# CS 360: Machine Learning

Sara Mathieson, Sorelle Friedler

Spring 2024

HAVERFORD COLLEGE

# Admin

- **Lab 3** due Thursday
- **Lab 1** graded on Moodle
  - You might have received a note about results being slightly off, that's okay since you might have used different hyperparameters (we didn't take off for this)
- By the end of lab today, make sure you can **build the decision tree**

# Outline for Feb 13

- Ensemble methods introduction
- Bagging
- Random forests
- AdaBoost

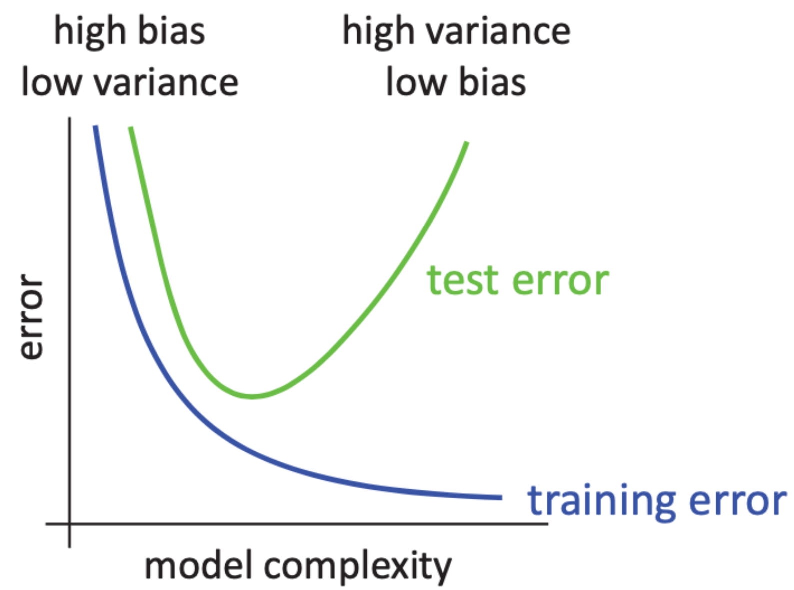
# Outline for Feb 13

- Ensemble methods introduction
- Bagging
- Random forests
- AdaBoost

# Recap Bias-Variance Tradeoff

Recall ML balances

- estimation error (variance)
  - precision of match
  - sensitivity to training data
- structural error (bias)
  - distance from true relationship

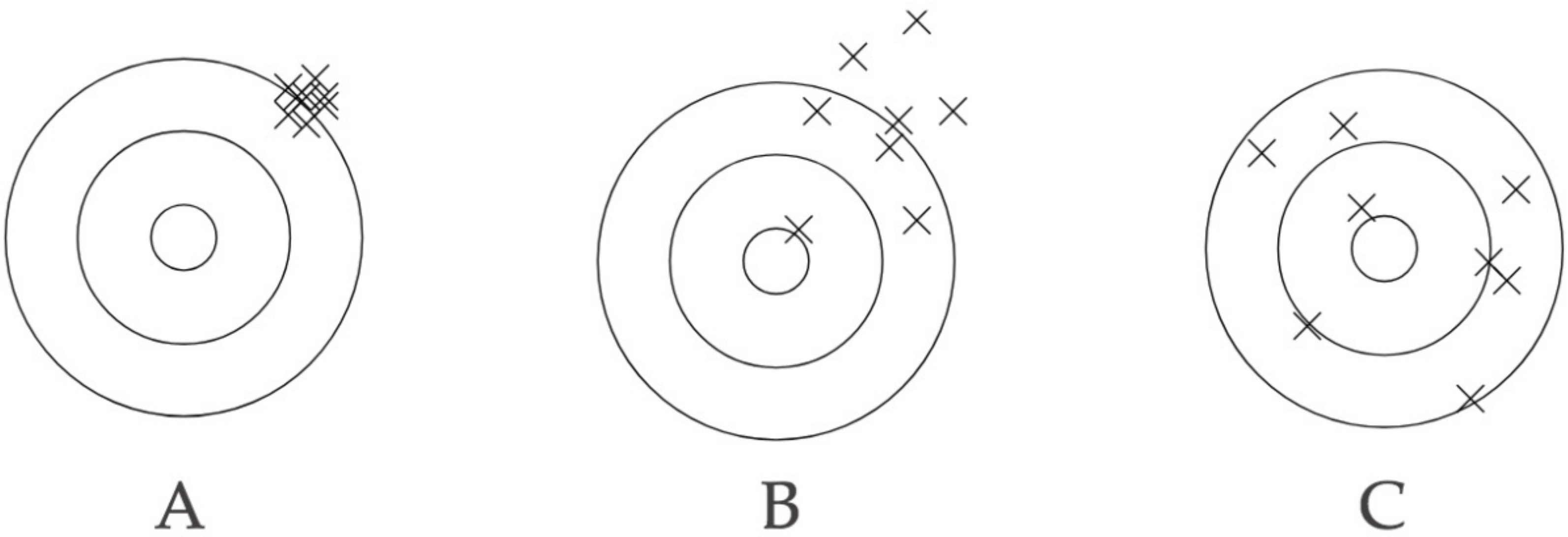



Goals

- Reduce variance without increasing bias
- Reduce bias and variance

Slide: Jessica Wu

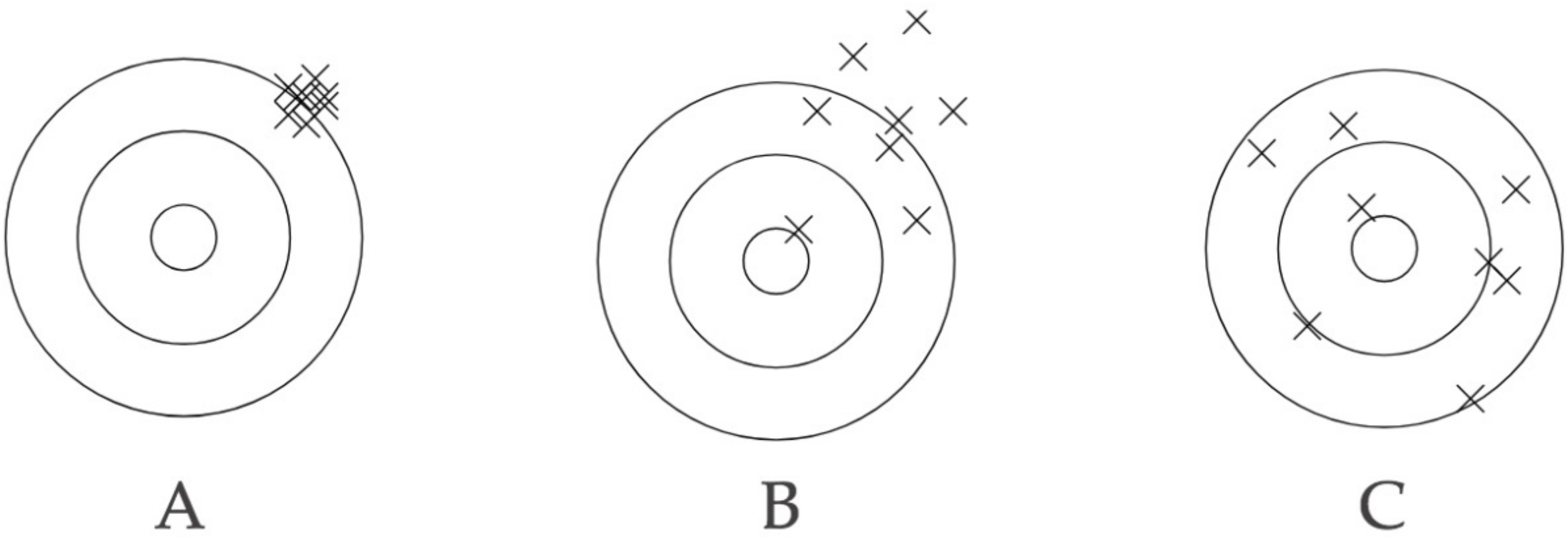
# Quiz: recap bias and variance




Label each picture with variance (high or low) and bias (high or low)

Example from Ameet Soni

# Quiz: recap bias and variance

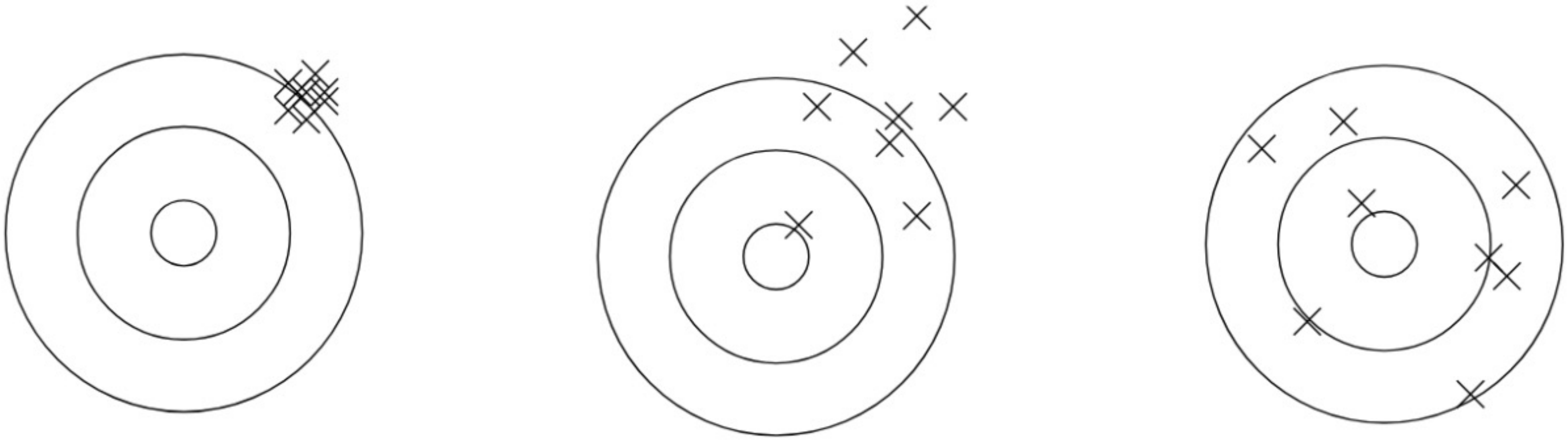



Variance: low
Bias: high

Label each picture with variance (high or low) and bias (high or low)

Example from Ameet Soni

# Quiz: recap bias and variance



A B C

**A**

Variance: low
Bias: high

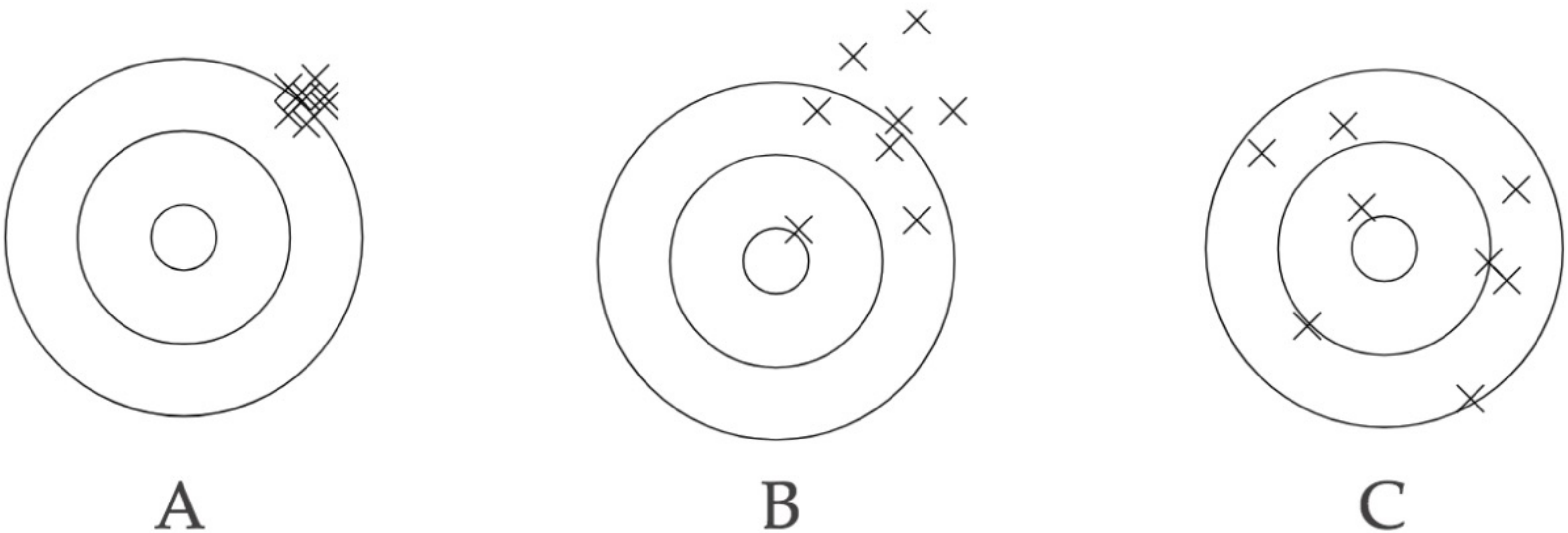
**B**

Variance: high
Bias: high

Label each picture with variance (high or low) and bias (high or low)

Example from Ameet Soni

# Quiz: recap bias and variance

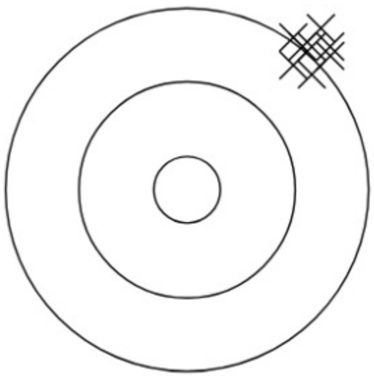



| | A | B | C |
|---|---|---|---|
| Variance: | low | high | high |
| Bias: | high | high | low |

Label each picture with variance (high or low) and bias (high or low)

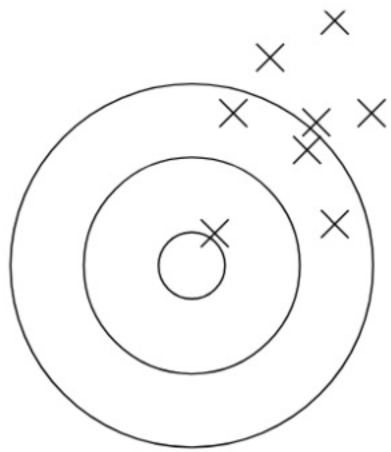
Example from Ameet Soni

# Quiz: recap bias and variance



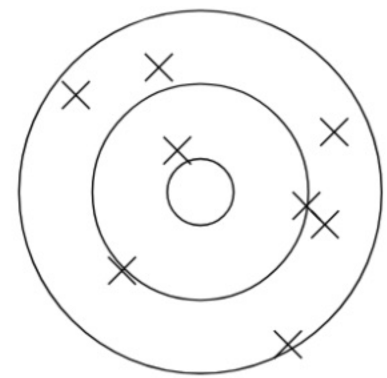
A

Variance: low
Bias: high



B

Variance: high
Bias: high



C

Variance: high
Bias: low

This is the type of classifier we want to average!

Label each picture with variance (high or low) and bias (high or low)

Example from Ameet Soni

# Ensemble Idea

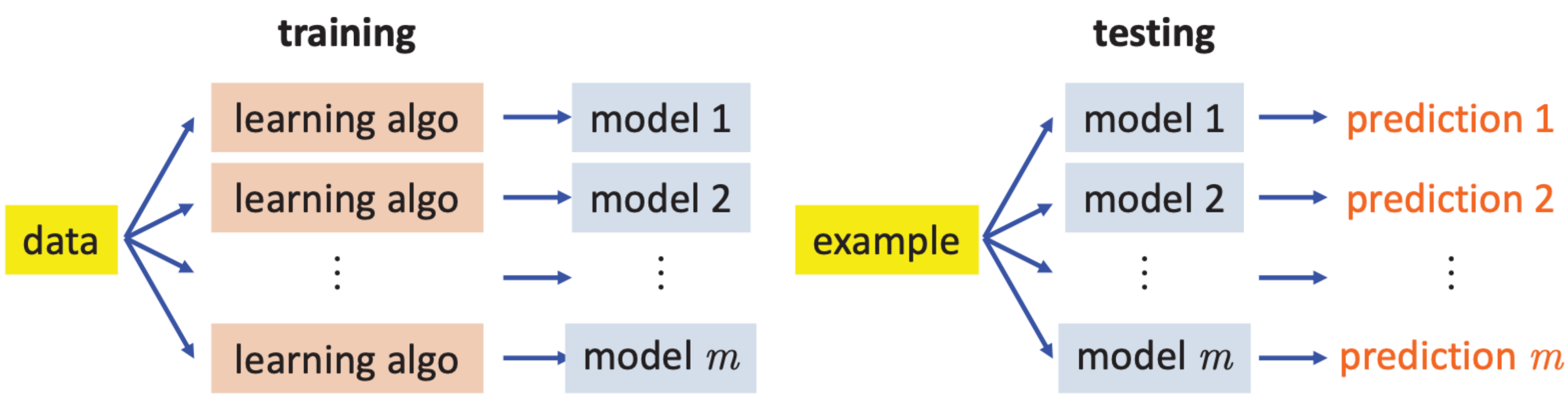
If one classifier works well…

why not use multiple classifiers!?!

Ensemble: "All the parts of a thing taken together so that each part is considered in relation to the whole."

ML Ensemble:

**Combine multiple classifiers** to improve prediction




Based on slide by David Kauchak

Slide: Jessica Wu

# Ensemble Idea

- Average the results from several models with high variance and low bias
  - Important that models be diverse (don't want them to be wrong in the same ways)

- If $n$ observations each have variance $s^2$, then the mean of the observations has variance $s^2/n$ (reduce variance by averaging!)

# Learning Theory

Let $H$ be the hypothesis space

Three sources of $\boxed{\text{limitations}}$ for traditional classifiers:

- Statistical - *H* is too large relative to size of data
  - Many hypotheses can fit the data by chance
- Computational - *H* is too large to completely search for "best" model
- Representational - *H* is not expressive enough

# Learning Theory

- Statistical: Average of unstable models (high variance) has more stability
- Computational: searching from multiple starting points is better approximation than one starting point
- Representational: sum of many models can represent more hypotheses than an individual model

# Learning Theory

- Statistical: Average of unstable models (high variance) has more stability
- Computational: searching from multiple starting points is better approximation than one starting point
- Representational: sum of many models can represent more hypotheses than an individual model

Ensembles can address all 3!

# Learning Theory

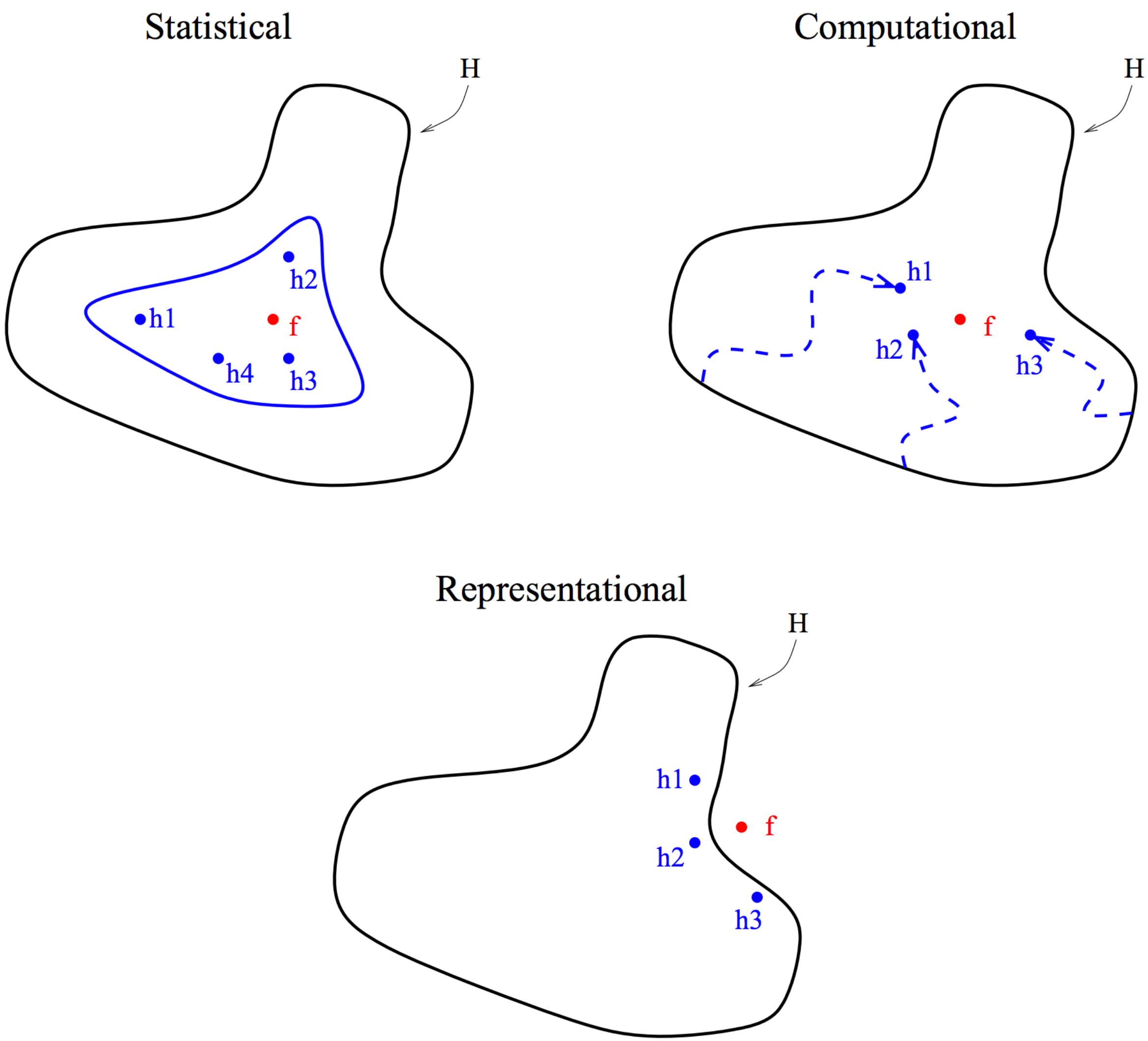


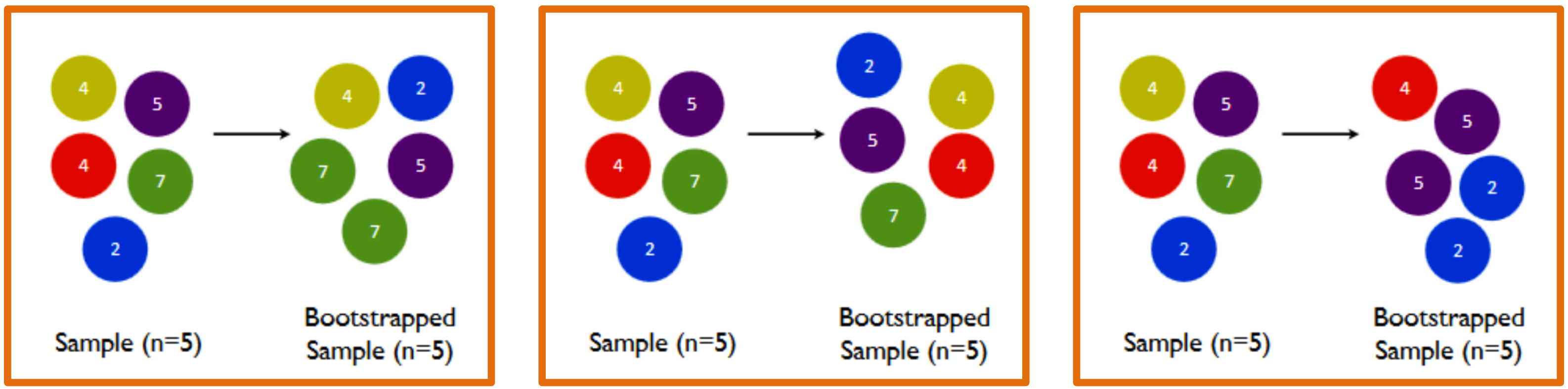

Figure from Tom Dietterich

# Outline for Feb 13

- Ensemble methods introduction
- Bagging
- Random forests
- AdaBoost

# Bagging Algorithm

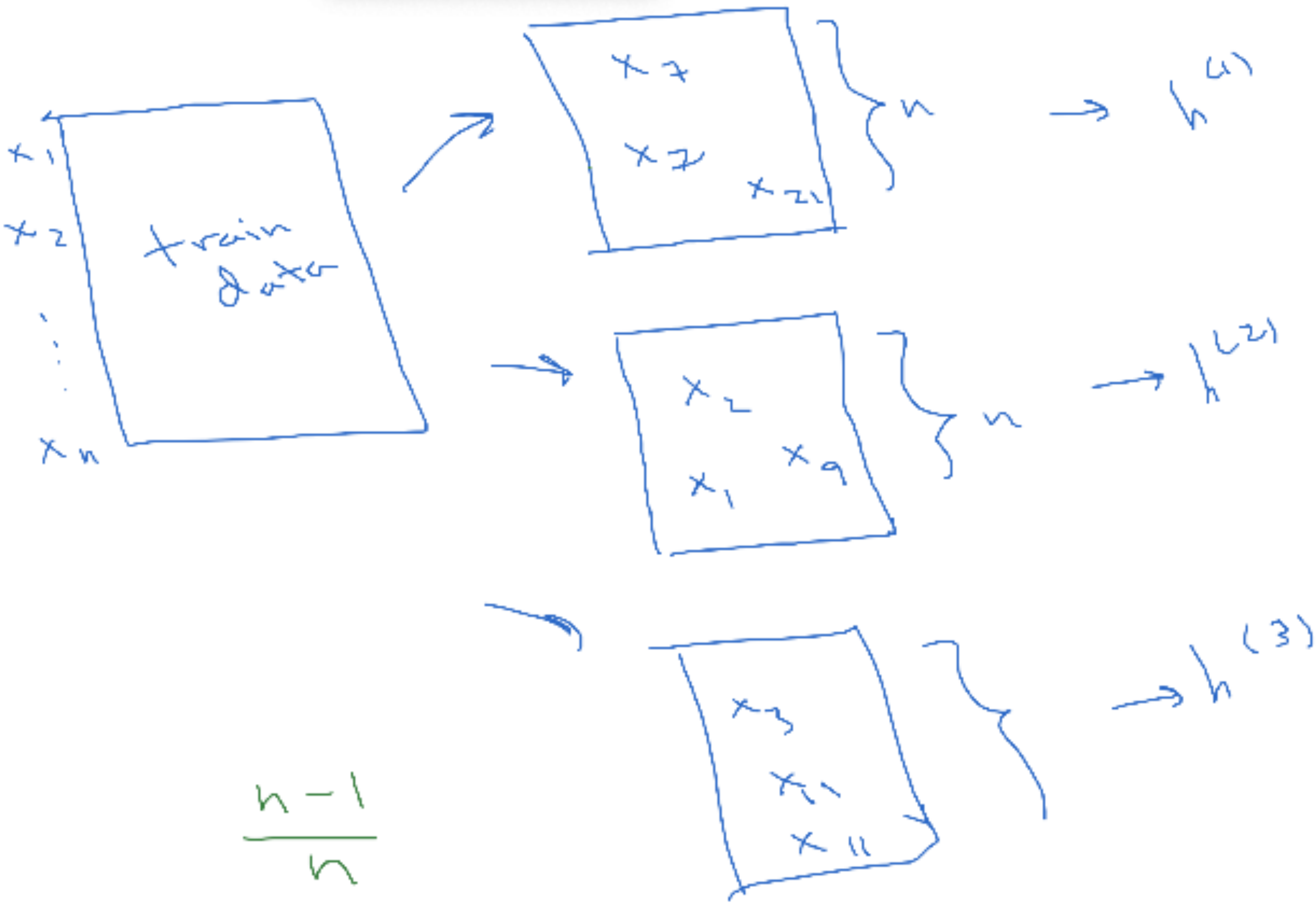
- Bagging = Bootstrap Aggregation [Brieman, 1996]
- *Bootstrap* (randomly sample with replacement) original data to create many different training sets
- Run base learning algorithm on each new data set independently




*Desmond Ong, Stanford*

Modified from Ameet Soni

# Bootstrap (Handout 7, Q1)



# Bootstrap (Handout 7, Q1)

(a) $\frac{n-1}{n}$

(b) $\left(\frac{n-1}{n}\right)^n$

(c) $\left(1-\frac{1}{n}\right)^n \approx e^{-1}$

(d) 0.37

# Ensemble Notation

- $T$ = number of models/classifiers (index $t$)
- $X^{(t)}$ = bootstrap training set $t$
- $x$ = generic test example (could be a vector)
- $y$ = test label (binary for now)
- $h^{(t)}(x)$ = hypothesis about $x$ from model $t$
- $r$ = probability of error of individual model
- $R$ = number of votes for wrong class

# Bagging (bootstrap aggregation)

__Train__ for $t$ in $T$:

- create bootstrap dataset $X^{(t)}$
- train on $X^{(t)}$ to get $h^{(t)}$

__Test__ for $x$ in test data

$$h(x) = \underset{y \in \{0,1\}}{\operatorname{argmax}} \sum_{t=1}^{T} \mathbb{1}\left(h^{(t)}(x) = y\right)$$

indicator function

# Bagging (bootstrap aggregation)

$r$ = prob base classifier is wrong

$R$ = # votes for the wrong class

$$P(R=k) = \binom{T}{k} \underbrace{r^k}_{\substack{\text{wrong} \\ k \text{ times}}} \underbrace{(1-r)^{T-k}}_{\substack{\text{right} \\ T-k \\ \text{times}}} \Bigg\} \text{ binomial distribution}$$

$$\left(\frac{1}{4}\right)^2 \quad \left(\frac{3}{4}\right)^1$$

$$\boxed{P(\text{overall wrong})} = P\left(R > \frac{T}{2}\right)$$

$$= \sum_{k=\frac{T+1}{2}}^{T} \binom{T}{k} r^k (1-r)^{T-k}$$

$T=3$

$k=2$

$r = \frac{1}{4}$

$\boxed{Y_{true} = 1}$

| $h$ | $h^{(1)}$ | | $h^{(2)}$ | | $h^{(3)}$ | |
|---|---|---|---|---|---|---|
| 0 | 0 | $\frac{1}{4}$ | 1 | $\frac{3}{4}$ | 0 | $\frac{1}{4}$ |

# Bagging (Handout 7, Q2)

| $h^{(1)}$ | $h^{(2)}$ | $h^{(3)}$ | |
|---|---|---|---|
| 0 | 0 | 0 | $\rightarrow \left(\frac{1}{4}\right)^3$ |
| 0 | 0 | 1 | $\rightarrow \left(\frac{1}{4}\right)^2 \left(\frac{3}{4}\right)$ |
| 0 | 1 | 0 | $\rightarrow$ " |
| 1 | 0 | 0 | $\rightarrow$ " |

$$p\left(\begin{matrix}\text{overall}\\ \text{wrong}\end{matrix}\right) = \binom{3}{2}\left(\frac{1}{4}\right)^2 \frac{3}{4} + \left(\frac{1}{4}\right)^3 = \frac{9+1}{64} \approx \boxed{0.16}$$

$$\boxed{0.16 < \underbrace{0.25}_{\epsilon}}$$

☆ ☆

# Outline for Feb 13

- Ensemble methods introduction
- Bagging
- Random forests
- AdaBoost

# Random Forests

- Idea: choose a different subset of features for every classifier *t*
- Typically use *decision stumps* (depth 1)

- Goal: decorrelate models

- In practice: choose sqrt(p) features
  - Without replacement for each model
  - Every model: data points and features chosen independently

# Outline for Feb 13

- Ensemble methods introduction
- Bagging
- Random forests
- AdaBoost

# Boosting idea

**Train**

- begin with equal weights on all examples
- for T iterations
  - learn a classifier on the weighted examples
  - change example weights based on train error

**Test**

- get a prediction from each classifier (T)
- vote based on how well classifier did during training

## AdaBoost (adaptive boosting)

set $w_i^{(1)} = \frac{1}{n}$ (equal weight on all examples)

for $t = 1 \cdots T$

(a) fit a classifier to weighted training set $\Rightarrow h^{(t)}(\vec{x})$

(b) compute weighted classification error $\mathbb{1}$

$$\varepsilon_t = \sum_{i=1}^{n} w_i^{(t)} \underbrace{\mathbb{1}\left(y_i \neq h^{(t)}(\vec{x}_i)\right)}_{\text{prediction was wrong}}$$

(c) compute model score

weight models $\rightarrow$ $\alpha_t = \frac{1}{2} \log\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$