CS 360: Machine Learning

Prof. Sara Mathieson

Fall 2019

Admin

- Office hours **TODAY** 12:30-1:30pm

- **Midterm 1 Thursday** (in lab + take home)
  - In-lab: may use self-created "cheat-sheet" + calculator
  - Take-home due Tuesday by 6pm (in my office)

- **Lab 5 due October 22** (Tuesday after fall break)

- No reading quiz this Thurs (continue to review for the midterm and start logistic regression)
- **Lab 2** back before the midterm (hopefully)

Why do we have a exam?

- Process of synthesizing the material on your own is essential
- Preparing the “cheat-sheet” is designed to facilitate that process
- Take-home allows you to demonstrate your understanding (and see the material in a new way) in a more relaxed setting
- If we only had a take-home, might be less likely to study things not on the take-home ☺

Real-world example of Naïve Bayes

"A Comparison of Event Models for Naive Bayes Text Classification" (4097 citations!)

http://www.cs.cmu.edu/~knigam/papers/multinomial-aaaiws98.pdf

Goal: text classification (classify documents into topics based on the words as features)

Informal quiz: discuss with a partner

- Identify the evidence, prior, posterior, and likelihood in the equation below

$$p(y = k|\boldsymbol{x}) = \frac{p(y = k)p(\boldsymbol{x}|y = k)}{p(\boldsymbol{x})}$$

Bonus question: where have we seen $p(y = k \,|\boldsymbol{x})$ before)?

Components of a Bayesian Model

- Identify the evidence, prior, posterior, and likelihood in the equation below

$$p(y = k|\boldsymbol{x}) = \frac{p(y = k)p(\boldsymbol{x}|y = k)}{p(\boldsymbol{x})}$$

- **Evidence**: this is the data (features) we actually observe, which we think will help us predict the outcome we're interested in

Components of a Bayesian Model

- Identify the evidence, prior, posterior, and likelihood in the equation below

$$p(y = k|\boldsymbol{x}) = \frac{p(y = k)p(\boldsymbol{x}|y = k)}{p(\boldsymbol{x})}$$

- **Prior**: without seeing any evidence (data), what is our prior believe about each outcome (intuition: what is the outcome in the population as a whole?)

Components of a Bayesian Model

- Identify the evidence, prior, posterior, and likelihood in the equation below

$$p(y = k|\boldsymbol{x}) = \frac{p(y = k)p(\boldsymbol{x}|y = k)}{p(\boldsymbol{x})}$$

- **Posterior**: this is the quantity we are actually interested in. **Given** the evidence, what is the probability of the outcome?

Components of a Bayesian Model

- Identify the evidence, prior, posterior, and likelihood in the equation below

$$p(y = k|\boldsymbol{x}) = \frac{p(y = k)p(\boldsymbol{x}|y = k)}{p(\boldsymbol{x})}$$

- **Likelihood**: given an outcome, what is the probability of observing this set of features?

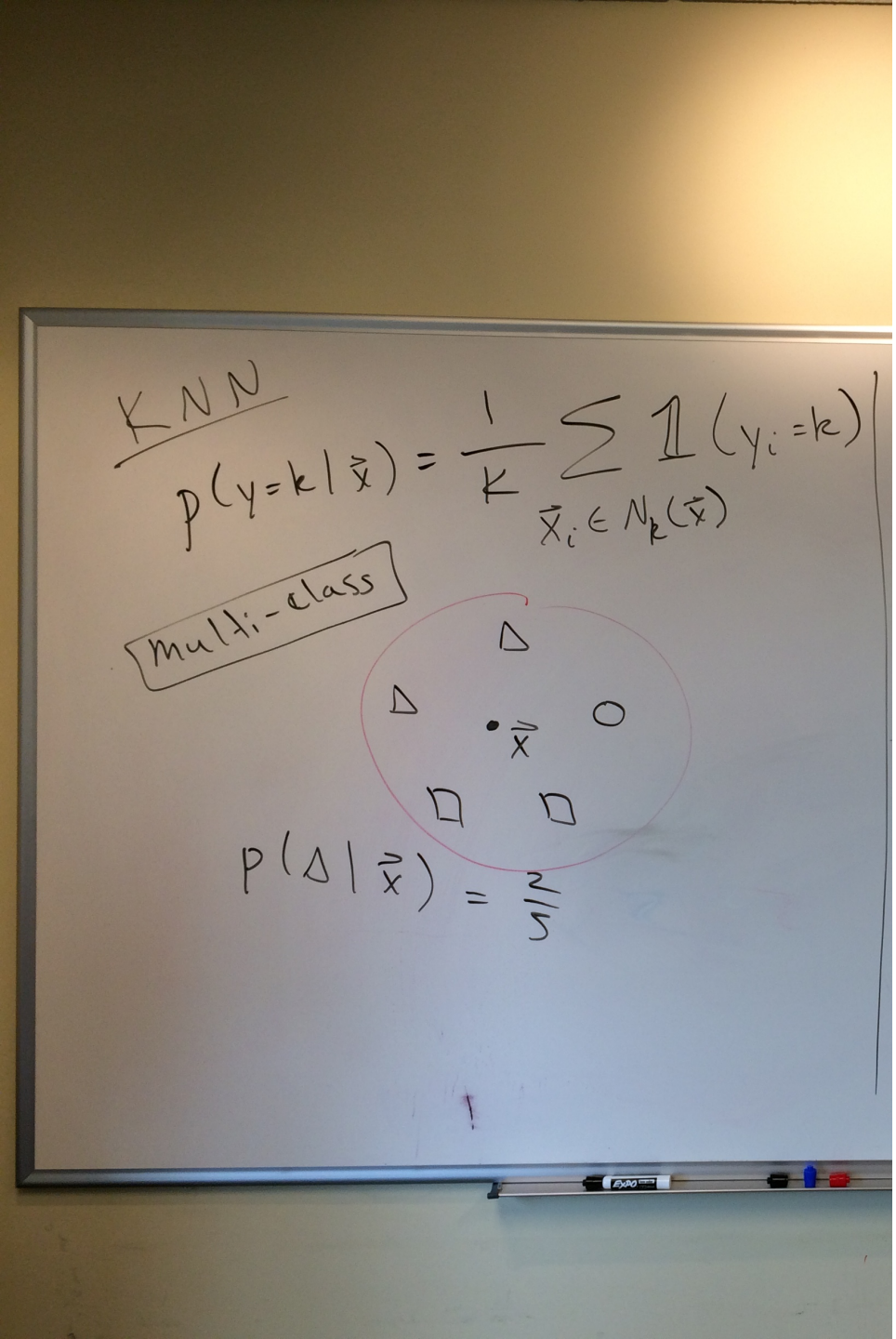
Bonus

- Identify the evidence, prior, posterior, and likelihood in the equation below

$$p(y = k|\boldsymbol{x}) = \frac{p(y = k)p(\boldsymbol{x}|y = k)}{p(\boldsymbol{x})}$$

Bonus question: where have we seen $p(y = k \,|\boldsymbol{x})$ before)?

KNN!



Bonus question: where have we seen $p(y = k \mid \boldsymbol{x})$ before)?

Naïve Bayes
KNN
Decision Trees

Outline for October 1

- Review
  - Lab 3 solutions
  - Entropy
  - Confusion matrices
  - SGD
  - Polynomial regression
  - Naïve Bayes
  - Loss functions & bias-variance tradeoff

- Begin: Logistic Regression

Outline for October 1

- Review
  - Lab 3 solutions
  - Entropy
  - Confusion matrices
  - SGD
  - Polynomial regression
  - Naïve Bayes
  - Loss functions & bias-variance tradeoff
- Begin: Logistic Regression

Review: overfitting

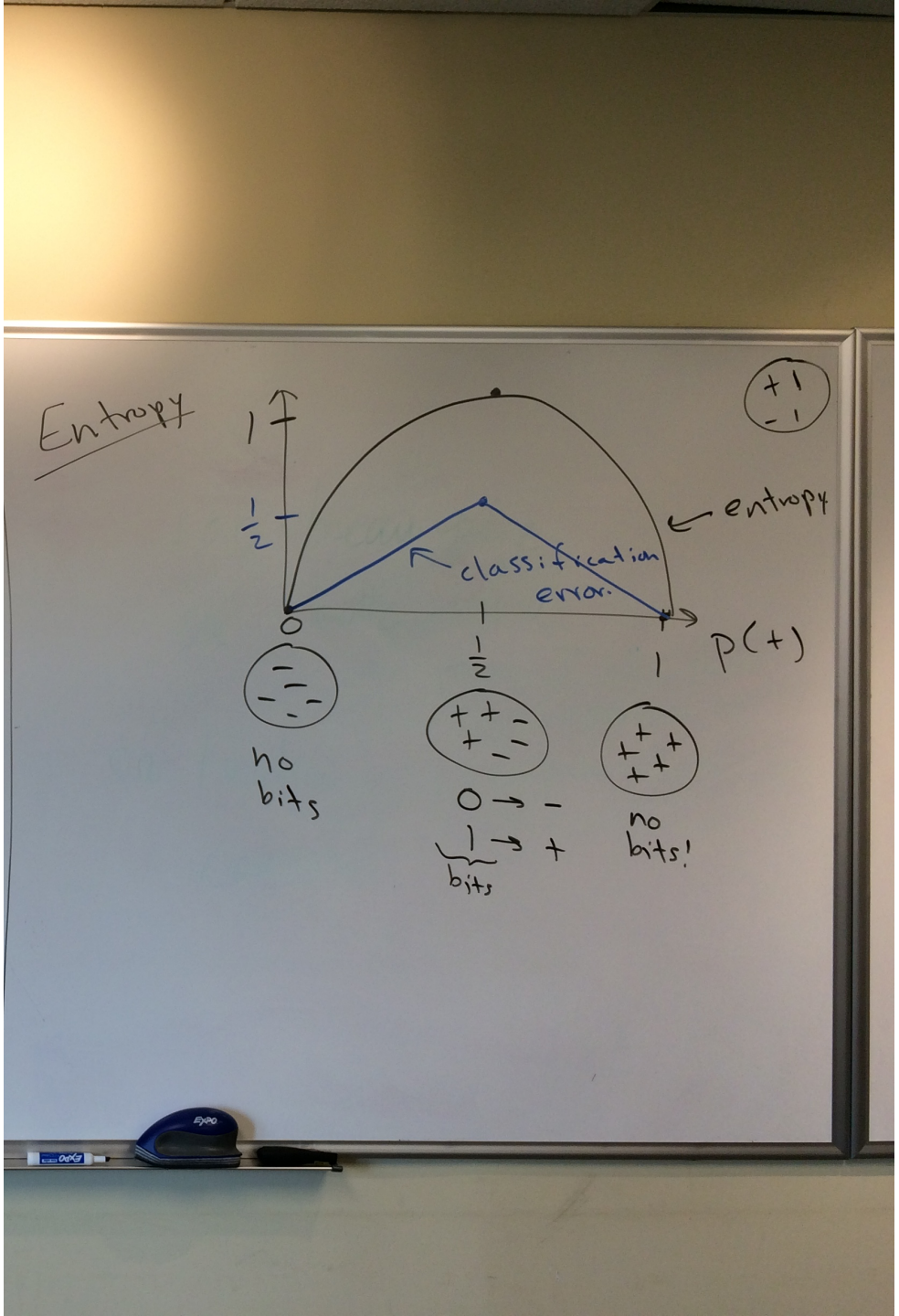
- Consider a hypothesis: *h*
  - Training error: $error_{train}(h)$
  - Error over all possible data: $error_{D}(h)$

- A hypothesis *h* **overfits** training data if there exists another hypothesis *h'* s.t.
  - $error_{train}(h) < error_{train}(h')$ AND
  - $error_{D}(h) > error_{D}(h')$

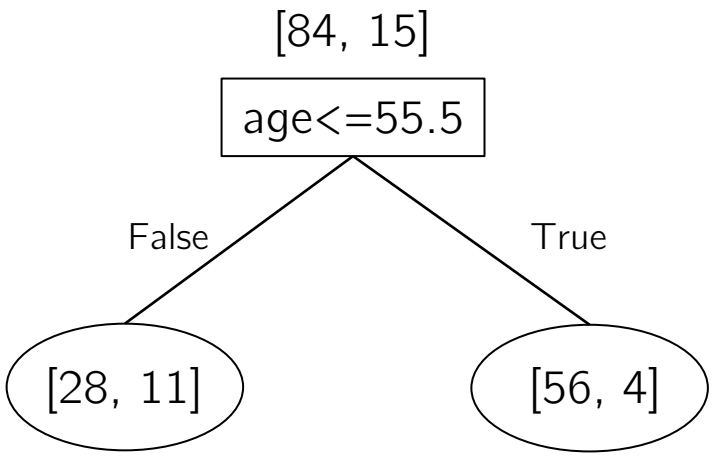
Lab 3 Solution
(not posted online)

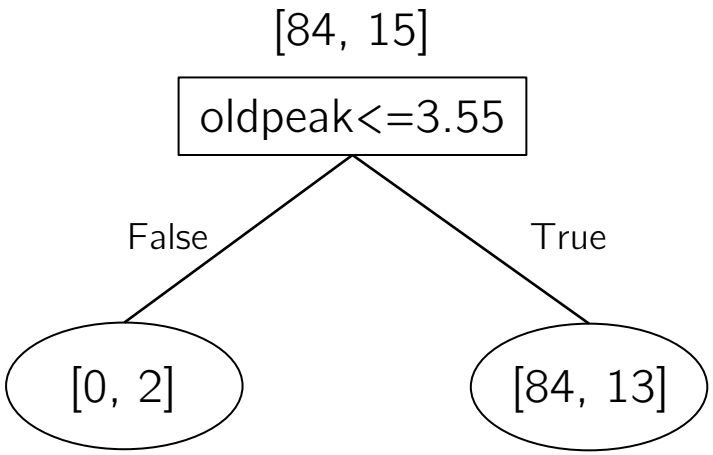
Outline for October 1

- Review
  - Lab 3 solutions
  - Entropy
  - Confusion matrices
  - SGD
  - Polynomial regression
  - Naïve Bayes
  - Loss functions & bias-variance tradeoff
- Begin: Logistic Regression


Entropy
1
1/2
0
1/2
1
p(+)
entropy
classification error
no bits
0 → -
1 → +
1 bits
no bits!

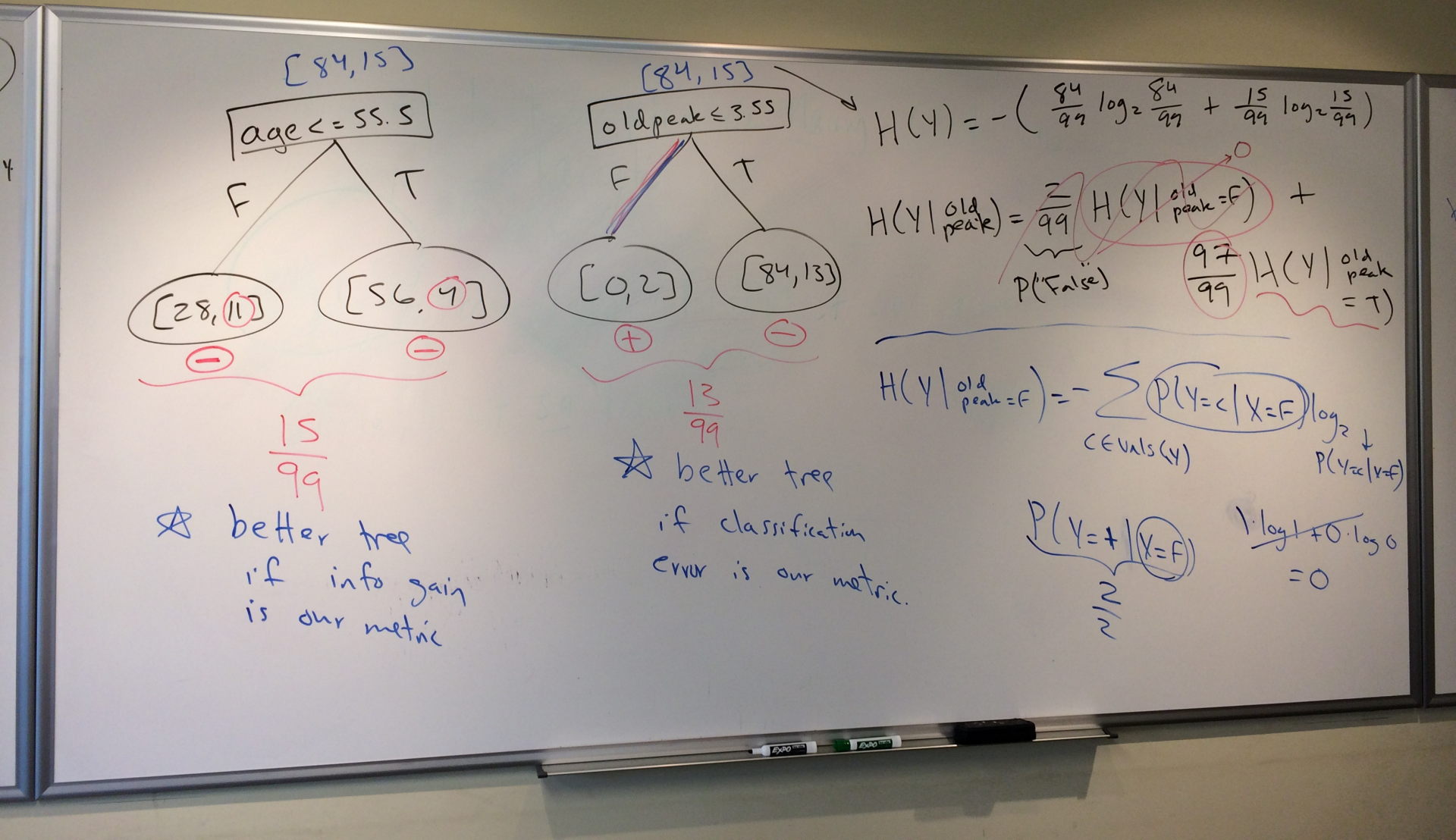
Decision Trees: information gain vs. classification error



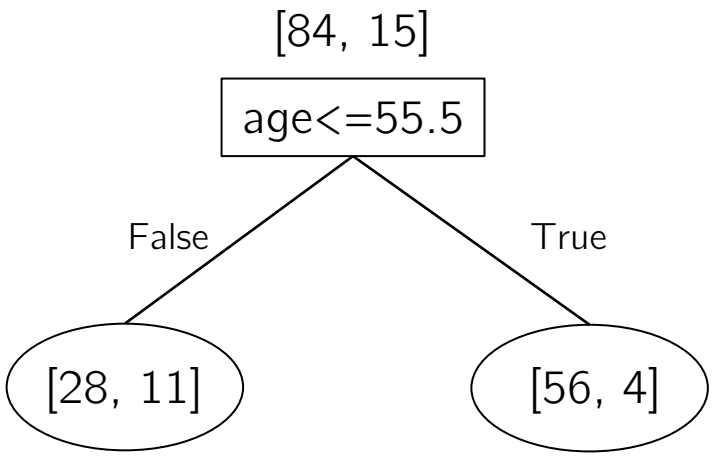


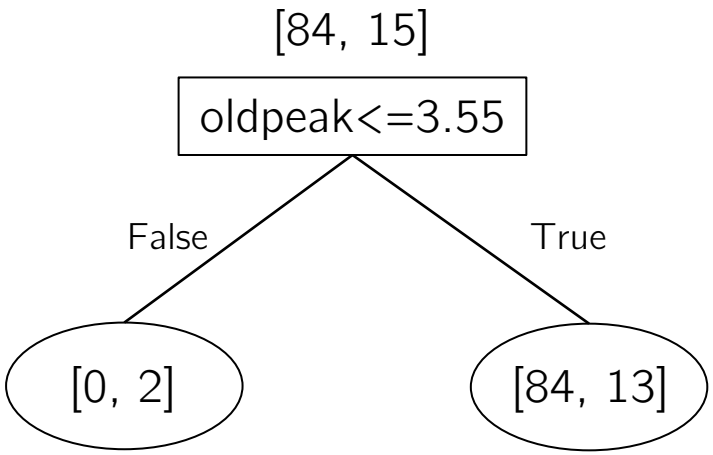
Work on first page of review problems with a partner/group

Handout 9, Question 1



Decision Trees: information gain vs. classification error





H(Y) = 0.6136190195993708

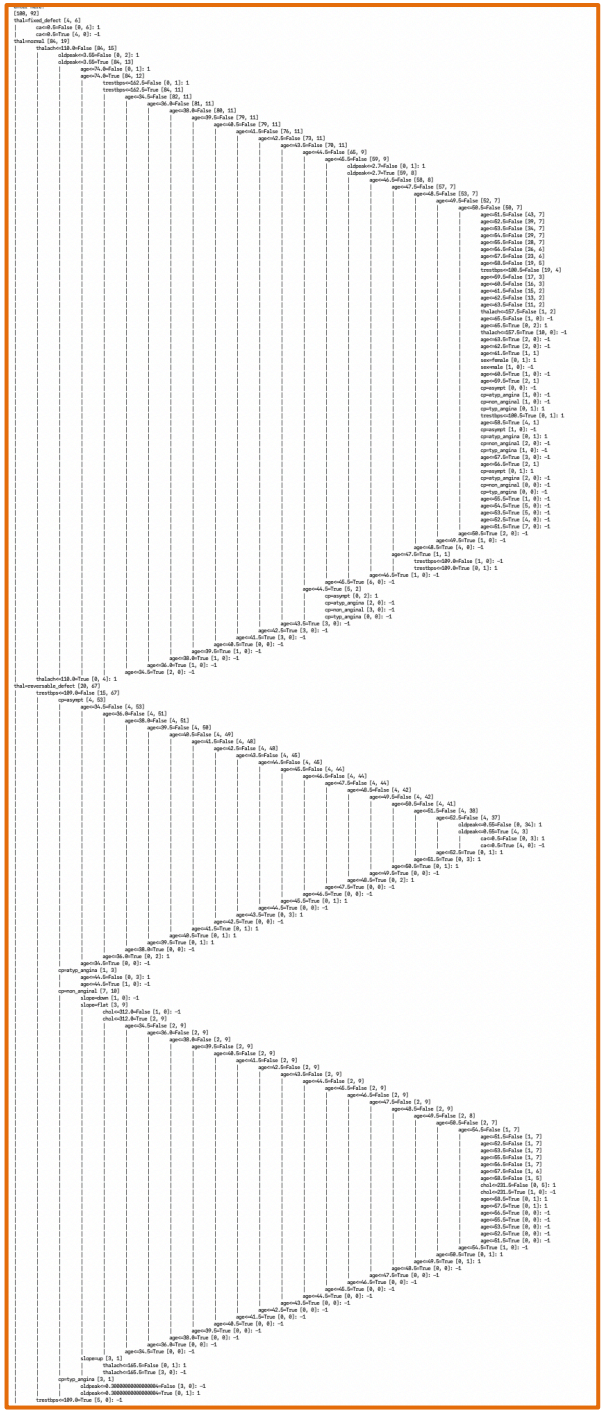
H(Y|age<=55.5) = 0.552248091053432 2

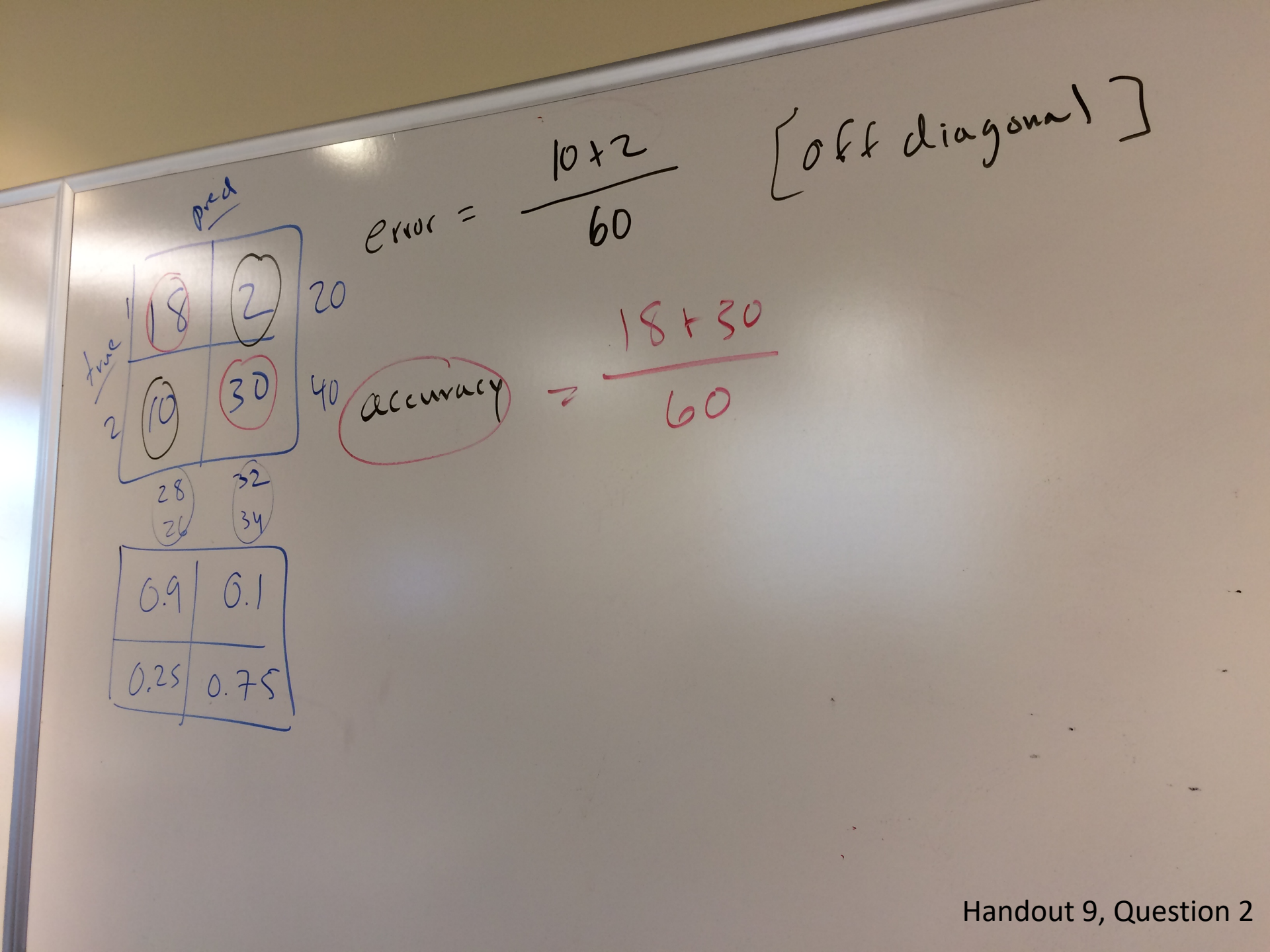
H(Y|oldpeak<=3.55) = 0.556880463059609 3

=> Age feature produces more information gain!

Decision trees from entropy (info gain) vs. classification error! (unlimited depth)

```
[108, 92]
thal=fixed_defect [4, 6]
|       ca<=0.5=False [0, 6]: 1
|       ca<=0.5=True [4, 0]: -1
thal=normal [84, 19]
|       thalach<=110.0=False [84, 15]
|       |       age<=55.5=False [28, 11]
|       |       |       chol<=248.5=False [14, 10]
|       |       |       |       sex=female [13, 3]
|       |       |       |       |       cp=asympt [3, 3]
|       |       |       |       |       |       age<=57.5=False [1, 3]
|       |       |       |       |       |       |       chol<=337.5=False [1, 0]: -1
|       |       |       |       |       |       |       chol<=337.5=True [0, 3]: 1
|       |       |       |       |       |       age<=57.5=True [2, 0]: -1
|       |       |       |       |       cp=atyp_angina [2, 0]: -1
|       |       |       |       |       cp=non_anginal [7, 0]: -1
|       |       |       |       |       cp=typ_angina [1, 0]: -1
|       |       |       |       sex=male [1, 7]
|       |       |       |       |       age<=65.5=False [1, 2]
|       |       |       |       |       |       age<=66.5=False [0, 2]: 1
|       |       |       |       |       |       age<=66.5=True [1, 0]: -1
|       |       |       |       |       age<=65.5=True [0, 5]: 1
|       |       |       chol<=248.5=True [14, 1]
|       |       |       |       oldpeak<=2.7=False [0, 1]: 1
|       |       |       |       oldpeak<=2.7=True [14, 0]: -1
|       |       age<=55.5=True [56, 4]
|       |       |       trestbps<=113.5=False [47, 1]
|       |       |       |       oldpeak<=3.55=False [0, 1]: 1
|       |       |       |       oldpeak<=3.55=True [47, 0]: -1
|       |       |       trestbps<=113.5=True [9, 3]
|       |       |       |       oldpeak<=0.05=False [6, 0]: -1
|       |       |       |       oldpeak<=0.05=True [3, 3]
|       |       |       |       |       cp=asympt [0, 2]: 1
|       |       |       |       |       cp=atyp_angina [2, 0]: -1
|       |       |       |       |       cp=non_anginal [1, 1]
|       |       |       |       |       |       age<=41.5=False [0, 1]: 1
|       |       |       |       |       |       age<=41.5=True [1, 0]: -1
|       |       |       |       |       cp=typ_angina [0, 0]: -1
|       thalach<=110.0=True [0, 4]: 1
thal=reversable_defect [20, 67]
|       cp=asympt [5, 53]
|       |       oldpeak<=0.55=False [0, 43]: 1
|       |       oldpeak<=0.55=True [5, 10]
|       |       |       chol<=237.5=False [0, 8]: 1
|       |       |       chol<=237.5=True [5, 2]
|       |       |       |       chol<=179.5=False [4, 0]: -1
|       |       |       |       chol<=179.5=True [1, 2]
|       |       |       |       |       age<=59.5=False [1, 0]: -1
|       |       |       |       |       age<=59.5=True [0, 2]: 1
|       cp=atyp_angina [3, 3]
|       |       age<=46.5=False [1, 3]
|       |       |       trestbps<=109.0=False [0, 3]: 1
|       |       |       trestbps<=109.0=True [1, 0]: -1
|       |       age<=46.5=True [2, 0]: -1
|       cp=non_anginal [9, 10]
|       |       oldpeak<=1.85=False [0, 5]: 1
|       |       oldpeak<=1.85=True [9, 5]
|       |       |       trestbps<=121.0=False [3, 5]
|       |       |       |       chol<=232.5=False [0, 4]: 1
|       |       |       |       chol<=232.5=True [3, 1]
|       |       |       |       |       trestbps<=128.5=False [3, 0]: -1
|       |       |       |       |       trestbps<=128.5=True [0, 1]: 1
|       |       |       trestbps<=121.0=True [6, 0]: -1
|       cp=typ_angina [3, 1]
|       |       oldpeak<=0.30000000000000004=False [3, 0]: -1
|       |       oldpeak<=0.30000000000000004=True [0, 1]: 1
```

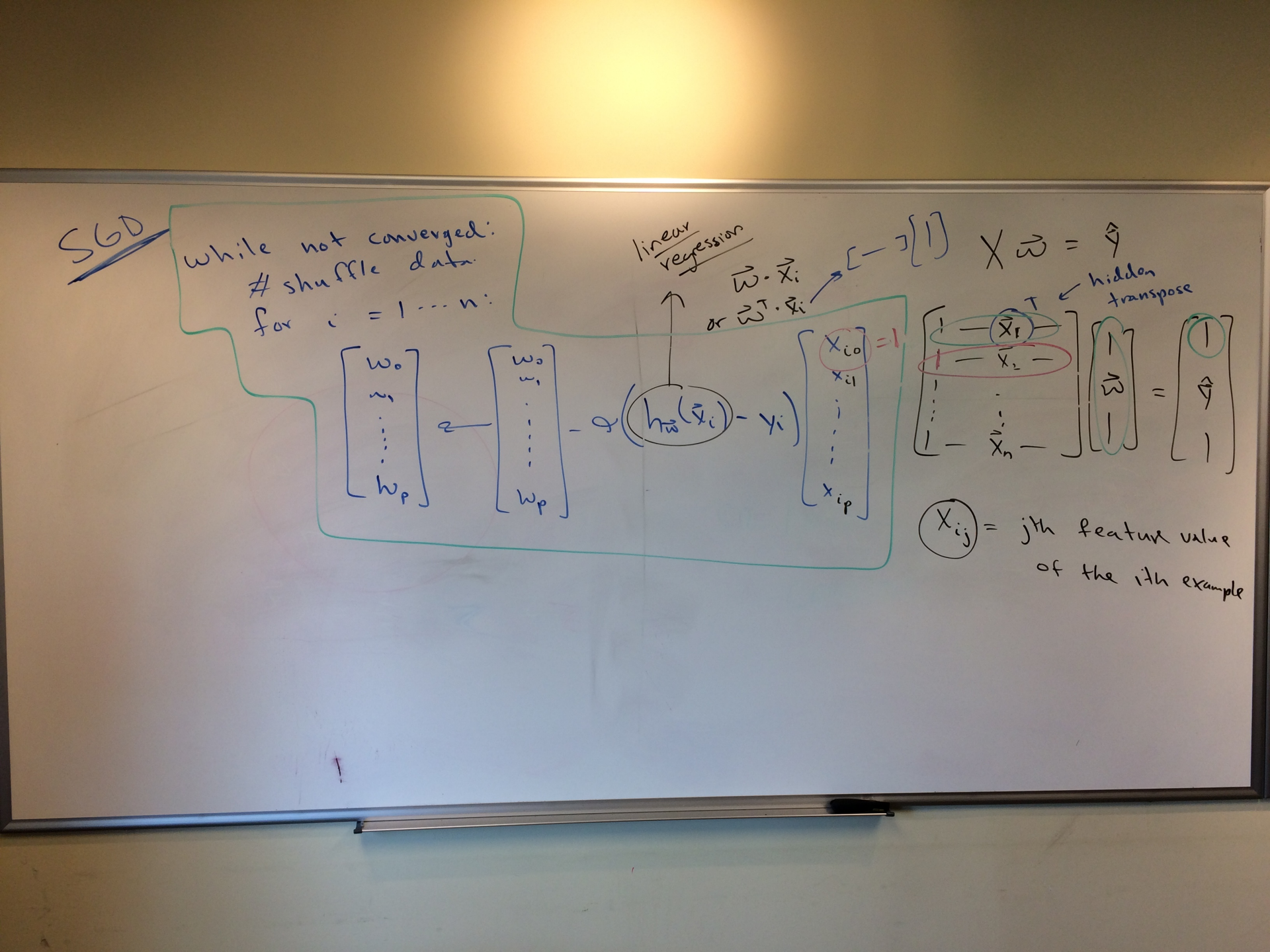




Handout 9, Question 2

Outline for October 1

- Review
  - Lab 3 solutions
  - Entropy
  - Confusion matrices
  - SGD
  - Polynomial regression
  - Naïve Bayes
  - Loss functions & bias-variance tradeoff
- Begin: Logistic Regression


SGD
while not converged:
# shuffle data
for i = 1 ... n:
linear regression
$\vec{w} \cdot \vec{x}_i$
or $\vec{w}^T \cdot \vec{x}_i$
$X\vec{w} = \hat{y}$
hidden
transpose
T
$\vec{x}_1$
$\vec{x}_2$
$\vec{x}_n$
$\vec{w}$
$\hat{y}$
x_{ij} = jth feature value of the ith example

polynomial regression

$p = 1$

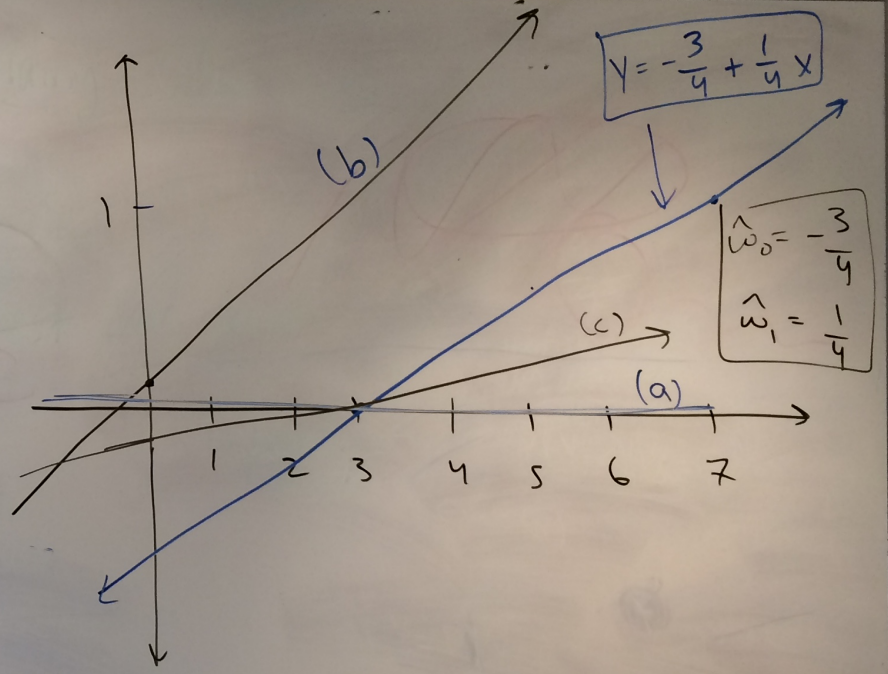
deg: d

$$h_{\vec{w}}(\vec{x}) = w_0 x^0 + w_1 x^1 + w_2 x^2 + \cdots w_d x^d$$

$$\phi(x) = \begin{bmatrix} x^0 \\ x^1 \\ \vdots \\ x^d \end{bmatrix}$$

change of basis

$$\Phi = \begin{bmatrix} - & \phi(x_1)^T & - \\ - & \phi(x_2) & - \\ & \vdots & \\ - & \phi(x_n) & - \end{bmatrix}$$



$\hat{w}_0 = -\frac{3}{4}$

$\hat{w}_1 = \frac{1}{4}$

$(x_2, y_2) = (7, 1)$

$$\begin{bmatrix} 0.1 \\ 0.7 \end{bmatrix} \leftarrow \begin{bmatrix} 0 \\ 0 \end{bmatrix} - 0.1\,(0 - 1) \begin{bmatrix} 1 \\ 7 \end{bmatrix}$$

$\begin{bmatrix} 0 & 0 \end{bmatrix}$ $\vec{w}^T$ $\begin{bmatrix} 1 \\ 7 \end{bmatrix}$

y_2

weights from prev datapoint

scalar

$\vec{x}_2$

$(x_1, y_1) = (3, 0)$

$$\begin{bmatrix} -0.12 \\ 0.04 \end{bmatrix} \leftarrow \begin{bmatrix} 0.1 \\ 0.7 \end{bmatrix} - 0.1 \left(\begin{bmatrix} 0.1 & 0.7 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \end{bmatrix} - 0 \right) \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

$\begin{bmatrix} 1 \\ 3 \\ 8 \\ \vdots \end{bmatrix}$

$\vec{x}_i$ $\vec{x}_i$

Handout 9, Question 3

Handout 9 last questions

3d:

1	3	9
1	7	49

4a: depth

4b: degree

Outline for October 1

- Review
  - Lab 3 solutions
  - Entropy
  - Confusion matrices
  - SGD
  - Polynomial regression
  - Naïve Bayes
  - Loss functions & bias-variance tradeoff

Next Time!

- Begin: Logistic Regression