CS 383: Machine Learning

Prof Adam Poliak

Fall 2024

10/29/2024

Lecture 17

Announcements

Lecture tomorrow Wednesday 10/30

Thursday reading quiz: Duame textbook Chapter 13 (Ensemble chapter)

Outline

Logistic Regression

Ensemble Methods

- Bagging
- Boosting
- Weighted Entropy

Multi-class prediction

$$J(\theta) = \sum_{i}^{n} y_i * \log\left(h_\theta(x_i)\right) + (1 - y_i) * log\left(1 - h_\theta(x_i)\right)$$

What should our loss be in multi-class prediction with k categories?

$$J(\theta) = \sum_{i}^{n} \sum_{j}^{k} y_{i,j} * \log\left(h_{\theta,j}(x_i)\right)$$

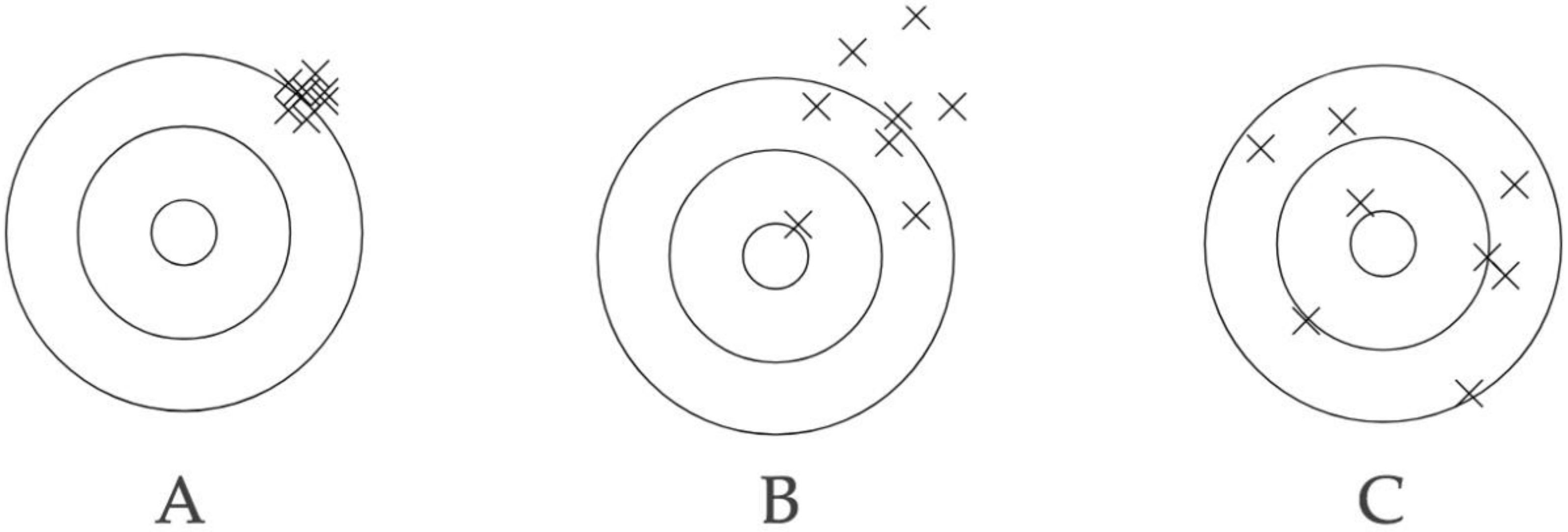
Outline

Logistic Regression

Ensemble Methods

- Bagging
- Boosting
- Weighted Entropy

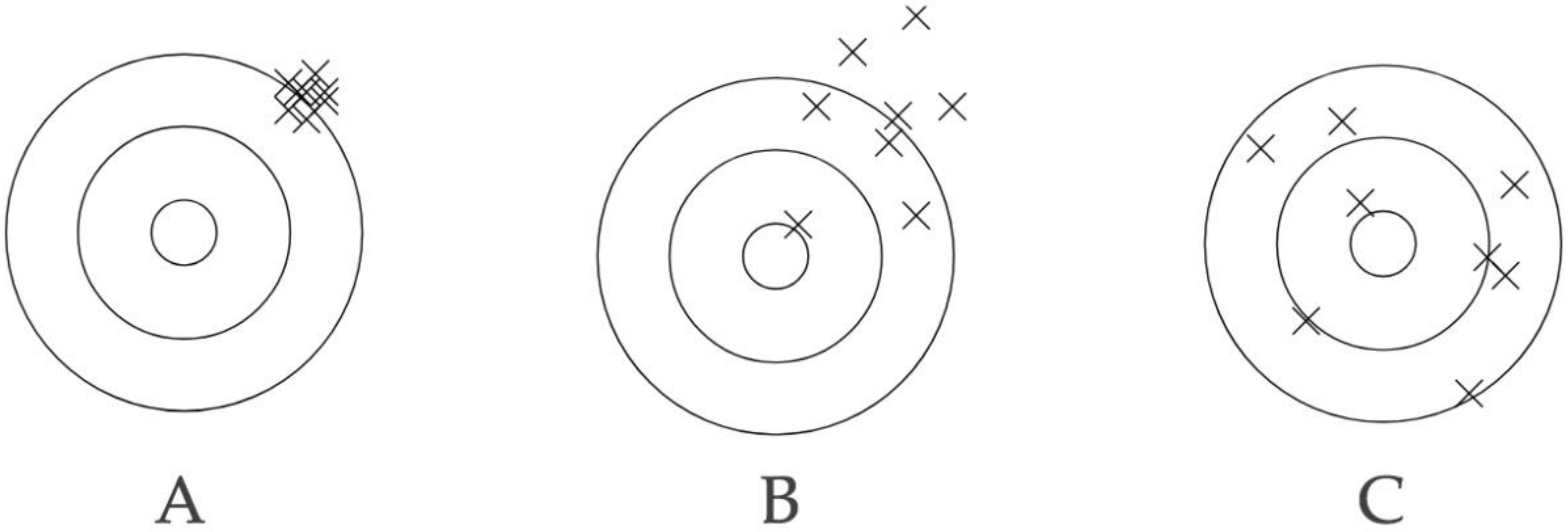
Quiz: recap bias and variance



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

Example from Ameet Soni

Quiz: recap bias and variance

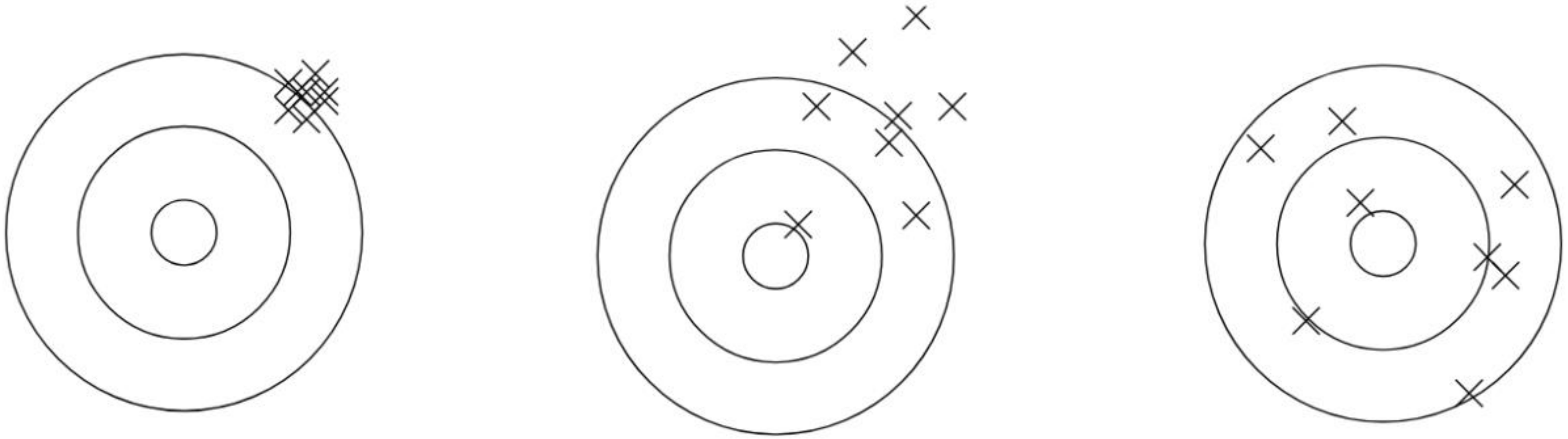


Variance: low
Bias: high

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

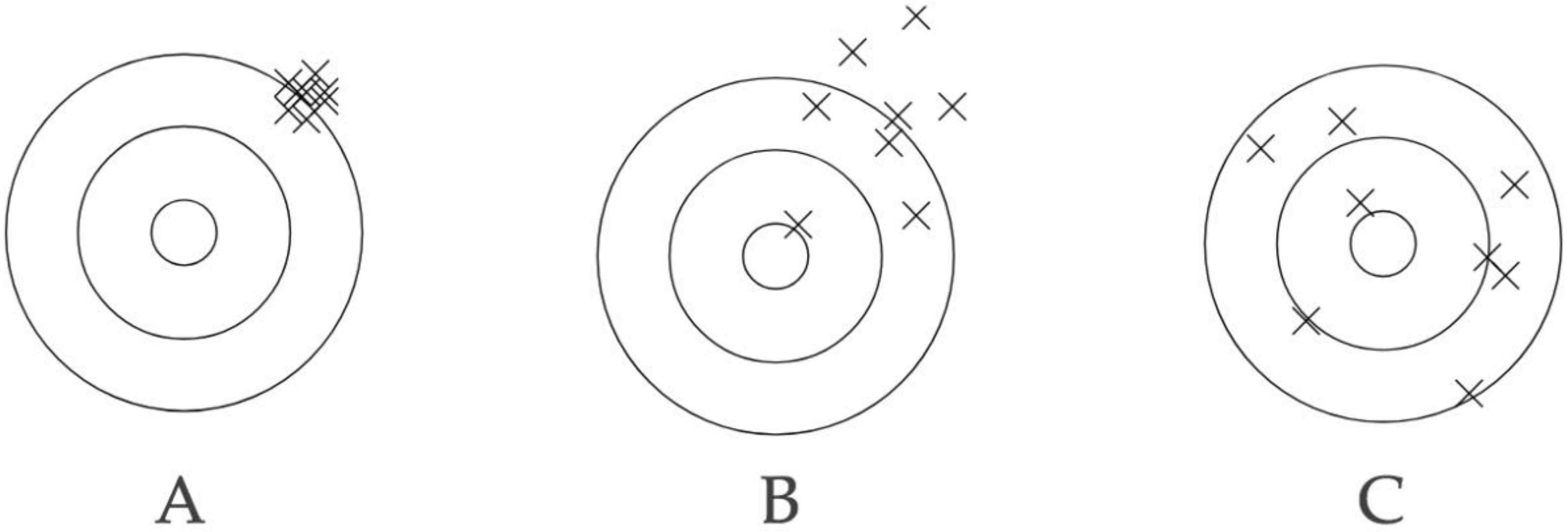
Variance: low
Bias: high

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

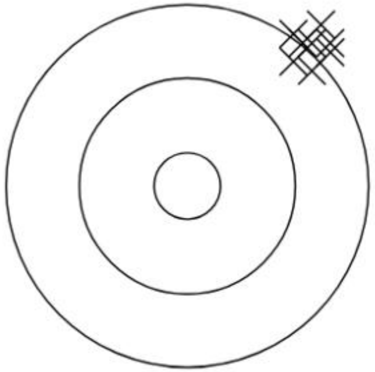


Variance:	low	Variance:	high	Variance:	high
Bias:	high	Bias:	high	Bias:	low

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

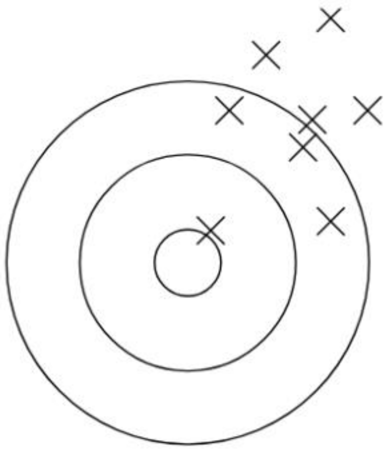
Example from Ameet Soni

Quiz: recap bias and variance



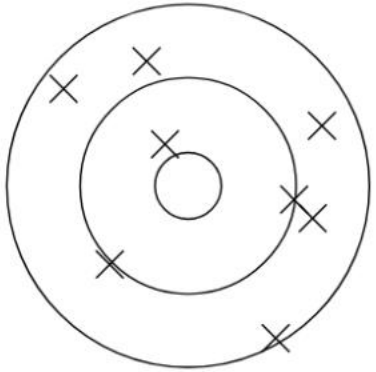
A

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 Intuition

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

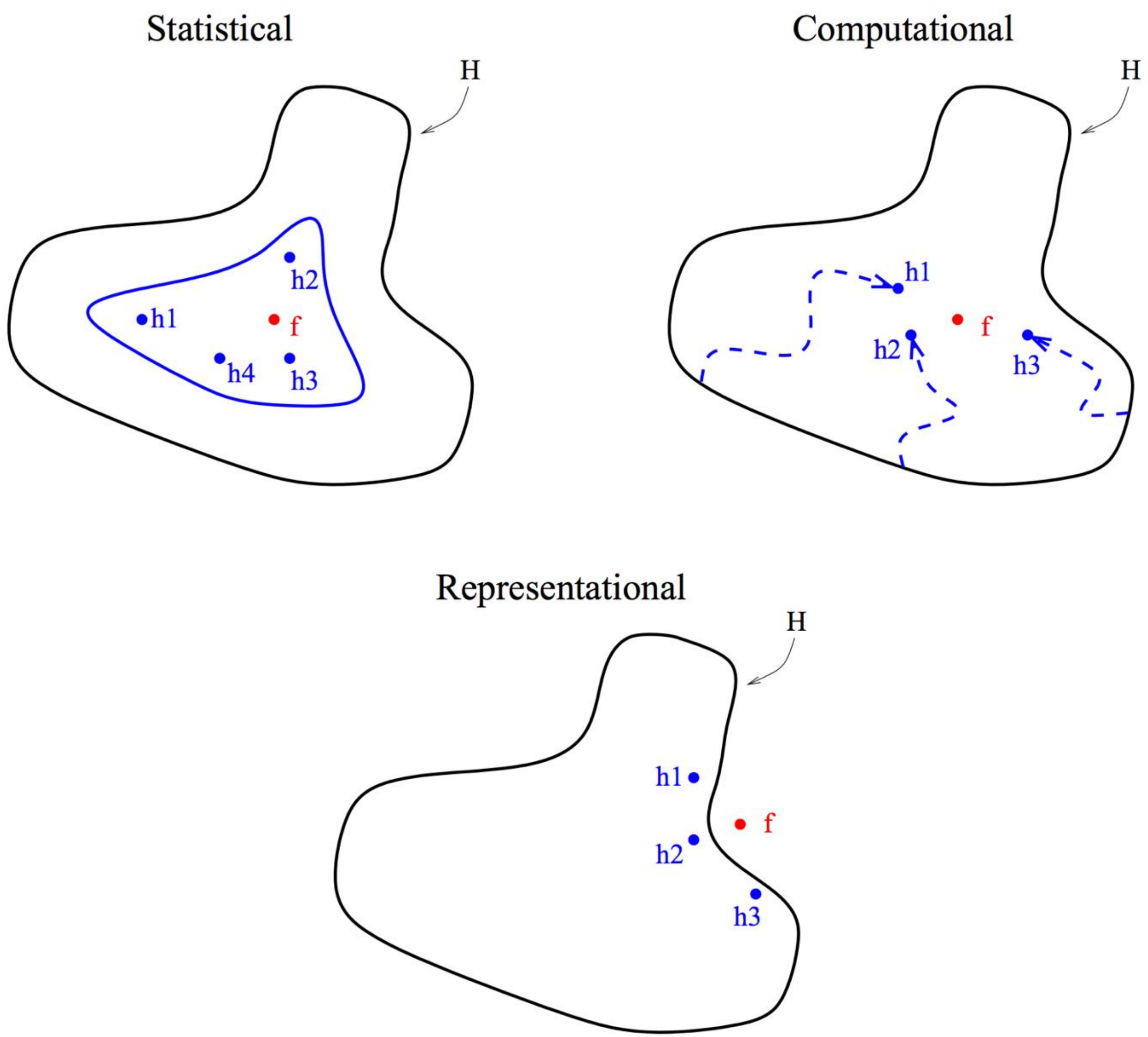
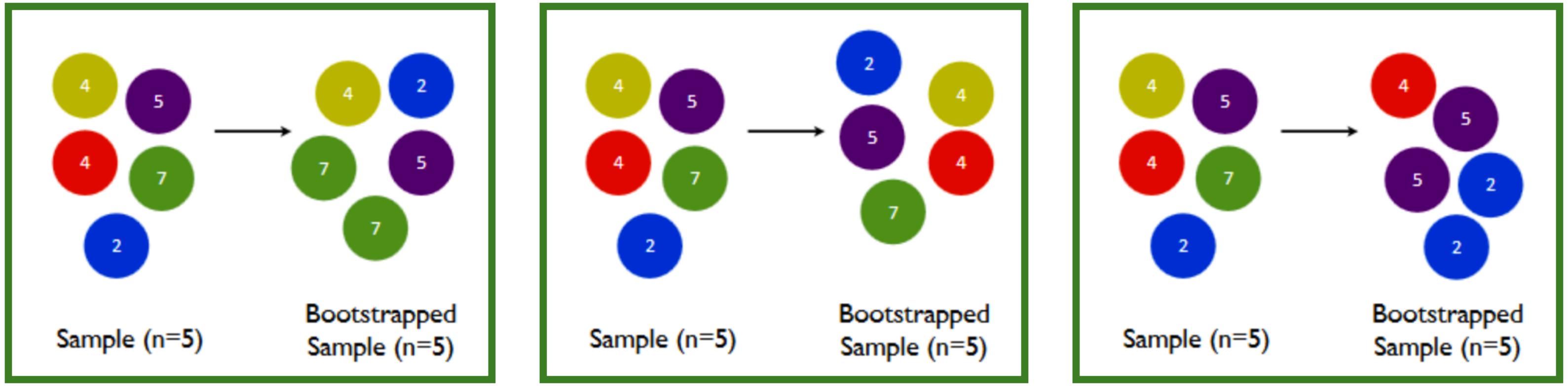


Figure from Tom Dietterich

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

Notation

T: # of models/classifiers

x: test example

$X^{(t)}$: bootstrap training set t

$h^{(t)}(x)$: hypothesis about x from model t

r: probability of error of individual model

R: number of votes for wrong class

Bagging Algorithm

Train

Generate $X^{(t)}$ for $t = 1, \ldots, T$ using bootstrap sampling

Train classifier $h^{(t)}$ on $X^{(t)}$

Test

for x in test data:

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

Probability that $R = k$?

$$P(R = k) = \binom{T}{k} r^k (1-r)^{T-k}$$

What is probability that ensemble is wrong?

$$P\left(R > \frac{T}{2}\right) = \sum_{k=\frac{T+1}{2}}^{T} \binom{T}{k} r^k (1-r)^{T-k}$$

If $r < \frac{1}{2}, \lim_{T \to \infty} P\left(R > \frac{T}{2}\right) = 0$

Random Forest

Idea: choose a different subset of features for every classifier *t*

Choose weak/base classifiers

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

Logistic Regression

Ensemble Methods

- Bagging
- **Boosting**
- Weighted Entropy