CS 360: Machine Learning

Prof. Sara Mathieson

Fall 2020

Admin

- **Lab 5 due TODAY!**
  - Grace period until Wed night
  - No office hours today (were yesterday)
- Lab 6 posted today
- Welcome prospective students!

Outline for October 20

- Recap multi-class logistic regression
- Introduction to ensemble methods
- Bagging
- Random forests
- AdaBoost

Outline for October 20

- Recap multi-class logistic regression
- Introduction to ensemble methods
- Bagging
- Random forests
- AdaBoost

Multi-class Logistic Regression

K = num classes (political parties, blood groups, etc)

2 classes

$$h_{\vec{w}}(\vec{x}) = \frac{1}{1+e^{-\vec{w}\cdot\vec{x}}}$$

$1 - h(\vec{x})$ = prob 0

prob 1

$$= \frac{e^{\vec{w}\cdot\vec{x}}}{e^{\vec{w}\cdot\vec{x}} + 1}$$

$y = 1$

weight $y = 1$

weight $y = 0$

Generalize

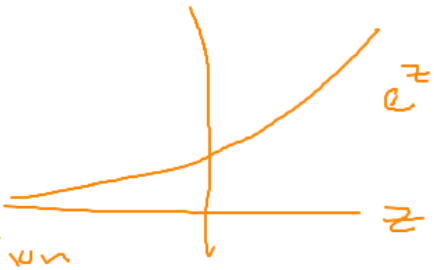
$k = 1, 2, 3 \cdots K$

$$\vec{h}_w(\vec{x}) = \begin{bmatrix} p(y=1|\vec{x}) \\ p(y=2|\vec{x}) \\ \vdots \\ p(y=K|\vec{x}) \end{bmatrix} = \frac{1}{\sum_{k=1}^{K} e^{\vec{w}^{(k)}\cdot\vec{x}}} \begin{bmatrix} e^{\vec{w}^{(1)}\cdot\vec{x}} \\ e^{\vec{w}^{(2)}\cdot\vec{x}} \\ \vdots \\ e^{\vec{w}^{(K)}\cdot\vec{x}} \end{bmatrix}$$

positive or negative

called logits

probability distribution



$$W = \begin{bmatrix} | & | & & | \\ \vec{w}^{(1)} & \vec{w}^{(2)} & \cdots & \vec{w}^{(K)} \\ | & | & & | \end{bmatrix}$$

$(p+1) \times K$

weight vector for each class

Multi-class Logistic Regression: cost function

still neg log likelihood $\{h^{y}(1-h)^{1-y}\}$

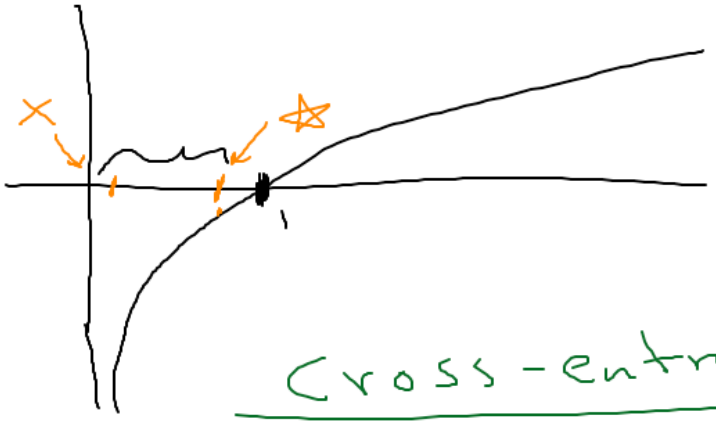
$\vec{w}^{(k)} \cdot \vec{x}$

$$J(W) = -\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik} \log p(y_i = k \mid \vec{x}_i)$$

indicator

e / normalizer

"one-hot"



$p \log p$

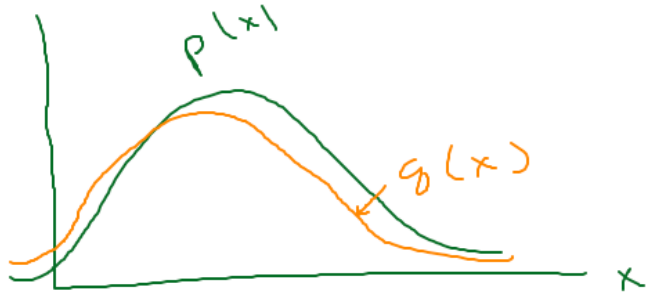
Cross-entropy

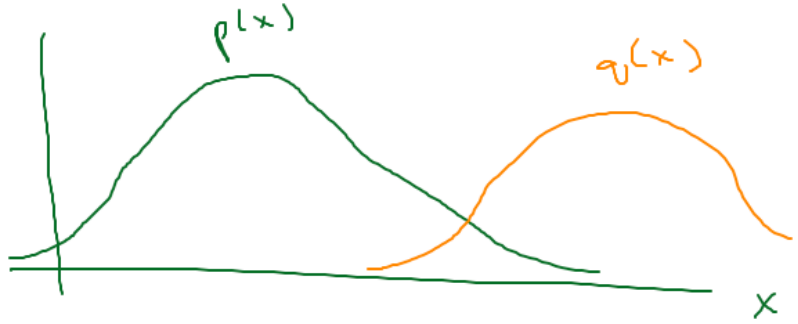
$y_i = 3$ (class 3)

$$\vec{y}_i = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$H(p, q) = -\sum_{x} p(x) \log q(x)$$

Cross Entropy

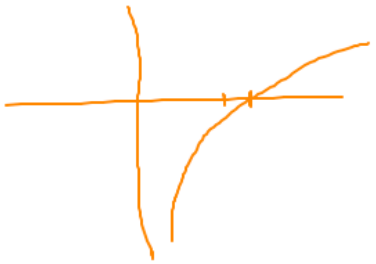


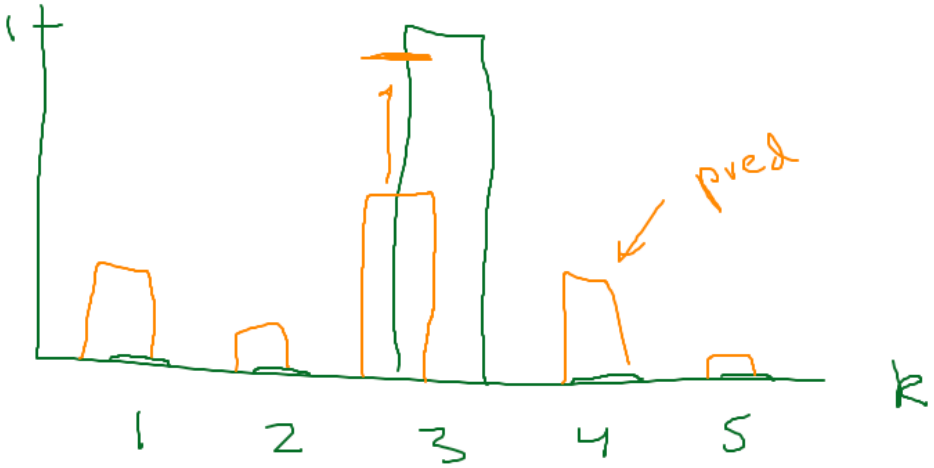


Log Regression :

$p(x) \rightarrow$ true

$q(x) \rightarrow$ pred





$K = 5$ ≈ 1 high

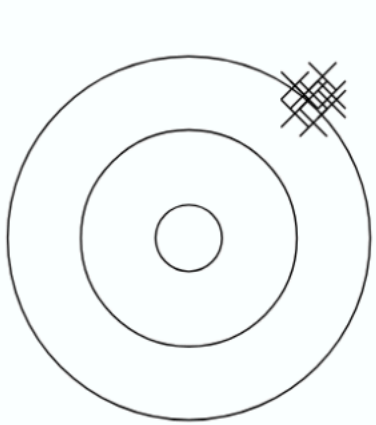
$-\sum_k true \cdot \log(pred)$ small

1 for class 3

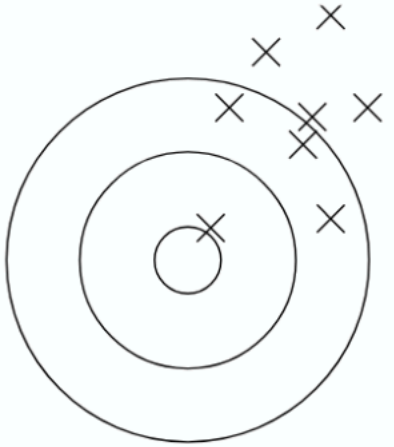
Outline for October 20

- Recap multi-class logistic regression
- Introduction to ensemble methods
- Bagging
- Random forests
- AdaBoost

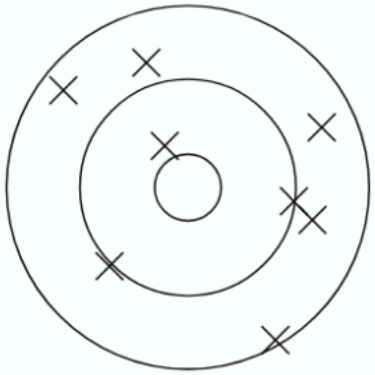
Quiz: recap bias and variance



A



B

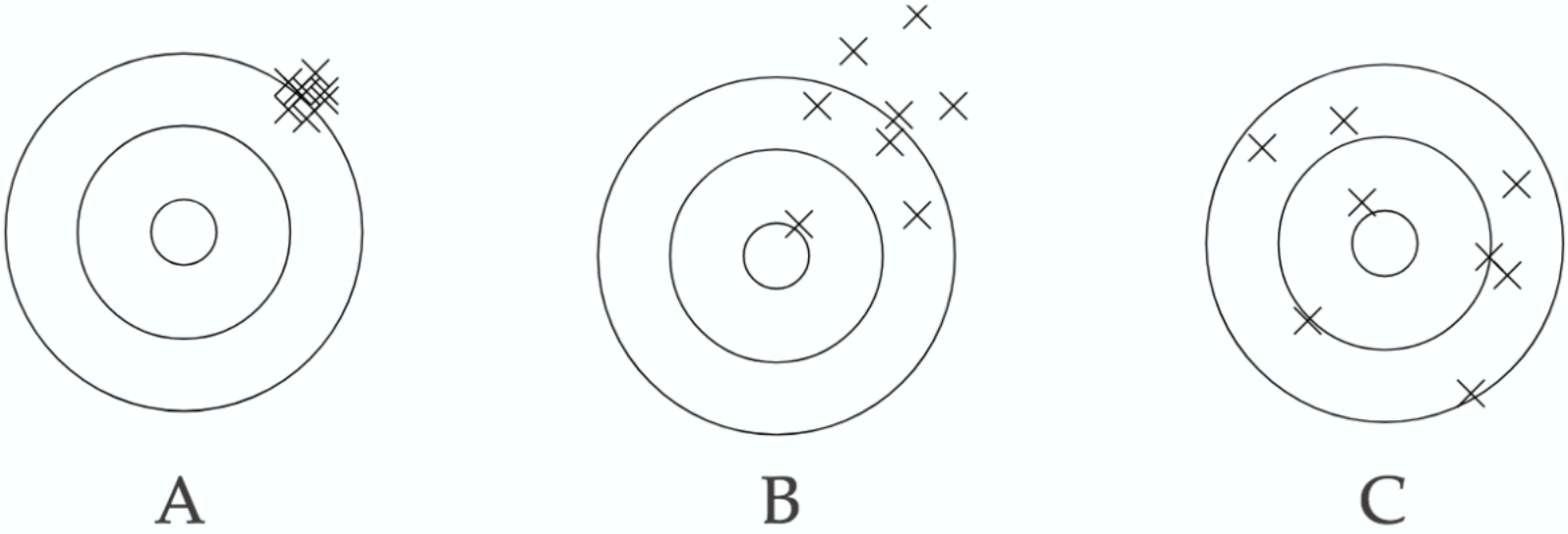


C

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

Example from Ameet Soni

Quiz: recap bias and variance

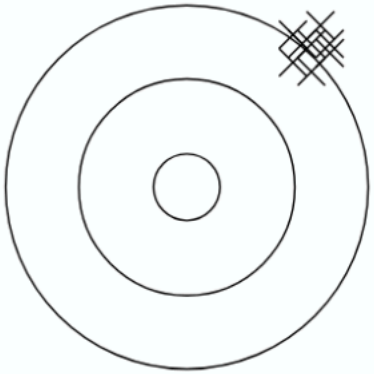


Variance: low
Bias: high

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

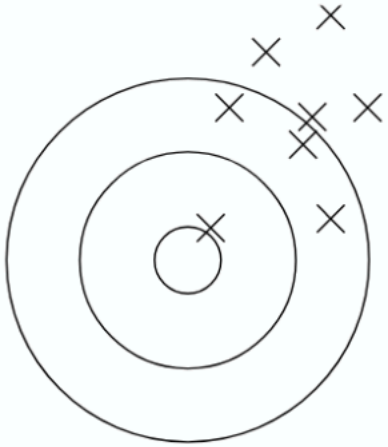
Example from Ameet Soni

Quiz: recap bias and variance



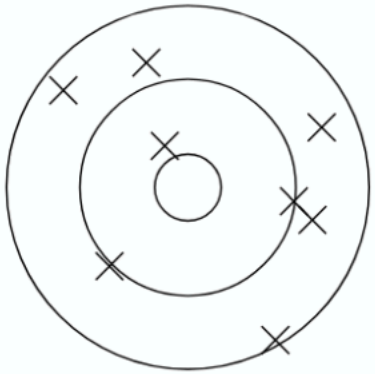
A

Variance: low
Bias: high



B

Variance: high
Bias: high

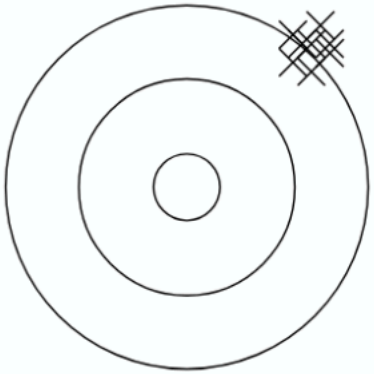


C

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

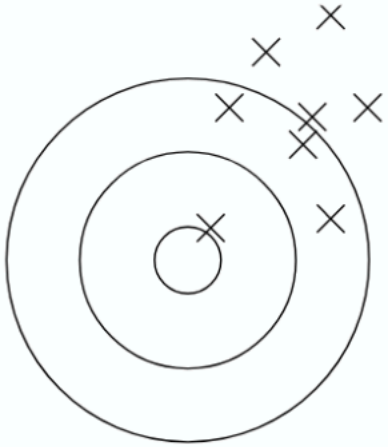
Example from Ameet Soni

Quiz: recap bias and variance



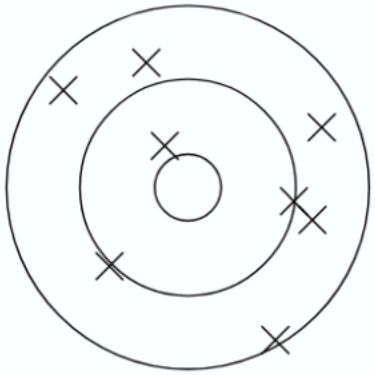
A

Variance: low
Bias: high



B

Variance: high
Bias: high



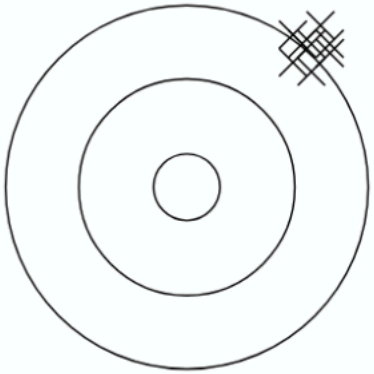
C

Variance: high
Bias: low

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

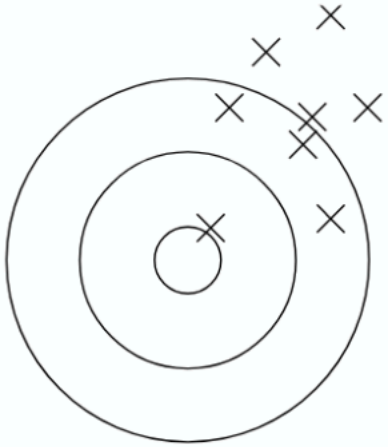
Example from Ameet Soni

Quiz: recap bias and variance



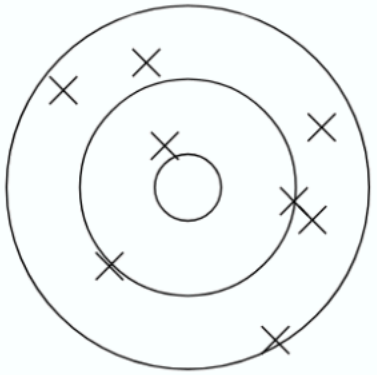
A

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

- 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 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

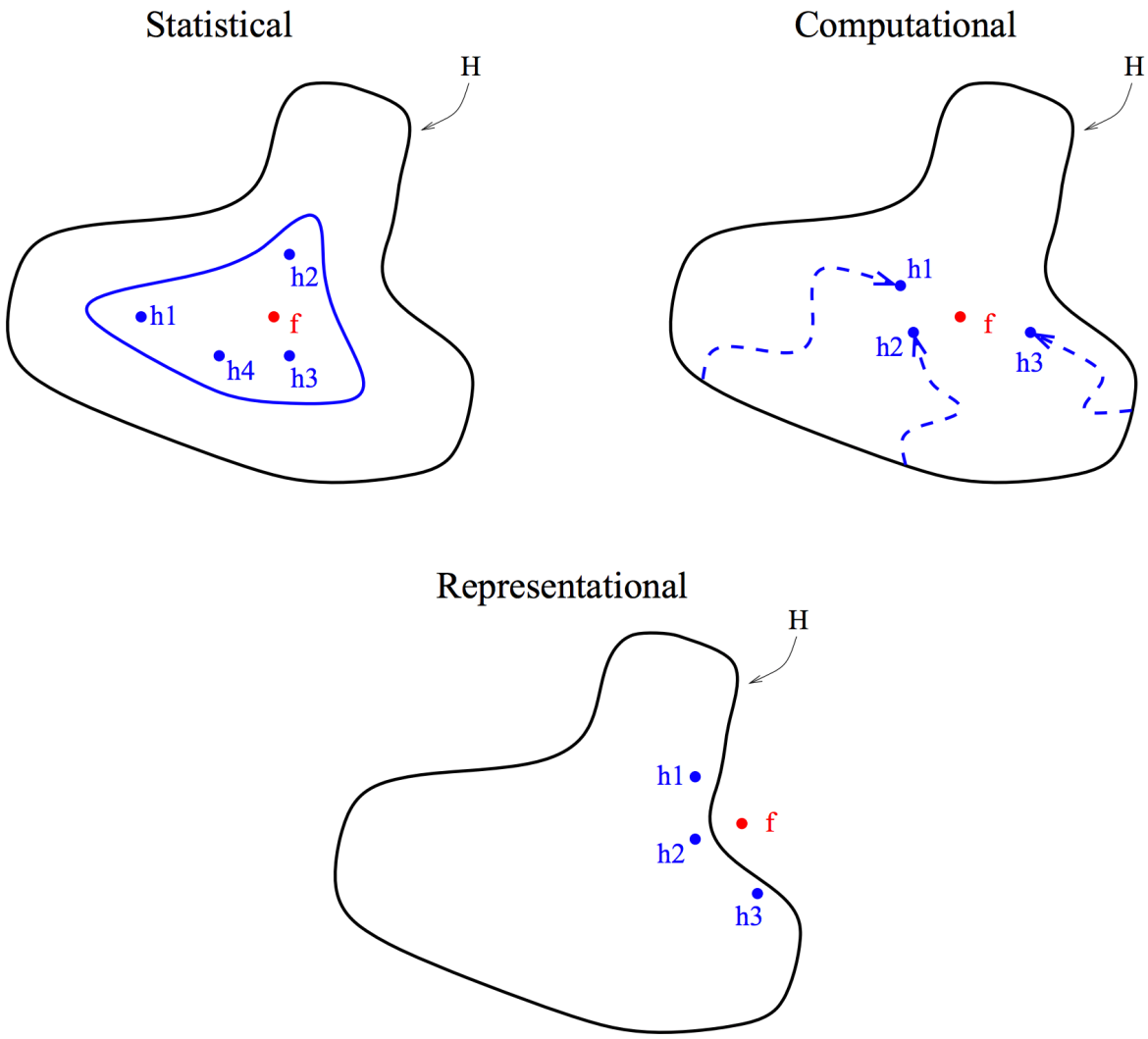


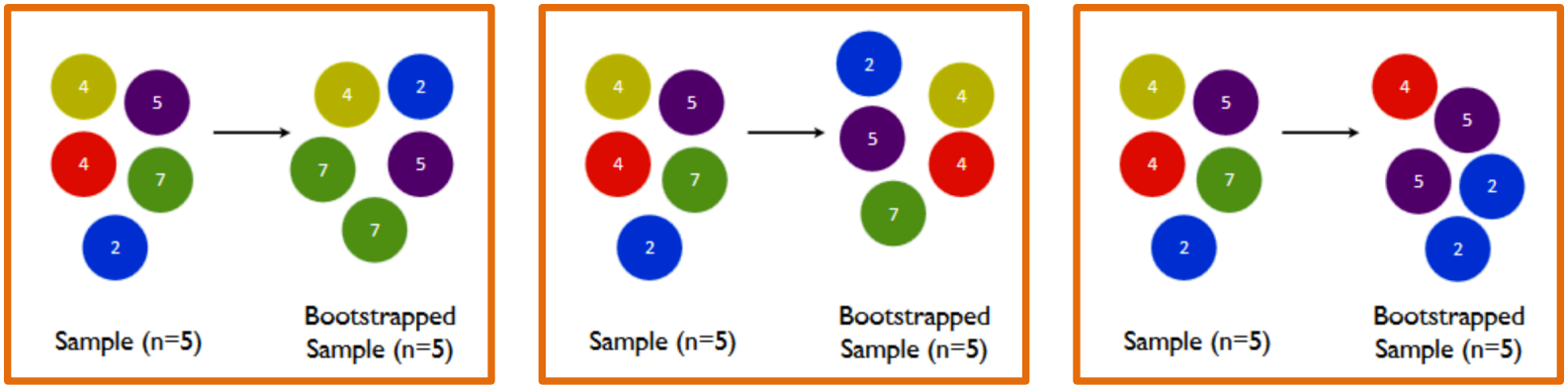
Figure from Tom Dietterich

Outline for October 20

- Recap multi-class logistic regression
- Introduction to ensemble methods
- Bagging
- Random forests
- AdaBoost

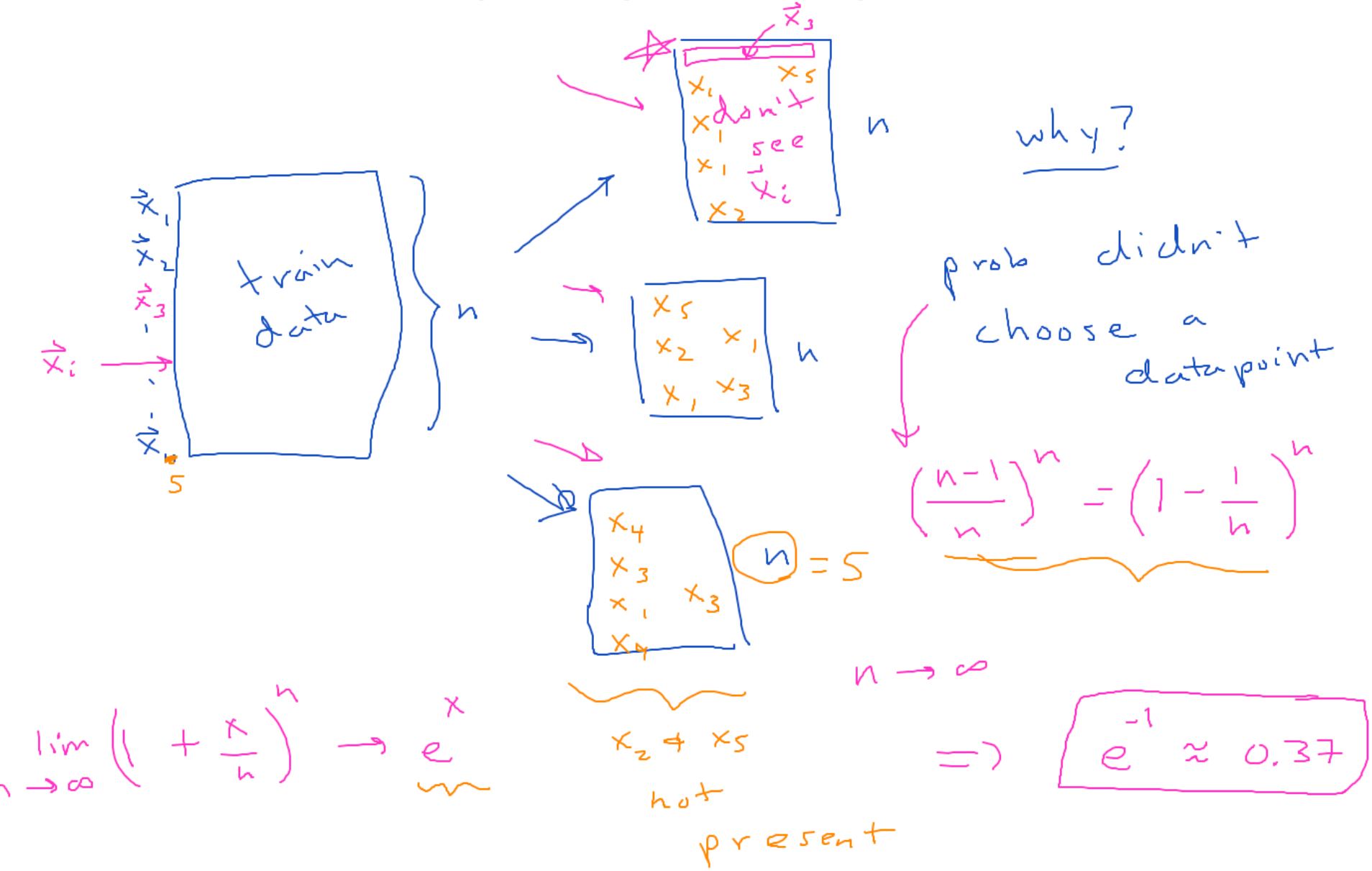
Bagging Algorithm

- 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

Bootstrap: sample with replacement



Notation for ensembles

- T = # models/classifiers (index t)
- x = test example (could be a vector)
- $X^{(t)}$ = bootstrap training dataset t
- $h^{(t)}(x)$ = hypothesis about x from model t $\in \{0, 1\}$ for binary problems
- r = probability of error of individual model
- R = # votes for wrong class

Bagging (Bootstrap Aggregation)

votes | Y
8 => 0
7 => 1

T = 15

h(x) = 0

Train:

for t in range(T):

* create bootstrap sample $X^{(t)}$ of size n from training data
* train on $X^{(t)}$ to get model $h^{(t)}$

Test:

for each test example x

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

Bagging: motivating example

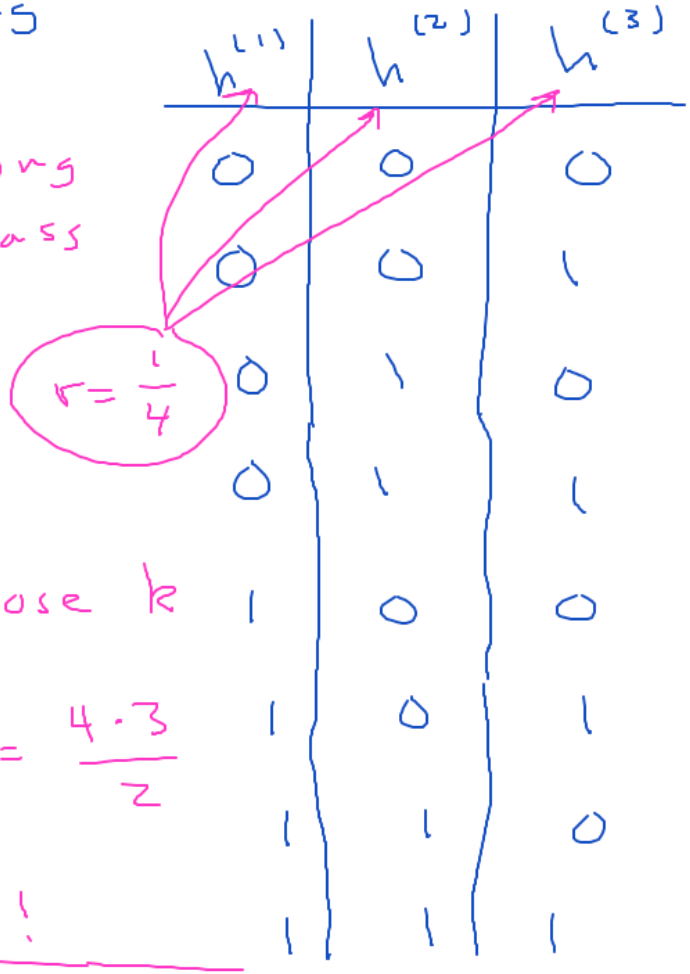
$T = 3$

test (x, y)

r = prob one model is wrong

R = # models that are wrong

$$R = \sum_{t=1}^{T} \mathbb{1}\left(h^{(t)}(x) = \tilde{y}\right)$$ ← wrong class



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

wrong · right

→ T choose k

incorrect overall

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

$$\binom{4}{2} = \frac{4 \cdot 3}{2}$$

more than half models wrong

$$\binom{T}{k} = \frac{T!}{k!(T-k!)}$$

Bagging: motivating example

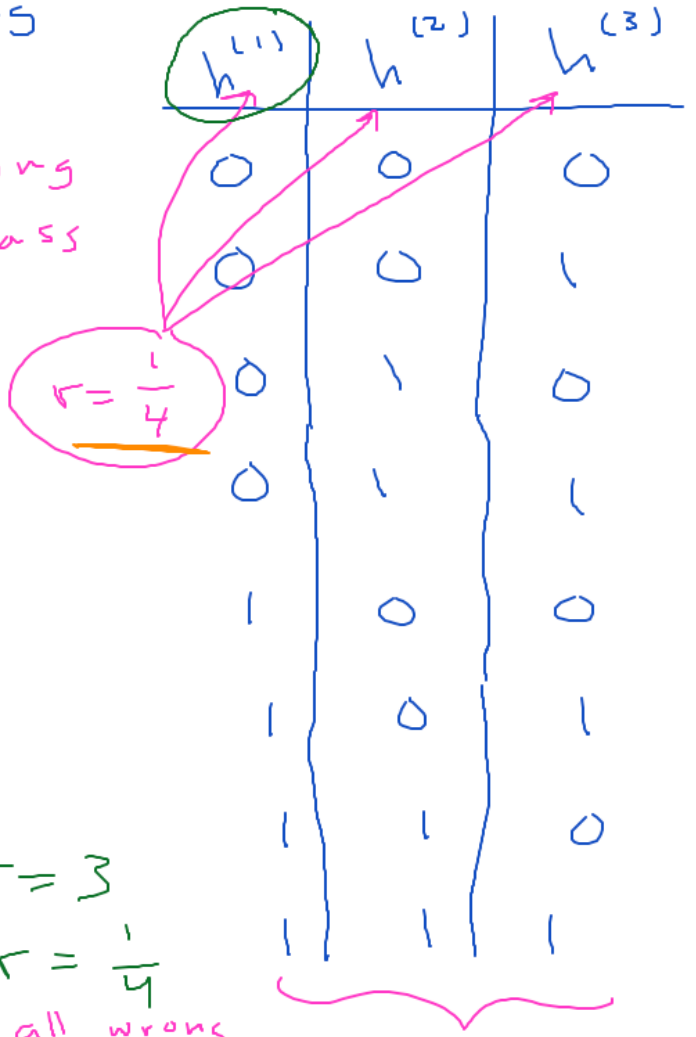
$T = 3$

test (x, y)

r = prob one model is wrong

R = # models that are wrong

$$R = \sum_{t=1}^{T} \mathbb{1}\left(h^{(t)}(x) = y\right)$$ ← wrong class



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

r^k: wrong; $(1-r)^{T-k}$: right

incorrect overall

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

more than half models wrong

Example $T = 3$, $r = \frac{1}{4}$

$$P(R > 1.5) = 3\left(\frac{1}{4}\right)^{2}\left(\frac{3}{4}\right) + \left(\frac{1}{4}\right)^{3}$$

2/3 wrong; all wrong

$$= 0.16$$ — 84% accuracy!

Outline for October 20

- Recap multi-class logistic regression
- Introduction to ensemble methods
- Bagging
- Random forests
- AdaBoost

Random Forests

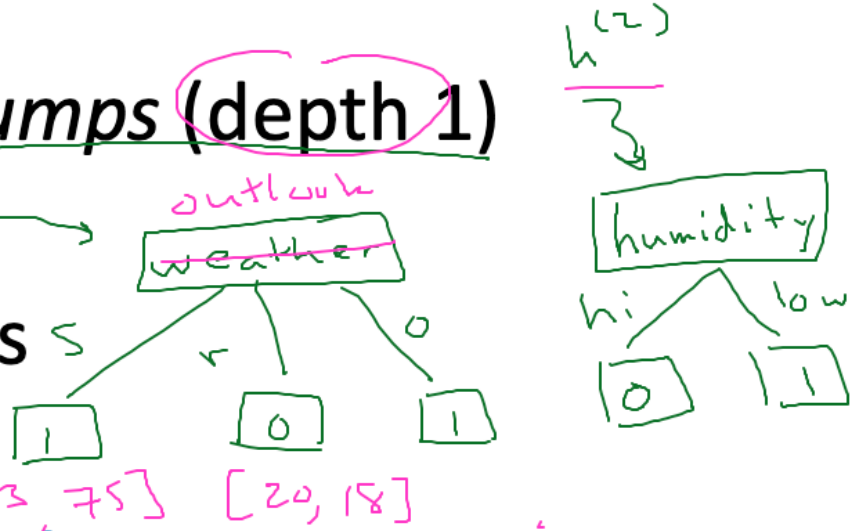
- Idea: choose a different **subset of features** for every classifier *t*
- Typically use *decision stumps* (depth 1)

- Goal: decorrelate models

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

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 October 20

- Recap multi-class logistic regression
- Introduction to ensemble methods
- Bagging
- Random forests
- AdaBoost

NEXT TIME!