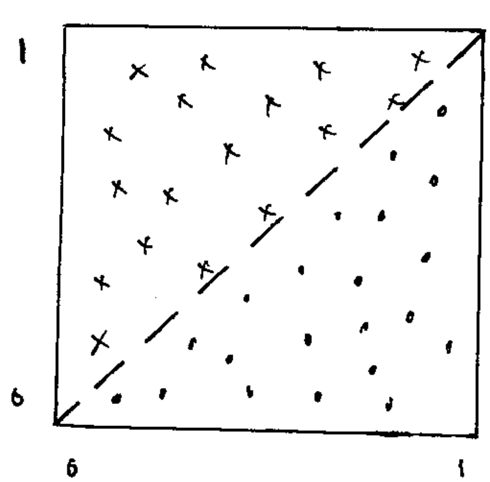
ENSEMBLE METHODS

The idea behind ensemble methods is to generate several classifiers $f_1, \dots, f_T$ using a variety of methods, and to combine them into a single classifier that performs better than any individual classifier.

Let's look at an ...

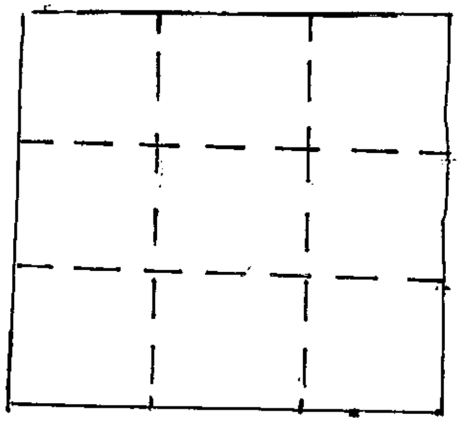
Example | Averaged Shifted Histograms

Suppose we observe two dimensional data, $x_i \in [0,1]^2$.



Bayes error $= 0$

A very basic (and not recommended!) classifier is a <u>histogram</u> rule:

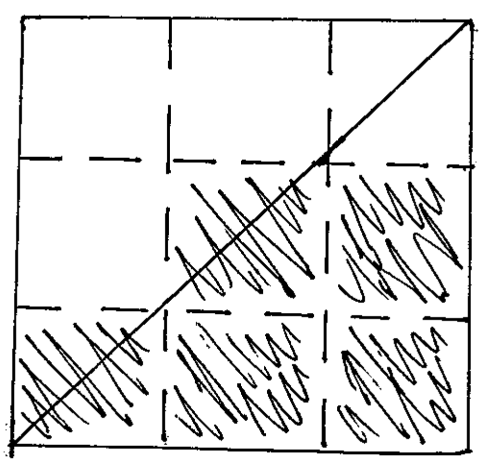


- assign the same label to patterns x in the same _________.

Ⓐ

- determine label by ______________ ________.

As you can imagine, this classifier will not perform very well.



Let's generate a whole bunch of equally

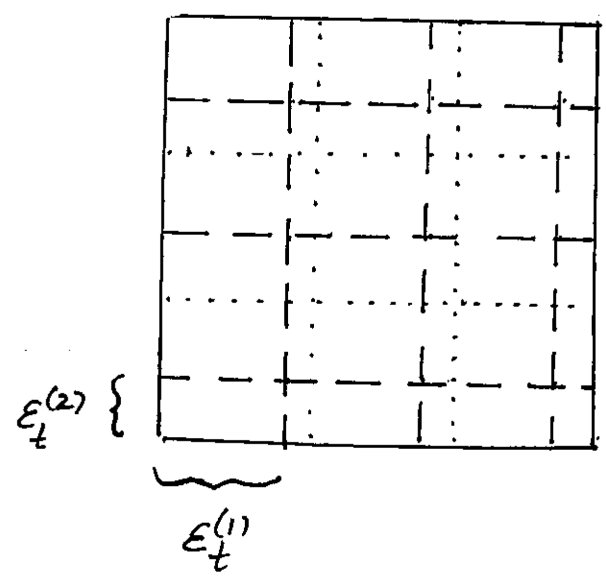
(B) ______________ classifiers as follows.

For $t = 1, \dots, T$

- generate $\varepsilon_t^{(1)}, \varepsilon_t^{(2)} \in [0 \; \tfrac{1}{3})$

  ______________ ______ ______________.

- shift the histogram by $[\varepsilon_t^{(1)}, \varepsilon_t^{(2)}]^T$
  and construct f_t based on the shifted partition.



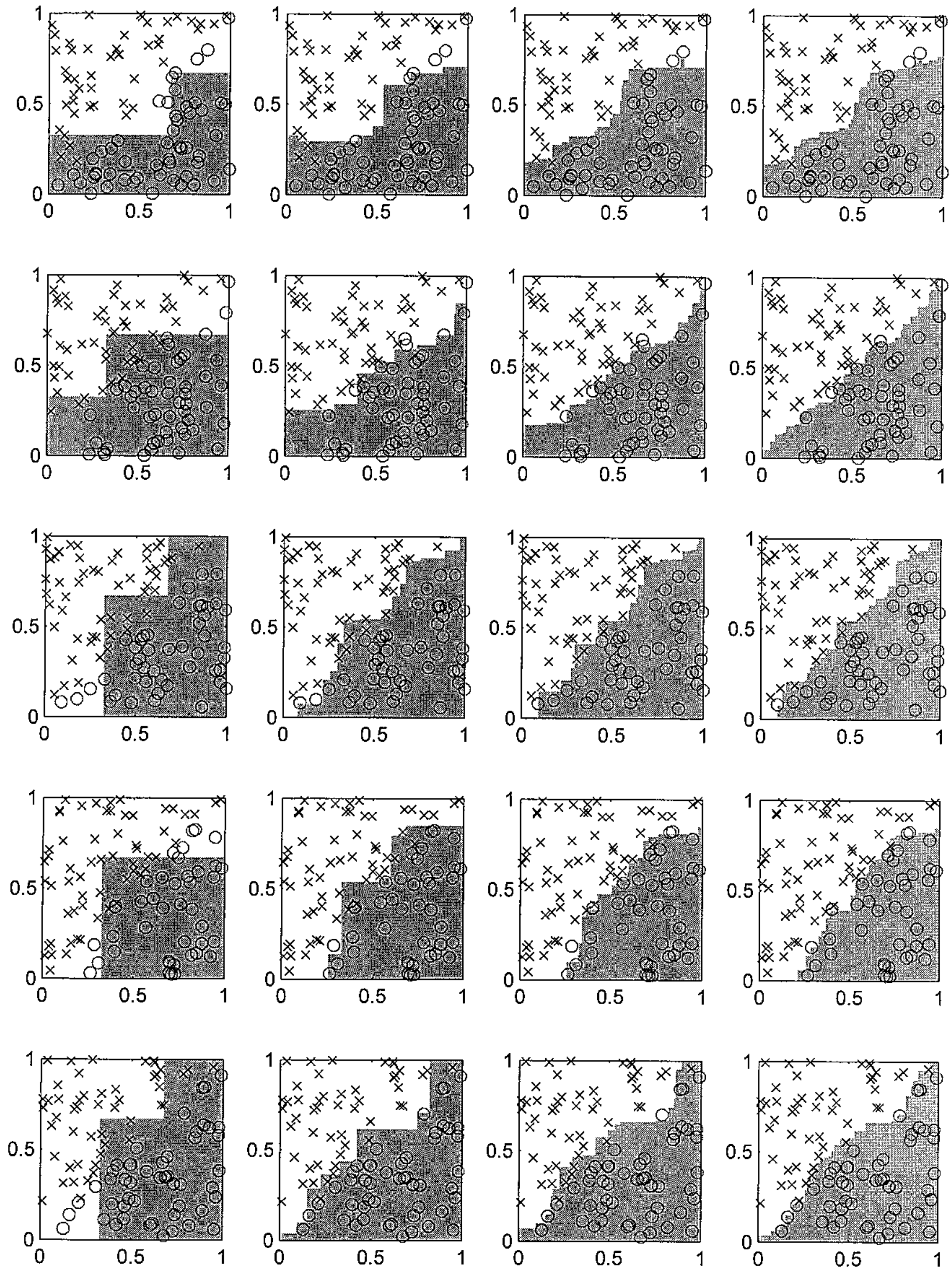
Now define the <u>ensemble classifier</u>

(C) $f(x) =$

This classifier is remarkably effective.

# of votes = 1 5 11 21

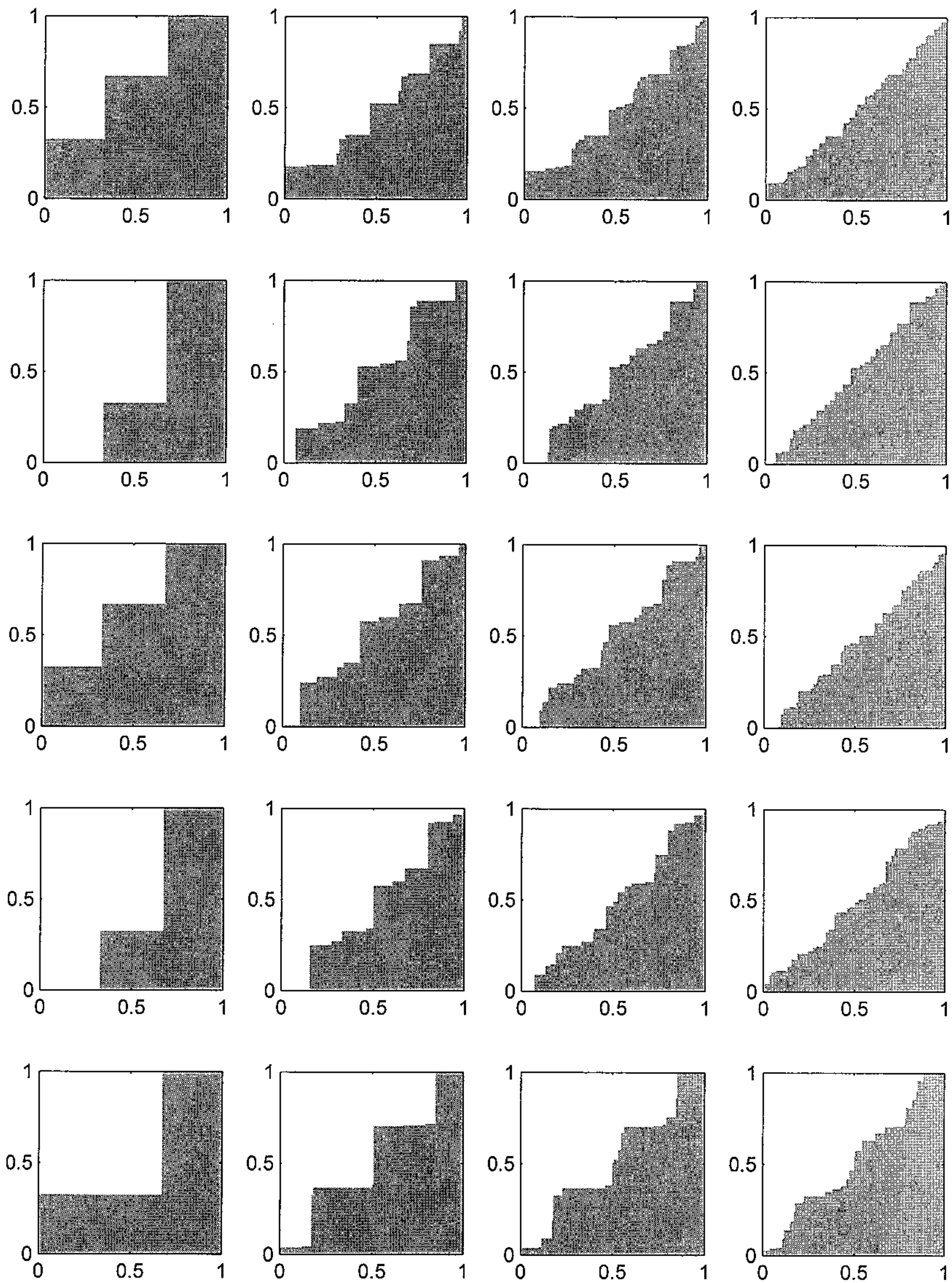
5 realizations of data



n = 100 points

# of votes = 1 | 5 | 11 | 21

5 realizations of data



n= 1000 points

Performance

Fix $x \in [0,1]^2$. Let $f^*(x)$ = correct label.

For any $t = 1, \dots, T$ we have

$$\Pr\{ f_t(x) \neq f^*(x) \}$$

with respect to choice of $\varepsilon_t^{(1)}, \varepsilon_t^{(2)}$

$$= \Pr\left\{ \begin{array}{l} \text{cell containing } x \text{ has } < \frac{1}{2} \\ \text{its area in same class as } x \end{array} \right\}$$

$$=: p(x) < \frac{1}{2}$$

Introduce the variable $Z_x \sim \text{binom}(T, p(x))$.

Then

$$\Pr\{ f(x) \neq f^*(x) \}$$

$$= \Pr\left\{ Z_x > \frac{T}{2} \right\}$$

$$= \Pr\left\{ Z_x > T \cdot p(x) + T\left(\tfrac{1}{2} - p(x)\right) \right\}$$

Chernoff's bound

$$\leq e^{-T\left(\frac{1}{2} - p(x)\right)^2} \to 0 \quad \text{as } T \to \infty$$

This simple example illustrates two important properties of ensemble rules:

1. Combining classifiers that are

(D) __________ and __________ to

form a classifier that is

__________ and __________.

2. Increased __________.

<u>Definition</u> A classifier (or model) is <u>stable</u> if small changes in the training data do not result in large changes to the final classifier.

Our primary example of an unstable

(E) classifier is a __________ ________.

On the downside, we lose __________.

Bagging

(F) Bagging stands for __________ ______________.

Fix $B \geq 1$. Let I_b be a subset of $\{1,2,\dots,n\}$ of size n, obtained by sampling w replacement.

Suppose we have adopted a specific learning strategy (e.g., decision trees, LDA) and set

$$f_b =$$

The bagging classifier is

$$f(x) =$$

Random forests

Random forests are ensemble methods that combine decision trees and some kind of randomization or resampling.

In addition to bagging, the most notable other random forest grows a large number of trees using a greedy procedure such that, at each node, the split is selected from a ________ ________ of ________.

Among other advantages, this allows the application of trees to very ______ ________ data.

$\boxed{\text{Key}}$ A. cell, majority vote

B. poor, uniformly at random

C. majority vote over $f_1(x), \ldots, f_T(x)$

D. simple, poor ; complex, accurate ; stability

E. decision tree ; interpretability

F. Bootstrap aggregation,

$f_b(x)$ = classifier based on $\{(x_i, y_i)\}_{i \in I_b}$

$f(x)$ = majority vote over $f_b(x)$, $b = 1, \ldots, B$

G. random subset of features ; high dimensional