from the k−-N-N classifier (which is the same as the $(k - \lceil k/2 \rceil)$-N-N classifier).

One could be displeased by the previous theorem: after all, we are comparing apples (the BDR, that does not reject samples) with oranges (the $(k - \lceil k/2 \rceil + 1)$-N-N classifier, that rejects samples and does not pay a penalty for it.) The following theorem should take care of this dissatisfaction.

Consider now the Bayes Decision Rule with Rejection:

- reject if $\max\{\eta(\mathbf{X}), 1 - \eta(\mathbf{X})\} < 1 + \lambda$;
- apply BDR otherwise.

Denote the corresponding risk by $R^*(\lambda)$, and call $A(\lambda)$ the rejection rate (i.e., the probability of rejecting a sample $\mathbf{X}$). Let $A_{k,l}$ be the rejection rate of the (k, l)-N-N.

The following theorem, valid for the M class case, compares the BDR with rejection to the $(k - \lceil k/2 \rceil + 1)$-N-N classifier when the corresponding rejection rates are the same.

Theorem [Loizu and Maybank 87]*If* λ *is such that* $A(\lambda) = A_{k,l}$, $l < k$,

$$R^{(k,l+1)} \leq R^*(\lambda) \leq R^{(k,l)}$$

It is worth noting in that, in the theorem, $A_{k,l+1} > A_{k,l} = A(\lambda)$. Therefore, the risk of the BDR with the same rejection rate as a $(k - l)$−N-N classifier is sandwiched between that of the $(k - l)$−N-N classifier and that of the $(k - l + 1)$−N-N classifier.

6 Edited Nearest Neighbor

What is wrong with the k-N-N classifier? In general, the fact that it is computationally intensive. In particular, the larger the value of k the more difficult is to find the k nearest neighbors of the point to be classified *even*

with the help of indexing structures. So, one would like to keep the value k to a minimum, ideally to 1 (there are very efficient methods for finding the single nearest neighbor of the query point).

Recall the main reason why we use k-N-N rather than $1-$N-N: for large training sets, the k nearest neighbors of $\mathbf{X}$ are close to $\mathbf{X}$ and we can approximate the sign of $\eta(\mathbf{X}) - 1/2$ by taking majority vote. In the 0-Bayes Risk case the $1-$N-N performs well for large sample sizes, and in fact there is little need to use k neighbors. However, when the Bayes Risk is greater than 0, the training set could contain points of class 1 in the region where the BDR decides 0, and vice versa.

Clearly, the $1-$N-N would benefit noticeably if the points of class 1 for which $\eta(\mathbf{x}) < 1/2$ and the points of class 0 for which $\eta(\mathbf{x}) > 1/2$ could be removed from the training set. This would be equivalent to sampling the training set from "edited" class conditional distributions such that the corresponding Bayes Risk is zero, and the corresponding decision regions coincide to the decision regions of the original problem.

Unfortunately, it is in general impossible to edit the training set without knowing the Bayes Decision Rule, in which case the use of a nearest-neighbor classifier would not be warranted anyway. There are, nevertheless, several algorithms that try to approximate the optimal editing scheme, and that in practice substantially improve the performance of the $1-$N-N classifier. These algorithms are approximate, in the sense that they are bound to incorrectly remove points for which $1_{\{\eta(\mathbf{x})>1/2\}} = Y$, and to fail to remove points for which $1_{\{\eta(\mathbf{x})>1/2\}} \neq Y$.

We now describe a few common such methods.

6.1 Generalized Editing Scheme

Let $\mathcal{R}$ be the original training set, let $g_{|\mathcal{R}|}(\mathbf{x}; \mathcal{R})$ be the classification rule[9], let ξ be an estimator of the error rate of the classification rule, and let σ be a stopping criterion that decides when the iterative algorithm described below terminates.

The generalized editing scheme consists of the following steps

1. Use the estimator ξ to estimate the error rate of the rule $g_{|\mathcal{R}|}(\mathbf{x}; \mathcal{R})$ trained with the training set $\mathcal{R}$. Call $\mathcal{E}$ the set of training samples that are misclassified during step 1.

2. Let $\mathcal{R} = \mathcal{R} \setminus \mathcal{E}$.

[9]Each classifier in the rule is a mapping from $\mathcal{X}$ to $\mathcal{Y}$, selected using the training set $\mathcal{R}$. Since we are dealing with a rule, the hypothesis space and the algorithm to select an element of the hypothesis space using the training set depends on the training set size, $|\mathcal{R}|$.

3. Evaluate the stopping rule σ, and if necessary terminate, otherwise go back to step 1.

Clearly, the generalized editing scheme is not a specific algorithm, but rather a class of algorithms, which depend on ξ and σ. Different error estimators have been used in conjunction with this scheme, including different variants of the Cross-validation estimate. It is worth noting that the actual error rate is not used explicitly, but might be used, for example, in the terminating condition: one might decide to terminate when the estimate of the error rate falls below a threshold. Other terminating conditions that could be used rely on a threshold on the fraction of training points that are eliminated, or on the desired size of the training set, etc.

6.2 Wilson Editing Algorithm

This is a very simple editing scheme.

1. Let $\mathcal{E}$ be the subset of $\mathcal{R}$ that are misclassified using the k-N-N classifier trained with $\mathcal{R} \setminus \mathcal{E}$.
2. Return $\mathcal{R} \setminus \mathcal{E}$.

This is an interesting algorithm, which is somewhat deeper than it appears at a first glance. Here one needs to find a subset $\mathcal{E}$ of $\mathcal{R}$ such that all the samples of $\mathcal{E}$ and no sample of $\mathcal{R} \setminus \mathcal{E}$ are misclassified by the $k-$N-N algorithm is trained using $\mathcal{R} \setminus \mathcal{E}$. The search for $\mathcal{E}$ is perforce iterative, and, since the samples of $\mathcal{R}$ are used alternatively in the training and edited test, the method in practice tends to be over optimistic.

6.3 The Multiedit Algorithm

The multiedit algorithm consists of the following steps

1. Let $\mathcal{E} = \emptyset$, partition randomly the training set into G groups, $\mathcal{R}_1, \ldots, \mathcal{R}_G$;
2. For $i = 1, \ldots, G$ do:
   (a) Train the $k-$N-N classifier with the samples from the group $\mathcal{R}_{(i+1) \bmod G}$;
   (b) Add to the edit set $\mathcal{E}$ the samples from $\mathcal{R}_i$ that were misclassified by the classifier trained in step (a);
3. If $\mathcal{E} = \mathcal{R}$ during the last I iteration, then terminate;
4. Else, let $\mathcal{R} = \mathcal{E}$ and go back to step 1.

It can be shown that the multiedit algorithm is optimal if applied to an infinite number of training points (I do not know if there is a universal limit theorem for this algorithm, there might be). However, in practice this method, like others using the Holdout estimate, appear to behave poorly when the training set size is finite.

6.4 Cross-Validation Multiedit Algoritm

The cross-validation multiedit algorithm is somewhere in between the previous two.

1. Let $\mathcal{E} = \emptyset$, partition randomly the training set into G groups, $\mathcal{R}_1, \ldots, \mathcal{R}_G$;
2. For $i = 1, \ldots, G$ do:
   (a) Train the $k-$N-N classifier with the samples from the group $\mathcal{R} \backslash \mathcal{R}_i$;
   (b) Add to the edit set $\mathcal{E}$ the samples from $\mathcal{R}_i$ that were misclassified by the classifier trained in step (a);
3. If $\mathcal{E} = \mathcal{R}$ during the last I iteration, then terminate;
4. Else, let $\mathcal{R} = \mathcal{E}$ and go back to step 1.