パターン認識
Pattern Recognition

佐藤真一
Shin'ichi Satoh

国立情報学研究所
National Institute of Informatics

May 9, 2023

k-Nearest-Neighbor Classifier

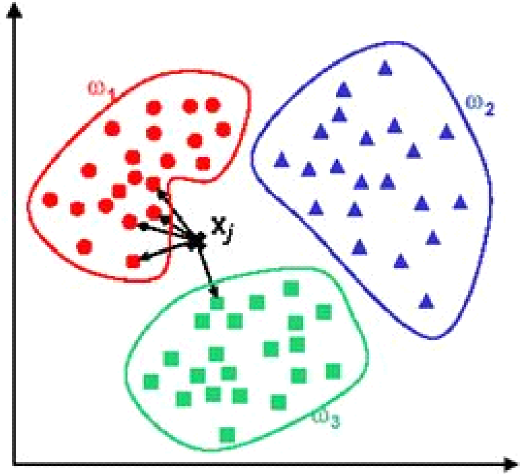
Suppose that we have c classes and that class ω_i contains n_i points in the training data with $n_1 + n_2 + \cdots + n_c = n$

$$P(\omega_i|\mathbf{x}) = \frac{p_n(\mathbf{x}|\omega_i)P(\omega_i)}{p_n(\mathbf{x})}.$$

Given a point x, we find the k_n nearest neighbors.
Suppose that k_i points from k_n belong to class ω_i then

$$p_n(x|\omega_i) = \frac{k_i}{n_i V_n}.$$

k-Nearest-Neighbor Classifier



k-Nearest-Neighbor Classifier

The prior probabilities can be computed as:

$$P(\omega_i) = \frac{n_i}{n}.$$

Using the Bayes' rule, the posterior probabilities can be computed as follows:

$$\begin{aligned} P(\omega_i|x) &= \frac{p_n(x|\omega_i)P(\omega_i)}{p_n(x)} = \frac{\frac{k_i}{n_i V_n}\frac{n_i}{n}}{\frac{k_n}{nV_n}} \\ &= \frac{k_i}{k_n} \end{aligned}$$

where $p_n(x) = \dfrac{k_n}{nV_n}$ is used.

k-Nearest-Neighbor Classifier

k-nearest-neighbor classification rule:

> Given a data point x, find a hypershpere around it that contains k points and assign x to the class having the largest number of representatives inside the hypersphere.

$$P(\omega_i|x) = \frac{p_n(x|\omega_i)P(\omega_i)}{p_n(x)} = \frac{k_i}{k_n}$$

When $k = 1$ we get the nearest-neighbor rule.

Error Rate for the Nearest-Neighbor Rule

We want to show that

$$P^* \le P \le P^*(2 - \frac{c}{c-1}P^*) \le 2P^*.$$

where P is the error rate based on the nearest-neighbor rule with infinitely many samples, and P^* is the minimum possible error rate (Bayes error rate).

Error Rate for the Nearest-Neighbor Rule

A set of n labeled prototypes: $\mathcal{D}^n = \{x_1, \cdots, x_n\}$.
The prototype nearest to a test point x: $x' \in \mathcal{D}^n$.
The nearest-neighbor rule: classifying x to the label associated with x'.
Random variable denoting the label of x': θ'.
The probability that $\theta' = \omega_i$: the *a posteriori* probability $P(\omega_i|x')$.
When the number of samples is very large, it is reasonable to assume that x' is sufficiently close to x that $P(\omega_i|x') \simeq P(\omega_i|x)$.
Thus the nearest-neighbor rule is effectively matching probabilities with nature.

Error Rate for the Nearest-Neighbor Rule

We define:

$$P(\omega_m|x) = max_i P(\omega_i|x).$$

The Bayes decision rule always selects ω_m.
Defining the infinite-sample conditional average probability of error $P(e|x)$ and the unconditional average probability of error $P(e)$, we have:

$$P(e) = \int P(e|x)p(x)dx.$$

If we further let $P^*(e|x)$ be the minuimum possible value of $P(e|x)$ and P^* be the minimum possible value of $P(e)$, then

$$P^*(e|x) = 1 - P(\omega_m|x) \text{ and } P^* = \int P^*(e|x)p(x)dx.$$

Error Rate for the Nearest-Neighbor Rule

Assume $P_n(e)$ is the n-sample error rate, and if

$$P = \lim_{n\to\infty} P_n(e)$$

then we want to show that

$$P^* \leq P \leq P^*(2 - \frac{c}{c-1}P^*).$$

Given x' as the nearest-neighbor of x, we have

$$P(e|x) = \int P(e|x,x')p(x'|x)dx'.$$

It is very difficult to obtain the conditional density $p(x'|x)$.

Error Rate for the Nearest-Neighbor Rule

However, because x' is the nearest neighbor of x, we expect $p(x'|x)$ to approach a delta function centered at x.
Consider the probability that any sample falls with in a hyper sphere $\mathcal{S}$ centered about x

$$P_{\mathcal{S}} = \int_{x' \in \mathcal{S}} p(x') dx'.$$

The probability that all n samples fall outside $\mathcal{S}$ is $(1 - P_{\mathcal{S}})^n$ which approaches zero as n goes to infinity.
Thus x' converges to x in probability, and $p(x'|x)$ approaches a delta function, as expected.

Error Rate for the Nearest-Neighbor Rule

We now turn to the calculation of the conditional probability of error $P_n(e|x, x')$.

- x'_n: the nearest neighbor of x with the number of samples n.
- n independently drawn labeled samples $(x_1, \theta_1), \ldots, (x_n, \theta_n)$

We assume that these pairs were generated by

1. selecting a state ot nature ω_j for θ_j with probability $P(\omega_j)$,
2. then selecting an x_j according to the probability law $p(x|\omega_j)$,

with each pair selected independently.

Error Rate for the Nearest-Neighbor Rule

Suppose that during classification, nature selects a pair (x, θ) and also suppose that x'_n labeled θ'_n is the training sample nearest x.
Because the state of nature when x'_n was drawn is independent of the state of nature when x is drawn, we have

$$P(\theta, \theta'_n | x, x'_n) = P(\theta|x)P(\theta'_n|x'_n).$$

Then the conditional probability of error

$$\begin{aligned} P(e|x, x'_n) &= 1 - \sum_{i=1}^{c} P(\theta = \omega_i, \theta'_n = \omega_i | x, x'_n) \\ &= 1 - \sum_{i=1}^{c} P(\omega_i|x)P(\omega_i|x'_n). \end{aligned}$$

Error Rate for the Nearest-Neighbor Rule

Considering

$$P(e|x) = \int P(e|x, x')p(x'|x)dx'$$
$$p(x'|x) = \delta(x' - x)$$

we have:

$$\lim_{n\to\infty} P_n(e|x) = \int [1 - \sum_{i=1}^{c} P(\omega_i|x)P(\omega_i|x'_n)]\delta(x'_n - x)dx'_n$$
$$= 1 - \sum_{i=1}^{c} P^2(\omega_i|x).$$

Error Rate for the Nearest-Neighbor Rule

The asymptotic nearest-neighbor eror rate is given by

$$\begin{aligned} P &= \lim_{n\to\infty} P_n(e) \\ &= \lim_{n\to\infty} \int P_n(e|x)p(x)dx \\ &= \int [1 - \sum_{i=1}^{c} P^2(\omega_i|x)]p(x)dx. \end{aligned}$$

Error Rate for the Nearest-Neighbor Rule

Recall

$$P^*(e|x) = 1 - P(\omega_m|x).$$

We want to know how small $\sum_{i=1}^{c} P^2(\omega_i|x)$ can be for a given $P(\omega_m|x)$, i.e., a given P^*.

Error Rate for the Nearest-Neighbor Rule

We write

$$\sum_{i=1}^{c} P^2(\omega_i|x) = P^2(\omega_m|x) + \sum_{i \neq m} P^2(\omega_i|x).$$

We want to minimize this subject to:

- $P(\omega_i|x) \geq 0$
- $\sum_{i \neq m} P(\omega_i|x) = 1 - P(\omega_m|x) = P^*(e|x)$.

Error Rate for the Nearest-Neighbor Rule

We can minimize $\sum_{i=1}^{c} P^2(\omega_i|x)$ if all of the *a posteriori* probabilities except $P(\omega_m|x)$ are equal.

$$P(\omega_i|x) = \begin{cases} \dfrac{P^*(e|x)}{c-1} & i \neq m \\ 1 - P^*(e|x) & i = m \end{cases}$$

Thus

$$\sum_{i=1}^{c} P^2(\omega_i|x) \geq (1 - P^*(e|x))^2 + \frac{(P^*(e|x))^2}{c-1}$$

$$1 - \sum_{i=1}^{c} P^2(\omega_i|x) \leq 2P^*(e|x) - \frac{c}{c-1}(P^*(e|x))^2.$$

This immediately shows that $P \leq 2P^*$.

Error Rate for the Nearest-Neighbor Rule

To seek for a tighter bound:

$$Var[P^*(e|x)] = \int [P^*(e|x) - P^*]^2 p(x)dx$$
$$= \int (P^*(e|x))^2 p(x)dx - P^{*2} \geq 0$$

so that

$$\int (P^*(e|x))^2 p(x)dx \geq (P^*)^2$$

with equality holding if and only if the variance of $P^*(e|x)$ is zero.

Error Rate for the Nearest-Neighbor Rule

Then...

$$P^* \leq P \leq P^*(2 - \frac{c}{c-1}P^*).$$

Error Rate for the Nearest-Neighbor Rule

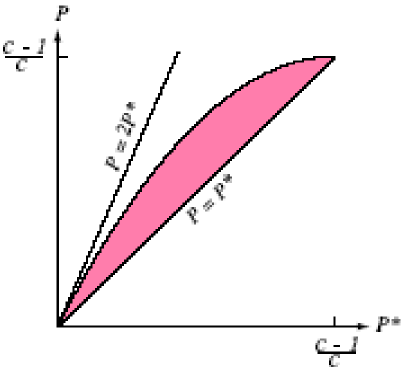


FIGURE 4.14. Bounds on the nearest-neighbor error rate P in a c-category problem given infinite training data, where P^* is the Bayes error (Eq. 52). At low error rates, the nearest-neighbor error rate is bounded above by twice the Bayes rate. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Error Rate for the Nearest-Neighbor Rule

- How well the nearest-neighbor rule works in the finite-sample case?
- How rapidly the performance converges to the asymptotic value?
- The convergence can be arbitrarily slow, and the error rate $P_n(e)$ need not even decrease monotonically with n.
- It is difficult to obtain anything other than asymptotic results without making further assumptions about the underliying probability structure.

The k-Nearest-Neighbor Rule

We can make decision by examining the labels on the k nearest neighbors and taking a vote. We can consider two-class case

- k odd: avoiding ties
- k even: reject ties

The error of k-NN P_{kNN} yields:

$$\frac{1}{2}P^* \leq P_{2NN} \leq P_{4NN} \leq \cdots \leq P^* \leq \cdots \leq P_{3NN} \leq P_{NN} \leq 2P^*.$$

If you are interested, see chapter 7 of Fukunaga "Statistical Pattern Recognition" for the detailed derivation.

The k-Nearest-Neighbor Rule

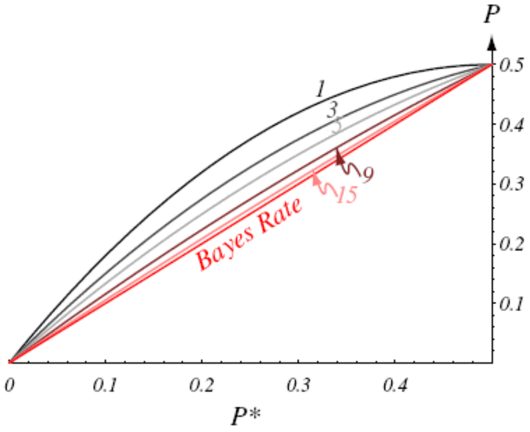


FIGURE 4.16. The error rate for the k-nearest-neighbor rule for a two-category problem is bounded by $C_k(P^*)$ in Eq. 54. Each curve is labeled by k; when $k = \infty$, the estimated probabilities match the true probabilities and thus the error rate is equal to the Bayes

Exercise

Two normal distributions

Class 1 $\Sigma_1 = I,\ \mu_1 = [s, 0, \cdots, 0]^T$

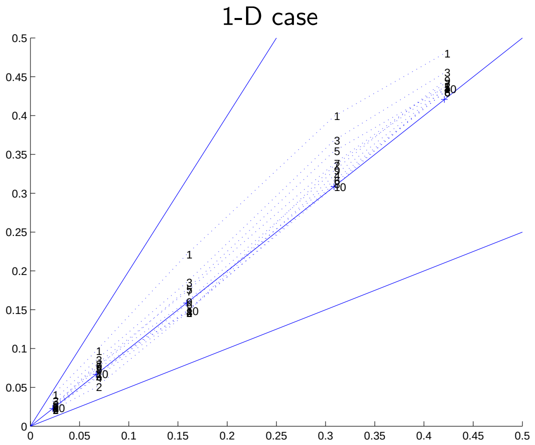
Class 2 $\Sigma_2 = I,\ \mu_2 = [-s, 0, \cdots, 0]^T$

Bayes error:

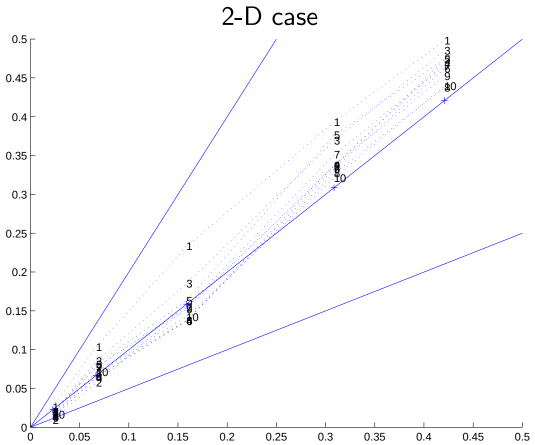
$$P^* = \int_s^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} N(0, I) dx_1 \cdots dx_n$$
$$= \frac{1}{2}(1 - \mathrm{erf}(\frac{s}{\sqrt{2}}))$$
$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

Given n class 1 and n class 2 data, observe the error rate by using k-NN classifier

Exercise (knnclass.py)

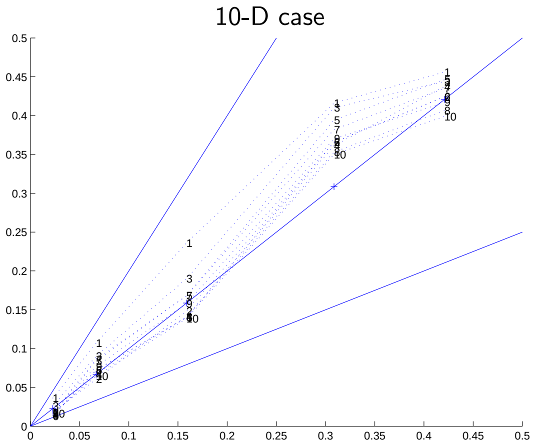


Exercise

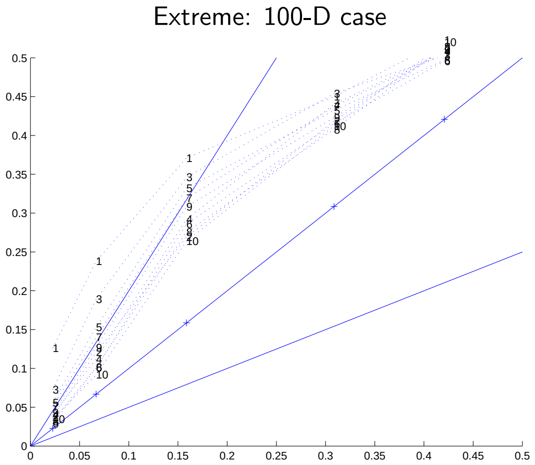


2-D case
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5

Exercise


10-D case
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5

Exercise



Exercise

Try k-NN classifier for MNIST data.
Very simple modification to mntest.py should work.
Strongly suggested (but not assignment).

My simple implementation of 1-NN classifier achieves 97% accuracy.
How about k-NN classifier?