ENSEMBLE METHODS

The idea behind ensemble methods is to generate several classifiers $f_1, \ldots, f_T$ 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,

$\mathbf{x}_i \in [0,1]^2$.

Bayes error = 0
A very basic (and not recommended!) classifier is a **histogram rule**:

- assign the same label to patterns $x$
in the same $\square$.
- determine label by $\square$.

As you can imagine, this classifier will not perform very well.
Let's generate a whole bunch of equally classifiers as follows:

For $t = 1, ..., T$

- generate $\xi_t^{(1)}, \xi_t^{(2)} \in [0, \frac{1}{3})$

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

Now define the **ensemble classifier**

$f(x) = \ldots$

This classifier is remarkably effective.
Number of votes: 1, 5, 11, 21

5 realizations of data

n = 100 points
$n = 1000$ points
Fix $x \in [0,1]^2$. Let $f^x(x) =$ correct label.

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

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

with respect to choice of $\xi_t^{(n)}$, $\xi_t^{(m)}$

$$= \Pr \left\{ \text{cell containing } x \text{ has } < \frac{1}{2} \text{ area} \right\}$$

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

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

Then

$$\Pr \left\{ f(x) \neq f^x(x) \right\}$$

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

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

$$\leq e^{-T \left( \frac{1}{2} - p(x) \right)^2} \rightarrow 0 \text{ as } T \rightarrow \infty$$
This simple example illustrates two important properties of ensemble rules:

1. Combining classifiers that are _____ and _____ to form a classifier that is _____ and _____.

2. Increased _____.

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

Our primary example of an unstable classifier is a _____.

On the downside, we lose _____.
Bagging

Bagging stands for _______ _______.

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

\[ f_b(x) = \text{classifier based on } \{(x_i, y_i)\}_{i \in I_b} \]
\[ f(x) = \text{majority vote over } f_b(x), \quad b = 1, \ldots, B \]

G. random subset of features; high dimensional