NEAREST NEIGHBORS CLASSIFICATION

Consider the binary classification problem: We are given training data

$$(x_1, y_1), \ldots, (x_n, y_n)$$

where

$$x_i \in \mathbb{R}^d$$

$$y_i \in \{-1, +1\}.$$

We assume that (x_i, y_i) are realizations of a random pair (X, Y).

Now suppose we are given an unlabeled point we want to classify, call it x.

Nearest Neighbor Classifier

The nearest neighbor (NN) classifier is easiest to state in words:

> Assign to x the same label as the closest training point x_i to x.

The NN rule defines a <u>partition</u> of the feature space:

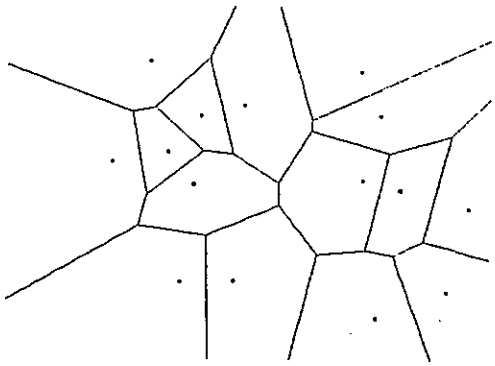
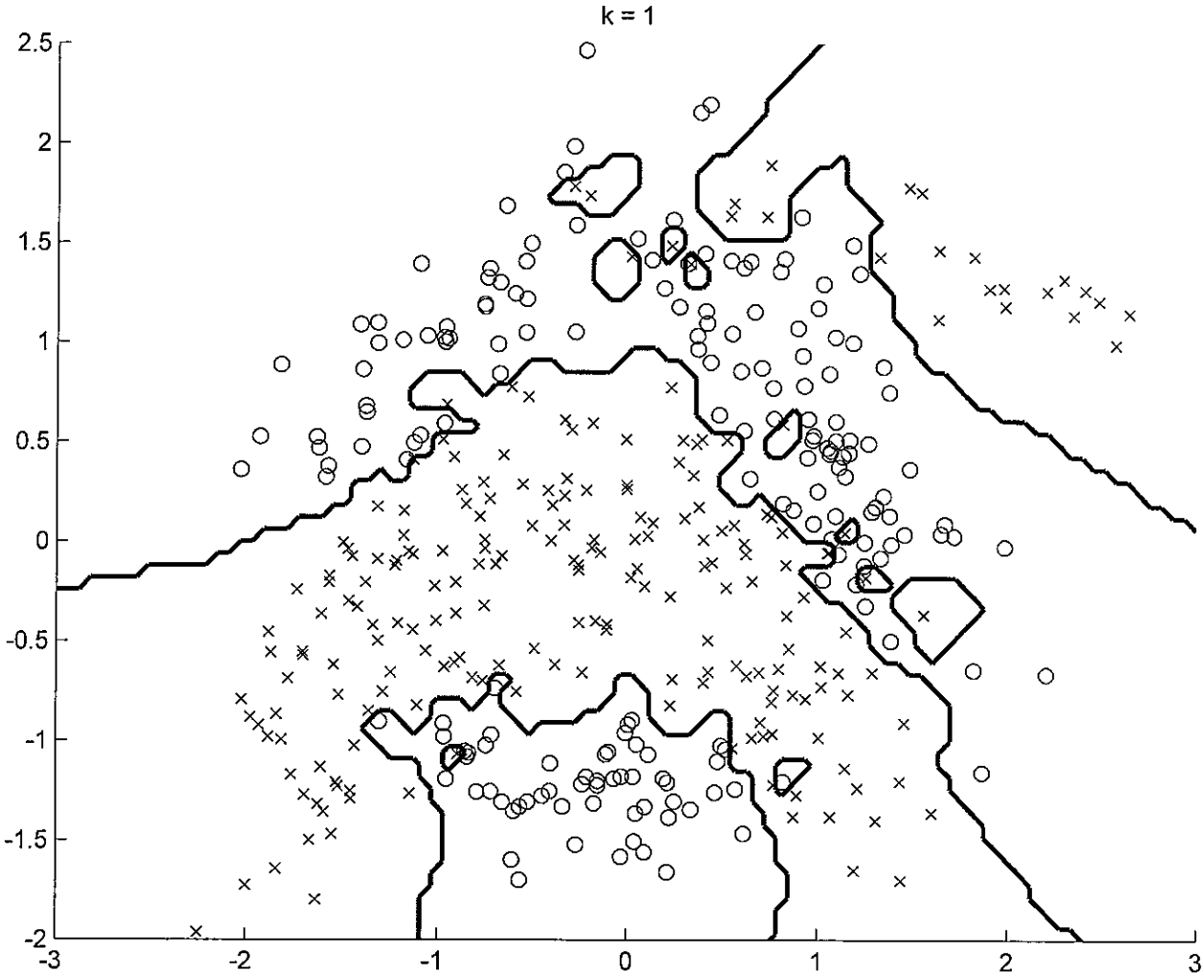


FIGURE 5.1. *At every point the decision is the label of the closest data point. The set of points whose nearest neighbor is X_i is called the Voronoi cell of X_i. The partition induced by the Voronoi cells is a Voronoi partition. A Voronoi partition of 15 random points is shown here.*

Figure from Devroye, Györfi & Lugosi, A Probabilistic Theory of Pattern Recognition.



Is the NN classifier

- generative or discriminative?
- linear or nonlinear?
- parametric or nonparametric?

(A)

k - Nearest Neighbors

For odd $k \geq 1$, the k-nearest neighbors (kNN) rule generalizes the NN rule:

> Assign a label to x by taking a majority vote over the k training points x_i closest to x

Mathematically, define

$$I_k(x) := \left\{ \begin{array}{l} \text{indices } 1 \leq i \leq n \text{ of the } k \text{ training} \\ \text{points closest to } x_i \end{array} \right\}$$

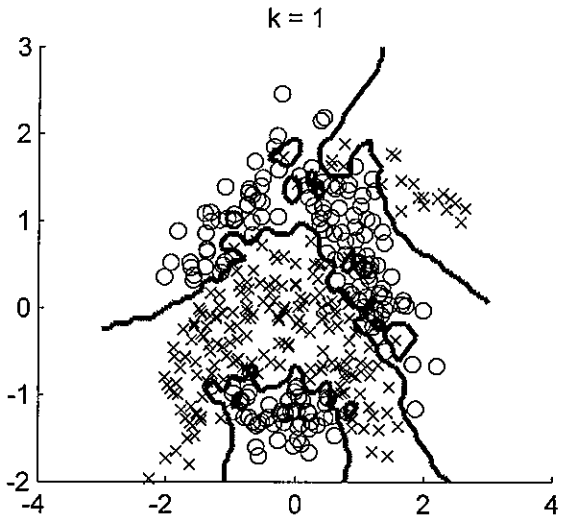
and

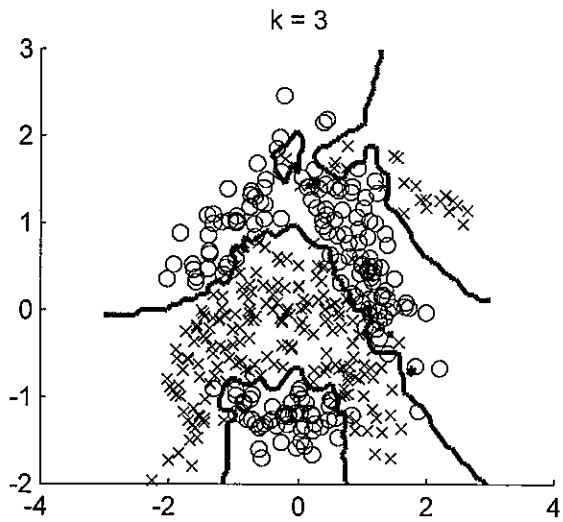
$$\text{sign}(t) = \begin{cases} 1 & \text{if } t \geq 0 \\ -1 & \text{if } t < 0 \end{cases}$$

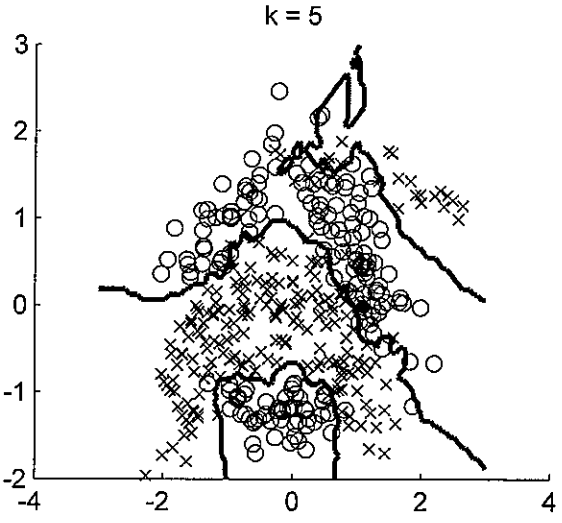
Then the kNN rule is

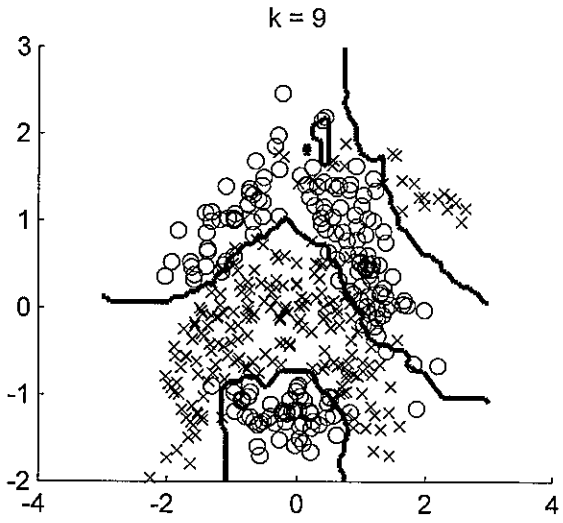
$$f_k(x) :=$$

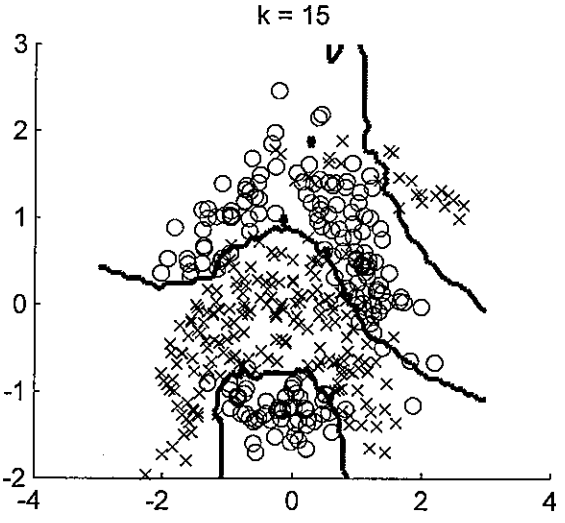
(B)

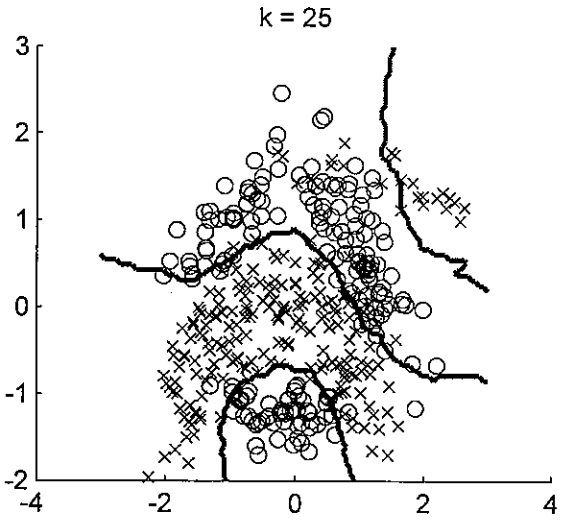

k = 1
3
2
1
0
-1
-2
-4
-2
0
2
4


k = 3
3
2
1
0
-1
-2
-4
-2
0
2
4


k = 5
3
2
1
0
-1
-2
-4
-2
0
2
4


k = 9
3
2
1
0
-1
-2
-4
-2
0
2
4


k = 15
3
2
1
0
-1
-2
-4
-2
0
2
4


k = 25
3
2
1
0
-1
-2
-4
-2
0
2
4

Which value of k should we use?

Evaluation

In theory, we would like to choose k to minimize the probability of error,

$$R(f_k) := \text{Prob}\{f_k(X) \neq Y\}.$$

However, we don't have knowledge of the true distribution governing (X, Y).

We could look at the training error

© $$\hat{R}_{\text{train}}(f_k) =$$

However, this is a biased estimate of $R(f_k)$.

Exercise

What is $\hat{R}_{\text{train}}(f_1)$?

Solution | $\hat{R}_{train}(f_1) = 0$. This estimate thinks the chances of f_1 making a mistake are zero.

For now, assume we have access to a set of test data

$$(x_{n+1}, y_{n+1}), \ldots, (x_{n+m}, y_{n+m}),$$

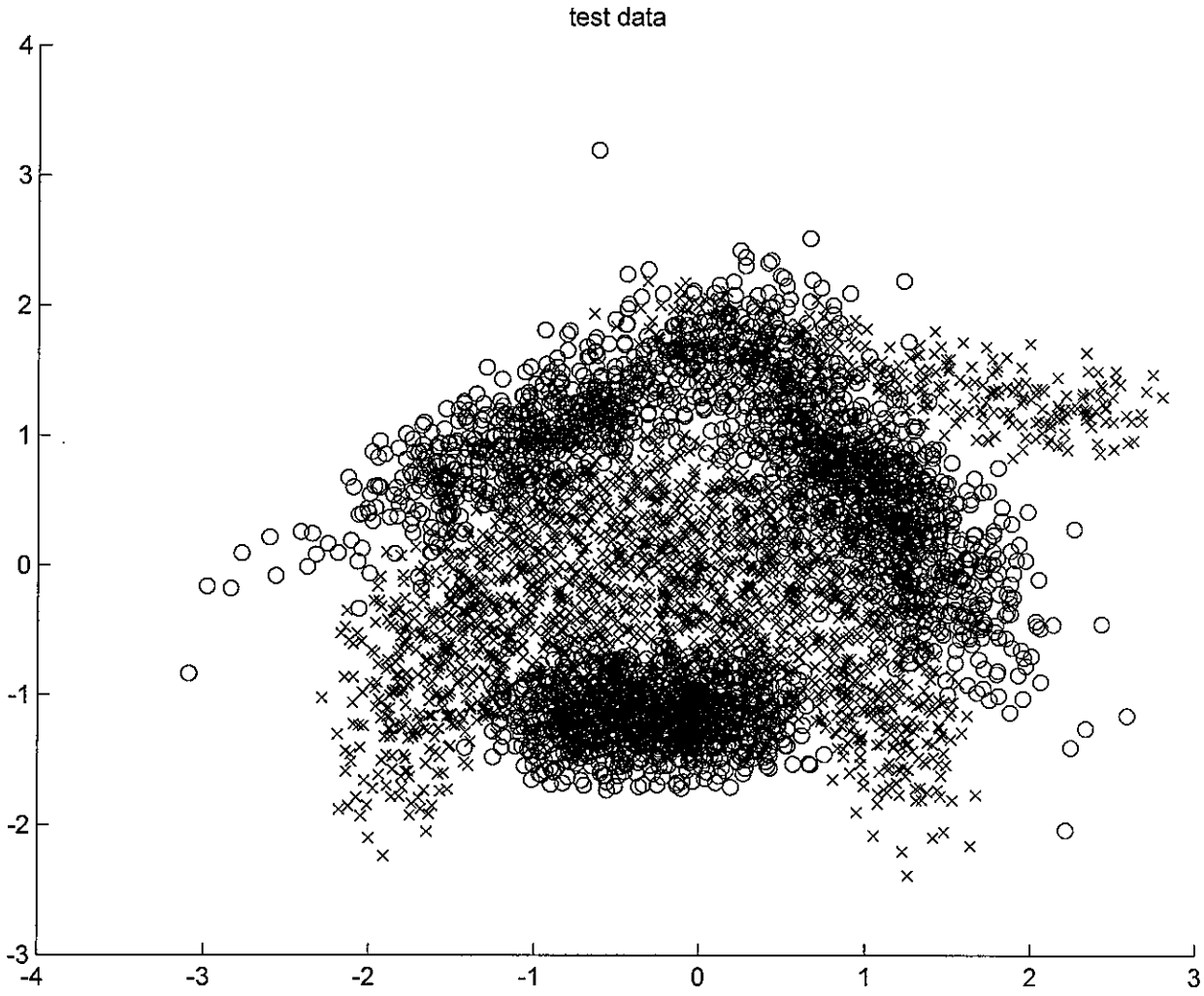
realizations of the same (X, Y) giving rise to the training data.

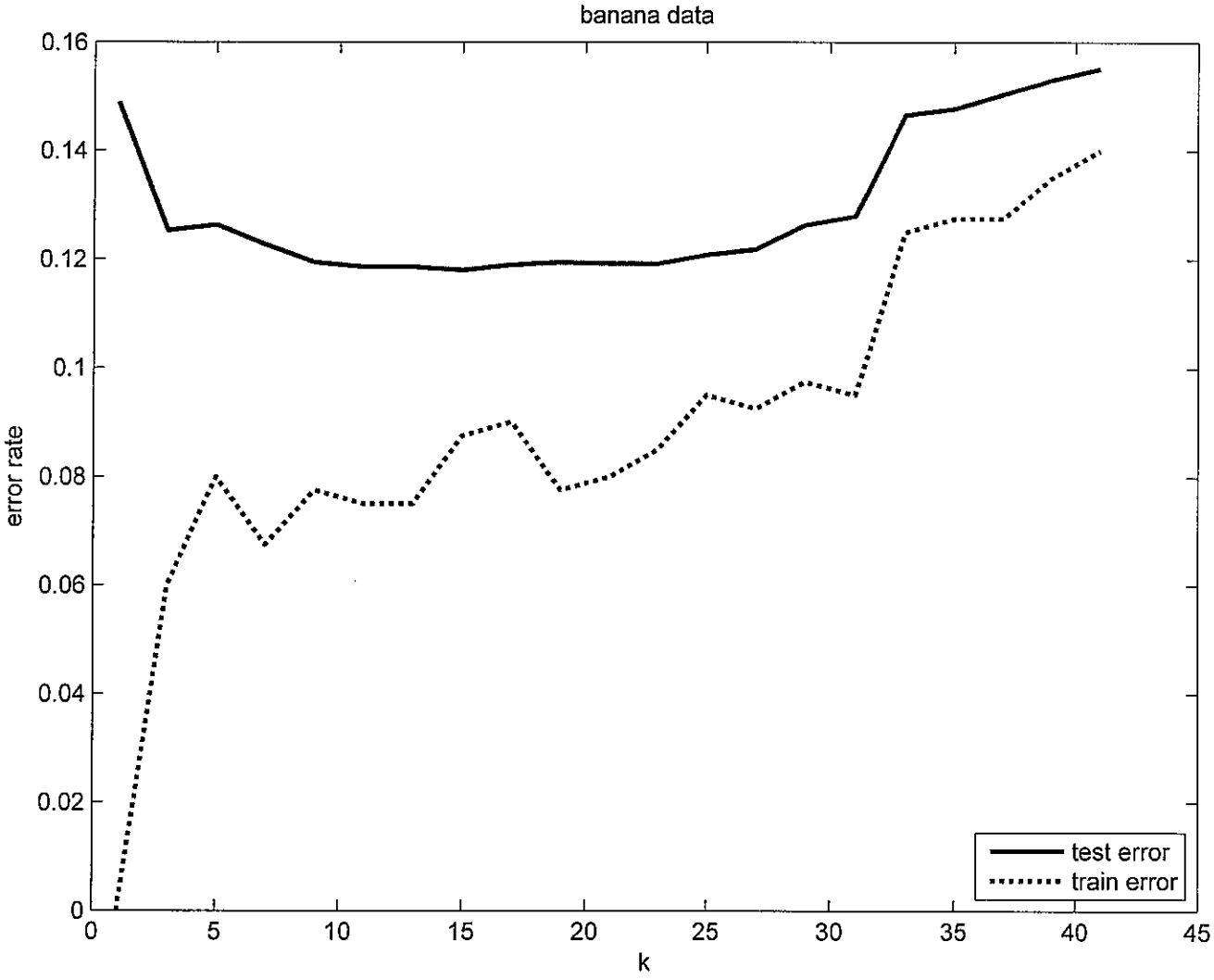
The test error is defined to be

$$\hat{R}_{test}(f_k) =$$

(D)

In most practical situations, we don't have access to test data -- if we did we'd want to treat it as training data. We'll return to this issue later in the course.


test data
4
3
2
1
0
-1
-2
-3
-4
-3
-2
-1
0
1
2
3


banana data
error rate
k
test error
train error
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
5
10
15
20
25
30
35
40
45

Theory

The **Bayes error** is

$$R^* := \min_{\{\text{all } f\}} R(f)$$

and is the best possible performance of a classifier for a given distribution.

Denote the kNN rule by $f_{k,n}$ to reflect its dependence on the sample size.

Theorem If $n \to \infty$, $k \to \infty$, and $k/n \to 0$, then $E\{R(f_{k,n})\} \to R^*$, i.e. $f_{k,n}$ is **consistent**.

↖ Proof: Devroye, Györfi, & Lugosi (1996)

Key

A. the latter in each case

B. $\text{sign}\left\{\sum_{i \in I_k(x)} y_i\right\}$

C. $\frac{1}{n}\sum_{i=1}^{n} 1_{\{f_k(x_i) \neq y_i\}}$

D. $\frac{1}{m}\sum_{i=1}^{m} 1_{\{f_k(x_{n+i}) \neq y_{n+i}\}}$