LINEAR NEAR-NEIGHBOR CLASSIFIERS FOR CORRELATED CATEGORIES

BY

JAY L. DEVORE

TECHNICAL REPORT NO. 68
JANUARY 1975

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT GP-30711X-2

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
LINEAR NEAR-NEIGHBOR CLASSIFIERS FOR CORRELATED CATEGORIES

by

JAY L. DEVORE

Oberlin College and Stanford University

TECHNICAL REPORT NO. 68
JANUARY 1975

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT GP-30711X-2

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Linear Near-Neighbor Classifiers for Correlated Categories

Jay L. Devore
Oberlin College and Stanford University

Abstract

We consider a classification problem in which $\theta = \theta_1, \theta_2, \ldots$ is a sequence of unknown categories, with each $\theta_i = 0$ or 1, and $X = X_1, X_2, \ldots$ is a sequence of observed random variables. The conditional distribution of $X_i$ given $\theta_i$ is assumed normal with variance 1 and mean $2\theta_i - 1$. We assume that the $\theta$ sequence is a realization of a two-state stationary Markov chain. Thus there is information about $\theta_i$ contained not only in $X_i$ but in all other $X_j$'s. Here we study various properties of classification rules for $\theta_i$ based on linear functions of the $X_j$'s with $|j-i|$ small. When the values of the Markovian parameters are unspecified, it is natural to consider using a plug-in rule based on substitution of parameter estimates. The large sample behavior of parameter estimates and plug-in rules is investigated, with particular attention to the probability that the plug-in rule yields a smaller risk than the usual minimax procedure appropriate for uncorrelated categories.
1. **Introduction**

In the usual formulation of the statistical classification problem, a sequence $X = X_1, X_2, \ldots$ of observed random variables is used to obtain an estimate of an unknown and unobservable sequence $\theta = \theta_1, \theta_2, \ldots$ of categories when the probability distribution of $X$ depends on $\theta$.

Here we assume that each $\theta_i = 0$ or 1 and that $X_i$ is conditionally (given $\theta_i$) independent of the other $X$'s and $\theta$'s, with the conditional distribution being normal with variance 1 and mean $2\theta_i - 1$. Thus when $\theta_i = 0$, $X_i \sim N(-1,1)$, while if $\theta_i = 1$ then $X_i \sim N(1,1)$.

In virtually all papers relating to the statistical classification problem, the successive categories ($\theta_i$'s) are assumed to be independent. The Bayes classification rule is obtained by assuming a prior probability distribution on the possible values of $\theta_i$, and taking as the Bayes estimate $\hat{\theta}_i$ that value of $\theta_i$ whose posterior probability given $X_i = x$ is the largest. For our problem, assuming that $P(\theta_i = 0) = P(\theta_i = 1) = 1/2$ yields the Bayes procedure

$$
\hat{\theta}_i = \begin{cases} 
1 & \text{if } X_i > 0 \\
0 & \text{if } X_i < 0 
\end{cases}
$$

which is also the minimax classification rule (Anderson, [1], chapter 6).
In many situations, though, it is unreasonable to assume that the successive categories are independent. For example, a production manager would want to know whether each component coming from a production line facility was defective \((\theta_i = 0)\) or free of defects \((\theta_i = 1)\), but might have at his disposal for such an inference only a test measurement on each component. One would expect the qualities of items produced at different times to have a correlation function which decreased as the difference between production times of two components increased. As another example, a geologist might be interested in ascertaining whether true rock type at a particular location was basaltic lava flow \((\theta_i = 0)\) or dry lake sediment \((\theta_i = 1)\). Because of inaccessibility, it might be more convenient to obtain measurements from an airplane rather than to take ground-based measurements yielding true rock type. A reasonable assumption is that the true rock types at locations close to one another are correlated, with the correlation decreasing as the distance between locations increased.

Here we model such situations by assuming that the sequence of actual categories is a realization of a stationary two-state Markov chain with \(P(\theta_i = 0) = P(\theta_i = 1) = 1/2\) and transition matrix

\[
P = \left( \begin{array}{cc}
\frac{1}{2} + \frac{\pi}{2} & \frac{1}{2} - \frac{\pi}{2} \\
\frac{1}{2} - \frac{\pi}{2} & \frac{1}{2} + \frac{\pi}{2}
\end{array} \right) \quad 0 \leq \pi \leq 1
\]

The parameter \(\pi\) is a measure of the degree to which like categories tend to cluster, with \(\pi = 0\) yielding the usual Bernoulli sequence of
categories and $\pi = 1$ corresponding to complete persistence of the initial state. Since the marginal probability of both states is $1/2$, the model is a generalization of that which led to the Bayes rule (1). It is easily verified that the $n$-step transition matrix $P^n$ is given by replacing $\pi$ in (2) by $\pi^n$ in all four entries, and that $\text{corr}(\theta_i, \theta_j) = \pi |i-j|.$

Because the $\theta$'s now form a correlated sequence, it follows that $\theta_i$ and $X_j$ are correlated for all $i,j$ and that inferences about $\theta_i$ should be based not only on $X_i$ but on at least some of the $X_j$'s with $j \neq i$. In section 2 we consider the simplest possible such procedure, appropriate when classifications must be made sequentially, which uses a linear combination of $X_i$ and $X_{i-1}$ to obtain $\hat{\theta}_i$. The best procedure in this class is obtained and its risk function is derived and studied. In section 3, it is assumed that the value of the parameter $\pi$ is unspecified. An estimator of $\pi$ is proposed, its asymptotic distribution derived, and the estimate is substituted for $\pi$ in the best classification rule to yield a plug-in rule. We use the asymptotic theory to determine the large sample probability that the resulting rule yields an improvement over the minimax rule (1). In section 4 we briefly consider classification in a non-sequential context in which $\hat{\theta}_i$ is based on a linear combination of $X_i$ and the two nearest neighbors $X_{i-1}$ and $X_{i+1}$.
2. One-Near-Neighbor Rules

Let the risk associated with using \( \hat{\theta}_i \) as an estimate of \( \theta_1 \) be the probability of misclassification \( P(\theta_1 \neq \hat{\theta}_i) \). Since \( \theta_1 = 0 \) or 1 we also have \( P(\theta_1 \neq \hat{\theta}_i) = E(\hat{\theta}_i - \theta_1)^2 \). The class of classification rules considered here is given by

\[
\hat{\theta}_i = \begin{cases} 
1 & \text{if } X_i + aX_{i-1} \geq k \\
0 & \text{otherwise}
\end{cases}
\] 

(3)

Among all rules in this class we derive that rule which minimizes \( P(\theta_1 \neq \hat{\theta}_i) \). We have

\[
P(\theta_1 \neq \hat{\theta}_i) = \frac{1}{2} P(X_i + aX_{i-1} \geq k \mid \theta_1 = 0) \]

\[
+ \frac{1}{2} P(X_i + aX_{i-1} < k \mid \theta_1 = 1)
\]

\[
= \frac{1}{4} \left\{ (1+\pi) P(X_i + aX_{i-1} \geq k \mid \theta_1 = 0, \theta_{i-1} = 0) \right. \\
\left. + (1-\pi) P(X_i + aX_{i-1} \geq k \mid \theta_1 = 0, \theta_{i-1} = 1) \right. \\
\left. + (1+\pi) P(X_i + aX_{i-1} < k \mid \theta_1 = 1, \theta_{i-1} = 1) \right. \\
\left. + (1-\pi) P(X_i + aX_{i-1} < k \mid \theta_1 = 1, \theta_{i-1} = 0) \right\}.
\]

But the conditional distribution of \( X_i + aX_{i-1} \) given that \( \theta_1 = 0, \theta_{i-1} = 0 \) is \( N(-(1+a),1+a^2) \), with similar conditional distributions for the other three cases. Therefore we can write
\[ P(\theta_1 \neq \hat{\theta}_1) = \frac{1}{4} \left\{ (1+\pi) \left[ 1 - \Phi \left( \frac{k+1-a}{(1+a^2)^{1/2}} \right) + \Phi \left( \frac{k-1-a}{(1+a^2)^{1/2}} \right) \right] \\
+ (1-\pi) \left[ 1 - \Phi \left( \frac{k+1-a}{(1+a^2)^{1/2}} \right) + \Phi \left( \frac{k-1-a}{(1+a^2)^{1/2}} \right) \right] \right\} \]

where \( \Phi(u) \) is the standard normal c.d.f. Minimizing (4) with respect to \( k \), it is easily verified that the unique minimizing value is \( k = 0 \), whence (4) reduces to

\[ P(\theta_1 \neq \hat{\theta}_1) = \frac{1}{2} \left\{ (1+\pi) \Phi \left( \frac{-1-a}{(1+a^2)^{1/2}} \right) + (1-\pi) \Phi \left( \frac{-1-a}{(1+a^2)^{1/2}} \right) \right\} \]

Minimizing the probability of misclassification with respect to \( a \) leads to the equation

\[ (1+\pi)(1-a) \phi \left( \frac{1+a}{(1+a^2)^{1/2}} \right) = (1-\pi)(1+a) \phi \left( \frac{1-a}{(1+a^2)^{1/2}} \right) \]

where \( \phi(u) \) is the standard normal density function, which gives implicitly the optimal value of \( a \) as a function of \( \pi \). We can solve (6) explicitly for \( \pi \) in terms of \( a \), obtaining

\[ \pi = \frac{1 - \frac{1-a}{1+a} \exp(-2a/(1+a^2))}{1 + \frac{1-a}{1+a} \exp(-2a/(1+a^2))} \]

Let the function in (7) be denoted by \( \pi = f(a) \). Then for each value of \( a \), \( f(a) \) gives the value of \( \pi \) for which that \( a \) specifies the
optimal rule within the class of rules given by (3) (with $k = 0$). Then $a = g(\pi) = f^{-1}(\pi)$ gives for each value of $\pi$ the optimal value of $a$ to be used in (3).

**Proposition 1:** The function $a = g(\pi)$, which gives the optimal value of $a$ for each $\pi$, is a strictly increasing function of $\pi$ for $0 \leq \pi \leq 1$, with $g(0) = 0$ and $g(1) = 1$.

**Proof:** The values of $g(\pi)$ for $\pi = 0$ and 1 are obtained by direct substitution. We can rewrite the function in (7) as

$$\pi = f(a) = \frac{1 - M(a)}{1 + M(a)}.$$

Then $f$ will be an increasing function of $a$ if $M(a)$ is a decreasing function of $a$. But

$$\log M = \log(1-a) - \log(1+a) - 2a/(1+a^2)$$

so that

$$\frac{d}{da} \log M(a) = -\left(\frac{1}{1-a} + \frac{1}{1+a} + \frac{2(1-a^2)}{(1+a^2)^2}\right) < 0$$

for $0 < a < 1$. That $g(\pi)$ is a decreasing function of $\pi$ now follows from the relations $g(\pi) = f^{-1}(\pi)$ and $g' = 1/f'$.

Recalling that large values of $\pi$ mean a high correlation between successive $\theta$ (and $X$) values, the proposition tells us that as $\pi$
increases we should weight the information contained in \( X_{i-1} \) more heavily in making an inference about \( \theta_i \), but that \( X_i \) contains more information about \( \theta_i \) than does \( X_{i-1} \).

**Proposition 2:** The probability of misclassification is, for fixed \( \pi > 0 \), a strictly decreasing function of \( a \) on the interval \([0, a^*] \) and a strictly increasing function of \( a \) on the interval \((a^*, 1] \), where \( a^* \) is the optimal value of \( a \) given by \( a^* = f^{-1}(\pi) \).

**Proof:**

\[
\frac{3}{2a} \mathbb{P}(\theta_i \neq \hat{\theta}_i) = \frac{2(1-\pi)(1+a)}{(1+a^2)^{3/2}} \phi \left( \frac{1-a}{(1+a^2)^{1/2}} \right) \\
- \frac{2(1+\pi)(1-a)}{(1+a^2)^{3/2}} \phi \left( \frac{1+a}{(1+a^2)^{1/2}} \right).
\]

Thus

\[
\left. \frac{3}{2a} \mathbb{P}(\theta_i \neq \hat{\theta}_i) \right|_{a=0} = -4\pi \phi(1) < 0
\]

and

\[
\left. \frac{3}{2a} \mathbb{P}(\theta_i \neq \hat{\theta}_i) \right|_{a=1} = \sqrt{2} \phi(0)(1-\pi) > 0.
\]

The proposition now follows since \( a^* \) is the unique stationary point (i.e. the unique value of \( a \) for which (8) is equal to 0).

**Corollary 1:** For \( \pi \geq \frac{\phi(-\sqrt{2}) + .5 - 2\phi(-1)}{.5 - \phi(-\sqrt{2})} \approx .620 \), any value of \( a > 0 \) will yield a smaller probability of misclassification than that attained using \( a = 0 \) (the usual minimax rule for uncorrelated categories).
Proof: Any value of $a > 0$ will yield a risk less than that for $a = 0$ when $P(\theta_i \neq \hat{\theta}_i)$, considered as a function of $a$, lies entirely below its value at $a = 0$. From proposition 2, this will be the case when $P(\theta_i \neq \hat{\theta}_i)\bigg|_{a=0} \geq P(\theta_i \neq \hat{\theta}_i)\bigg|_{a=1}$. Since

$$P(\theta_i \neq \hat{\theta}_i)\bigg|_{a=0} = \phi(-1)$$

(9)

$$P(\theta_i \neq \hat{\theta}_i)\bigg|_{a=1} = 0.5\{(1+\pi)\phi(-\sqrt{2}) + 0.5(1-\pi)\}$$

the result follows from equating the two expressions in (9) to obtain the boundary value of $\pi$.

Corollary 1 states that when the underlying correlation in the $\theta$ and $X$ processes is strong enough, any of the one-nearest-neighbor rules of (3) (for $k = 0$) with $a > 0$ is to be preferred to the rule which assumes no underlying correlation structure.

**Corollary 2:** For $\pi < .620$ there is an interval $(0,a')$ of $a$ values such that each $a$ in the interval yields a smaller value of $P(\theta_i \neq \hat{\theta}_i)$ than the value for $a = 0$.

**Proof:** This is immediate from proposition 2. The value of $a'$ is obtained by equating $\phi(-1)$ and $P(\theta_i \neq \hat{\theta}_i)$ to yield

$$\pi = \frac{\phi\left(\frac{-(1+a)}{(1+a^2)^{1/2}}\right) + \phi\left(\frac{-(1-a)}{(1+a^2)^{1/2}}\right) - 2\phi(-1)}{\phi\left(\frac{-(1+a)}{(1+a^2)^{1/2}}\right) - \phi\left(\frac{-(1-a)}{(1+a^2)^{1/2}}\right)}$$

(10)
which gives \( a' \) implicitly as a function of \( \pi \). That is, for fixed \( \pi \), (10) specifies that value \( a' \) for which any value of \( a \) in \((0,a')\) yields a better classification rule than the rule for \( a = 0 \).

Table 1 gives the values of \( a^*(\pi) \) and \( a'(\pi) \) for selected values of \( \pi \), Table 2 gives, for these same values of \( \pi \), the probability of misclassification for selected values of \( a \). From Table 2 we see that, as expected, the improvement obtained by using \( a^* \) rather than \( a = 0 \) is most significant for large values of \( \pi \).

3. Parameter Estimation and the Plug-In Rule

In the geology example mentioned in section 1, the experimenter may well have had enough experience with the two different rock types to know the values of the parameters in the conditional (normal) distributions; he may, however, have no prior information about the value of the Markovian parameter \( \pi \), so that it will have to be estimated from the data. Here we develop an estimator for \( \pi \), derive its asymptotic distribution, and use it to calculate the large sample probability that the plug-in rule improves upon the rule with \( a = 0 \). We note that the \( X_j \)'s do not form a sequence of independent random variables, so that a law of large numbers and a central limit theorem appropriate to correlated sequences must be invoked. In addition, the marginal distribution of the \( X_j \)'s does not involve \( \pi \) at all, so our estimator will involve combining adjacent \( X_j \)'s. Since the likelihood of the sample \((X_1, \ldots, X_n)\) is very complicated, making the M.L.E. equation intractable, we will use a moment estimator.
<table>
<thead>
<tr>
<th>$\pi$</th>
<th>0</th>
<th>.1</th>
<th>.2</th>
<th>.25</th>
<th>.3</th>
<th>.4</th>
<th>.5</th>
<th>.6</th>
<th>.7</th>
<th>.8</th>
<th>.9</th>
<th>.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a^*(\pi)$</td>
<td>0</td>
<td>.052</td>
<td>.102</td>
<td>.128</td>
<td>.156</td>
<td>.215</td>
<td>.282</td>
<td>.359</td>
<td>.455</td>
<td>.582</td>
<td>.758</td>
<td>.871</td>
</tr>
<tr>
<td>$a'(\pi)$</td>
<td>0</td>
<td>.103</td>
<td>.209</td>
<td>.265</td>
<td>.327</td>
<td>.467</td>
<td>.652</td>
<td>.926</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

$a^*(\pi) = \text{optimal value for } a \text{ in the one-nearest-neighbor rule}$

$a'(\pi) = \text{value of } a \text{ for which any } a \text{ in } (0, a') \text{ improves on } a = 0$
Table 2

<table>
<thead>
<tr>
<th>a</th>
<th>.1</th>
<th>.2</th>
<th>.3</th>
<th>.4</th>
<th>.5</th>
<th>.6</th>
<th>.7</th>
<th>.8</th>
<th>.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>.1</td>
<td>.1581</td>
<td>.1586</td>
<td>.1632</td>
<td>.1717</td>
<td>.1832</td>
<td>.1967</td>
<td>.2114</td>
<td>.2263</td>
<td>.2410</td>
</tr>
<tr>
<td>.2</td>
<td>.1562</td>
<td>.1562</td>
<td>.1584</td>
<td>.1644</td>
<td>.1736</td>
<td>.1849</td>
<td>.1973</td>
<td>.2103</td>
<td>.2231</td>
</tr>
<tr>
<td>.25</td>
<td>.1548</td>
<td>.1550</td>
<td>.1559</td>
<td>.1608</td>
<td>.1687</td>
<td>.1789</td>
<td>.1903</td>
<td>.2023</td>
<td>.2142</td>
</tr>
<tr>
<td>.3</td>
<td>.1531</td>
<td>.1538</td>
<td>.1535</td>
<td>.1572</td>
<td>.1640</td>
<td>.1730</td>
<td>.1833</td>
<td>.1942</td>
<td>.2052</td>
</tr>
<tr>
<td>.4</td>
<td>.1486</td>
<td>.1514</td>
<td>.1487</td>
<td>.1500</td>
<td>.1544</td>
<td>.1611</td>
<td>.1693</td>
<td>.1782</td>
<td>.1873</td>
</tr>
<tr>
<td>.5</td>
<td>.1427</td>
<td>.1490</td>
<td>.1438</td>
<td>.1427</td>
<td>.1448</td>
<td>.1492</td>
<td>.1552</td>
<td>.1621</td>
<td>.1694</td>
</tr>
<tr>
<td>.6</td>
<td>.1350</td>
<td>.1465</td>
<td>.1390</td>
<td>.1355</td>
<td>.1352</td>
<td>.1374</td>
<td>.1412</td>
<td>.1461</td>
<td>.1515</td>
</tr>
<tr>
<td>.7</td>
<td>.1253</td>
<td>.1441</td>
<td>.1342</td>
<td>.1283</td>
<td>.1256</td>
<td>.1255</td>
<td>.1272</td>
<td>.1300</td>
<td>.1336</td>
</tr>
<tr>
<td>.8</td>
<td>.1131</td>
<td>.1417</td>
<td>.1293</td>
<td>.1210</td>
<td>.1160</td>
<td>.1136</td>
<td>.1131</td>
<td>.1140</td>
<td>.1157</td>
</tr>
<tr>
<td>.9</td>
<td>.0978</td>
<td>.1393</td>
<td>.1245</td>
<td>.1138</td>
<td>.1064</td>
<td>.1017</td>
<td>.0991</td>
<td>.0979</td>
<td>.0978</td>
</tr>
</tbody>
</table>

Tabulated values of $P(\hat{\theta}_1 \neq \hat{\theta}_1)$ for selected values of $\pi$ and $a$ for the one-near-neighbor rule

(for $a = 0$, $P(\theta_1 \neq \hat{\theta}_1) = .1587$)
Proposition 3: Define new random variables by

\[ Z_i = I\{\text{sgn} \ X_i \neq \text{sgn} \ X_{i+1}\} \]

\[ = I\{X_i \geq 0, X_{i+1} < 0\} + I\{X_i < 0, X_{i+1} \geq 0\} \]

where \( I(A) = 1 \) if \( A \) occurs and 0 otherwise, and \( \text{sgn}(x) = +1 \) if \( x \geq 0 \) and \(-1 \) if \( x < 0 \). Let the estimator \( \hat{\pi} \) be given by

\[
(11) \quad \hat{\pi} = \frac{.5 - \frac{1}{n} \sum_{i=1}^{n} Z_i}{.5(\phi(1) - \phi(-1))^2} = 2.56 \frac{1}{n} \sum_{i=1}^{n} Z_i + 2.28.
\]

Then \( \hat{\pi} \) is an unbiased and consistent estimator of \( \pi \).

Proof: \( E(Z_i) = 2P(X_i > 0, X_{i+1} \leq 0) \), and this probability can be calculated by conditioning on (\( \theta_i, \theta_{i+1} \)) (as in the calculation of the probability of misclassification) to obtain

\[
E(Z_i) = .5 - .5\pi(\phi(1) - \phi(-1))^2
\]

from which the unbiasedness follows immediately. The consistency will follow from the ergodic theorem if we can show that the sequence of \( Z_j \)'s is ergodic. The ergodicity of this sequence follows from the ergodicity of the \( X_j \)'s via proposition 6.31 of Breiman [5]; since a mixing process is ergodic (Billingsley [2], page 12), the ergodicity of the \( X \) process can be demonstrated by
using the fact that the \( \theta \) process is mixing to show that the \( \chi \) process is mixing. The details are omitted.

**Proposition 4:** \( \sqrt{n} (\hat{\pi} - \pi) \xrightarrow{\mathcal{D}} N(0, \eta^2) \) where

\[
\eta^2 = 4\sigma^2/(c-d)^4, \quad c = \phi(1), \quad d = \phi(-1),
\]

and

\[
\sigma^2 = .75cd - .25 + .25(c^3 + d^3) \\
- .5\pi(c^3 - 2c^2 + 5cd - 2d^2 + d^3) \\
+ .25\pi^2(3cd + c^3 + d^3 - 3(c-d)^4) .
\]

**Proof:** Let \( V_i = Z_i - E(Z_i) \). Then, using the same line of argument as in the proof of proposition 3, it can be shown that the \( V \) process is \( \phi \)-mixing (Billingsley [3], section 20), from which it follows that

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i \xrightarrow{\mathcal{D}} N(0, \sigma^2)
\]

with

\[
\sigma^2 = E(V_i^2) + 2 \sum_{i=2}^{\infty} E(V_i V_{i-1}) .
\]
After some tedious algebra we have

\[ E(V_1^2) = \lambda (1-\lambda) \text{ where } \lambda = E(Z_i) \]
\[ E(V_1 V_2) = \frac{1}{8} (3-2\pi+3\pi^2)cd + \frac{1}{8} (1-\pi)^2 (c^3+d^3) - \lambda^2 \]
\[ E(V_1 V_i) = 0 \text{ for } i \geq 3 \]

which gives \( \sigma^2 \). If we let \( \hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} Z_i \), then

\[ \hat{\pi} = (1-2\hat{\lambda})/(c-d)^2 = h(\hat{\lambda}) \]

It then follows from a standard result in asymptotic distribution theory (c.f. Anderson [1], chapter 4) that

\[ \sqrt{n} (\hat{\pi} - \pi) = \sqrt{n} (h(\hat{\lambda}) - h(\lambda)) \xrightarrow{d} N(0, \eta^2) \]

with

\[ \eta^2 = [h'(\lambda) \sigma]^2, \quad h'(\lambda) = -2/(c-d)^2 \]

Recall now that the optimal value of \( a \) for a fixed value of \( \pi \) is given by \( a^* = g(\pi) = f^{-1}(\pi) \) where the function \( f \) is given in (7) (see also the statement and proof of proposition 1). When \( \pi \) is unknown then the optimal value of \( a \) is also unknown. In this situation we estimate \( \pi \) using (11) and obtain an estimate \( \hat{a}^* \) of the optimal value of \( a \) as \( \hat{a}^* = g(\hat{\pi}) \); we then substitute \( \hat{a}^* \) into (3) to obtain a plug-in classification rule. The next proposition details the asymptotic
behavior of \( \hat{a}^* \), while its corollary serves to compare the performance of the plug-in rule to that of the rule using \( a = 0 \).

Proposition 5: \( \sqrt{n} (\hat{a}^* - a^*) \overset{d}{\rightarrow} N(0, \tau^2) \) where \( \tau = n/f'(a) \) and \( f'(a) \) is given by

\[
f'(a) = \frac{-16a^2 \exp(-2a^2/(1+a^2))}{(1+a)^2(1+a^2)^2D^2}.
\]

where \( D \) is the denominator in (7); \( f' \) is to be evaluated at \( a^* = f^{-1}(\pi) \). The proof is immediate since \( a^* = g(\pi) \) and \( g' = 1/f' \).

Corollary 3: The large sample probability that the plug-in rule yields an improvement over the rule with \( a = 0 \) is given by

\[
P(\hat{a}^* \leq a') = \Phi\left( \frac{\sqrt{n} (a' - a^*)}{\tau} \right)
\]

where \( a' \) is that value of \( a \) for which all values of \( a \) in \((0,a']\) improve upon the rule with \( a = 0 \) and all \( a \) in \((a',1]\) give a larger value of \( P(\theta_1 \neq \theta_1) \) than the rule with \( a = 0 \). (The value of \( a' \), as a function of \( \pi \), is the solution to (10).)

Proof: The plug-in rule improves on the rule with \( a = 0 \) if and only if \( \hat{a}^* \leq a' \) (where \( a' \) is of course unknown, since \( \pi \) is unknown).

Therefore

\[
P(\text{improvement}) = P(\hat{a}^* = a') = P\left( \frac{\sqrt{n} (\hat{a}^* - a^*)}{\tau} \leq \frac{\sqrt{n} (a' - a^*)}{\tau} \right)
\]

which is the same as (12).
Table 3 gives the values of $\pi^2$ and $\tau^2$ for various values of $\pi$.

We now give several examples of the calculation of (12). For $\pi = .25$, table 1 gives $\ast = .128$ and $a' = .265$. Thus (12) becomes $\phi(.1636\sqrt{n})$. Thus for $n = 50$ we obtain a probability of .674, for $n = 100$ a probability of .739, and for $n = 900$ a probability of .972. This illustrates that for small values of $\pi$, $n$ must be rather large before the probability of improvement is near 1. But of course from table 2 we see that for small values of $\pi$ the improvement in using the optimal $a$ is not as great as when $\pi$ is large. For $\pi = .5$, (12) becomes .993 for $n = 50$ and 1 for $n = 100$, indicating a much larger probability of improvement when it counts.

4. Two-Near-Neighbor Rules

When the entire data set $X_1, \ldots, X_n$ is available for classification of each of $\theta_1, \ldots, \theta_n$, it is natural to consider using classification rules which use $X$'s occurring previous to as well as after $X_i$ in classifying $\theta_i$. The rule that we propose, which we call the two-near-neighbor rule, is the obvious analogue of (3), namely

\begin{equation}
\hat{\theta}_i = \begin{cases} 
1 & \text{if } aX_{i-1} + X_i + aX_{i+1} \geq 0 \\
0 & \text{otherwise}
\end{cases}
\end{equation}
<table>
<thead>
<tr>
<th>( \pi )</th>
<th>.1</th>
<th>.2</th>
<th>.25</th>
<th>.3</th>
<th>.4</th>
<th>.5</th>
<th>.6</th>
<th>.7</th>
<th>.8</th>
<th>.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \eta^2 )</td>
<td>.199</td>
<td>.431</td>
<td>.558</td>
<td>.694</td>
<td>.990</td>
<td>1.318</td>
<td>1.678</td>
<td>2.070</td>
<td>2.494</td>
<td>2.950</td>
</tr>
<tr>
<td>( \tau^2 )</td>
<td>21.324</td>
<td>6.806</td>
<td>4.640</td>
<td>3.285</td>
<td>1.886</td>
<td>1.149</td>
<td>.716</td>
<td>.418</td>
<td>.209</td>
<td>.068</td>
</tr>
</tbody>
</table>

Table 3

Asymptotic variance of \( \hat{\pi} = \eta^2 / n \)

Asymptotic variance of \( a^* = \tau^2 / n \)
To compute \( P(\theta_i \neq \hat{\theta}_i) \) we must condition on \((X_{i-1}, X_i, X_{i+1})\), obtaining

\[
P(\theta_i \neq \hat{\theta}_i) = 0.25(1+\pi)^2 \frac{1 - 2a}{(1+2a^2)^{1/2}} + 0.25(1-\pi)^2 \frac{1 + 2a}{(1+2a^2)^{1/2}} + 0.5(1-\pi^2) \Phi \left( \frac{-1}{(1+2a^2)^{1/2}} \right).
\]

To consider a rule based on \( x_j \)'s other than \((X_{i-1}, X_i, X_{i+1})\) would lead to more tedious algebra in computing (14), since it would involve conditioning on all possible values of at least five random variables (for four-near-neighbor rules). Since most of the information about \( \theta_i \) is contained in \((X_{i-1}, X_i, X_{i+1})\) unless \( \pi \) is quite large, the extra computation seems unjustified. In addition, the implicit functional relationship between \( \pi \) and \( a^* \) will involve at least a fourth degree polynomial in \( \pi \), which presents obvious difficulties.

If we differentiate (14) with respect to \( a \), we find that the optimal value of \( a \) is given implicitly by

\[
\pi = \frac{u + v - 2(uv + w)^2}{-u + v - 2w}^{1/2}
\]

where

\[
u = (1-a) \phi \left( \frac{1 + 2a}{(1+2a^2)^{1/2}} \right), \quad v = (1+a) \phi \left( \frac{1 - 2a}{(1+2a^2)^{1/2}} \right),
\]

and

\[
w = a \phi \left( \frac{1}{(1+2a^2)^{1/2}} \right).
\]
Similarly, we can obtain the value of \( a' \) from

\[
\pi = \frac{-B - (B^2 - 4AC)}{2A}
\]

where

\[
A = .25\phi_1 + .25\phi_2 - .5\phi_3,
\]

\[
B = .25\phi_1 - .25\phi_2,
\]

\[
C = .25\phi_1 + .25\phi_2 + .5\phi_3 - \phi(-1),
\]

and

\[
\phi_1 = \Phi\left(\frac{-1 - 2a}{(1 + 2a^2)^{1/2}}\right),
\]

\[
\phi_2 = \Phi\left(\frac{-1 + 2a}{(1 + 2a^2)^{1/2}}\right),
\]

\[
\phi_3 = \Phi\left(\frac{1}{(1 + 2a^2)^{1/2}}\right).
\]

The smallest value of \( \pi \) for which any \( a > 0 \) improves on the rule using \( a = 0 \) is \( \pi = .553 \). Tables 4 and 5 summarize the relevant information for this two-near-neighbor rule.
Table 4

<table>
<thead>
<tr>
<th>$\pi$</th>
<th>0</th>
<th>.1</th>
<th>.2</th>
<th>.25</th>
<th>.3</th>
<th>.4</th>
<th>.5</th>
<th>.6</th>
<th>.7</th>
<th>.8</th>
<th>.9</th>
<th>.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a^*(\pi)$</td>
<td>0</td>
<td>.052</td>
<td>.102</td>
<td>.128</td>
<td>.156</td>
<td>.215</td>
<td>.280</td>
<td>.355</td>
<td>.445</td>
<td>.560</td>
<td>.726</td>
<td>.840</td>
</tr>
<tr>
<td>$a'(\pi)$</td>
<td>0</td>
<td>.108</td>
<td>.213</td>
<td>.274</td>
<td>.344</td>
<td>.518</td>
<td>.782</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

$a^*(\pi)$ = optimal value for $a$ in the two-near-neighbor rule

$a'(\pi)$ = value of $a$ for which any $a$ in $(0,a')$ improves on $a = 0$
<table>
<thead>
<tr>
<th>a</th>
<th>(a^*)</th>
<th>.1</th>
<th>.2</th>
<th>.3</th>
<th>.4</th>
<th>.5</th>
<th>.6</th>
<th>.7</th>
<th>.8</th>
<th>.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>.1</td>
<td>.1575</td>
<td>.1586</td>
<td>.1671</td>
<td>.1819</td>
<td>.2000</td>
<td>.2193</td>
<td>.2380</td>
<td>.2554</td>
<td>.2711</td>
<td>.2851</td>
</tr>
<tr>
<td>.2</td>
<td>.1538</td>
<td>.1538</td>
<td>.1578</td>
<td>.1680</td>
<td>.1822</td>
<td>.1979</td>
<td>.2135</td>
<td>.2283</td>
<td>.2419</td>
<td>.2542</td>
</tr>
<tr>
<td>.3</td>
<td>.1478</td>
<td>.1491</td>
<td>.1485</td>
<td>.1547</td>
<td>.1649</td>
<td>.1771</td>
<td>.1898</td>
<td>.2022</td>
<td>.2136</td>
<td>.2242</td>
</tr>
<tr>
<td>.4</td>
<td>.1394</td>
<td>.1444</td>
<td>.1395</td>
<td>.1416</td>
<td>.1482</td>
<td>.1571</td>
<td>.1670</td>
<td>.1769</td>
<td>.1864</td>
<td>.1953</td>
</tr>
<tr>
<td>(\pi)</td>
<td>.5</td>
<td>.1288</td>
<td>.1398</td>
<td>.1306</td>
<td>.1289</td>
<td>.1319</td>
<td>.1377</td>
<td>.1449</td>
<td>.1525</td>
<td>.1601</td>
</tr>
<tr>
<td>.6</td>
<td>.1158</td>
<td>.1352</td>
<td>.1220</td>
<td>.1165</td>
<td>.1162</td>
<td>.1191</td>
<td>.1237</td>
<td>.1290</td>
<td>.1347</td>
<td>.1402</td>
</tr>
<tr>
<td>.7</td>
<td>.1007</td>
<td>.1307</td>
<td>.1135</td>
<td>.1044</td>
<td>.1010</td>
<td>.1011</td>
<td>.1032</td>
<td>.1064</td>
<td>.1102</td>
<td>.1142</td>
</tr>
<tr>
<td>.8</td>
<td>.0834</td>
<td>.1262</td>
<td>.1051</td>
<td>.0927</td>
<td>.0863</td>
<td>.0838</td>
<td>.0836</td>
<td>.0848</td>
<td>.0867</td>
<td>.0891</td>
</tr>
<tr>
<td>.9</td>
<td>.0639</td>
<td>.1218</td>
<td>.0970</td>
<td>.0814</td>
<td>.0722</td>
<td>.0672</td>
<td>.0647</td>
<td>.0640</td>
<td>.0650</td>
<td>.0655</td>
</tr>
</tbody>
</table>

Tabulated values of \(P(\theta_i \neq \hat{\theta}_i)\) for selected values of \(\pi\) and \(a\) for the two-near-neighbor rule

(for \(a = 0, \quad P(\theta_i \neq \hat{\theta}_i) = .1587)\)
REFERENCES


