CS 260: Foundations of Data Science

Prof. Sara Mathieson

Fall 2023

HAVERFORD
COLLEGE

Admin

- **Lab 6** posted (Information Theory)
  - Due next Wednesday Nov 1
- **Lab 4** grades up soon

Outline for October 26

- Continuous features
- Introduction to logistic regression
- Cost function and SGD for logistic regression
- Connection to cross entropy

Outline for October 26

- Continuous features
- Introduction to logistic regression
- Cost function and SGD for logistic regression
- Connection to cross entropy

Continuous Features

(do this for the TRAIN only!)

X	Y
10	Y
7	Y
8	N
3	Y
7	N
12	Y
2	Y

1) Sort examples based on given feature

2	3	7	7	8	10	12
Y	Y	Y	N	N	Y	Y

Continuous Features

(do this for the TRAIN only!)

X	Y
10	Y
7	Y
8	N
3	Y
7	N
12	Y
2	Y

1) Sort examples based on given feature

2	3	7	7	8	10	12
Y	Y	Y	N	N	Y	Y

2) Different label with same feature value, collapse to "None"

2	3	7	8	10	12
Y	Y	None	N	Y	Y

Continuous Features

(do this for the TRAIN only!)

X	Y
10	Y
7	Y
8	N
3	Y
7	N
12	Y
2	Y

1) Sort examples based on given feature

2	3	7	7	8	10	12
Y	Y	Y	N	N	Y	Y

2) Different label with same feature value, collapse to "None"

2	3	7	8	10	12
Y	Y	None	N	Y	Y

3) Whenever label changes, make a feature (use avg)

2	3	7	8	10	12
Y	Y	None	N	Y	Y

X <= 5

X <= 7.5

X <= 9

discrete $x \geq 5$	continuous feature	label
T	10	Y
T	7	Y
T	8	N
F	3	Y
T	7	N
T	12	Y
F	2	Y

$x \geq 7.5$

$x \geq 9$

①

2	3	7	7	8	10	12
Y	Y	Y	N	N	Y	Y

②

2	3	7	8	10	12
Y	Y	None	N	Y	Y

$\boxed{x \geq 5}$ $\boxed{x \geq 7.5}$ $\boxed{x \geq 9}$

Continuous Features (Handout 14)

(do this for the TRAIN only!)

temp	Y
80	Y
48	Y
60	N
48	Y
40	N
48	Y
90	Y

1) Sort examples based on feature “temp”

2) Different label with same feature value, collapse to “None”

3) Whenever label changes, make a feature (use avg)

Outline for October 26

- Continuous features
- Introduction to logistic regression
- Cost function and SGD for logistic regression
- Connection to cross entropy

Why is linear regression a bad choice for classification?

Case Study: you need to identify the medical condition of a patient in the emergency room on the basis of their symptoms.

Possible conditions (y) are:

- Stroke
- Drug overdose
- Epileptic seizure

1) If you were forced to use linear regression for this problem, how could you encode y to make it real-valued?

2) What issues arise with making y real-valued?

3) What if you just had two outcomes (i.e. stroke and drug overdose) -- why is linear regression still not a good choice?

Why is linear regression a bad choice for classification?

Case Study: you need to identify the medical condition of a patient in the emergency room on the basis of their symptoms.

Possible conditions (y) are:

- Stroke
- Drug overdose
- Epileptic seizure

1) If you were forced to use linear regression for this problem, how could you encode y to make it real-valued?

You could choose stroke=0, drug overdose=1, epileptic seizure=2 (or some permutation)

2) What issues arise with making y real-valued?

3) What if you just had two outcomes (i.e. stroke and drug overdose) -- why is linear regression still not a good choice?

Why is linear regression a bad choice for classification?

Case Study: you need to identify the medical condition of a patient in the emergency room on the basis of their symptoms.

Possible conditions (y) are:
- Stroke
- Drug overdose
- Epileptic seizure

1) If you were forced to use linear regression for this problem, how could you encode y to make it real-valued?

You could choose stroke=0, drug overdose=1, epileptic seizure=2 (or some permutation)

2) What issues arise with making y real-valued?

Assumes some *ordering* of the outcomes that is probably not there!

3) What if you just had two outcomes (i.e. stroke and drug overdose) -- why is linear regression still not a good choice?

Why is linear regression a bad choice for classification?

Case Study: you need to identify the medical condition of a patient in the emergency room on the basis of their symptoms.

Possible conditions (y) are:
- Stroke
- Drug overdose
- Epileptic seizure

1) If you were forced to use linear regression for this problem, how could you encode y to make it real-valued?

You could choose stroke=0, drug overdose=1, epileptic seizure=2 (or some permutation)

2) What issues arise with making y real-valued?

Assumes some *ordering* of the outcomes that is probably not there!

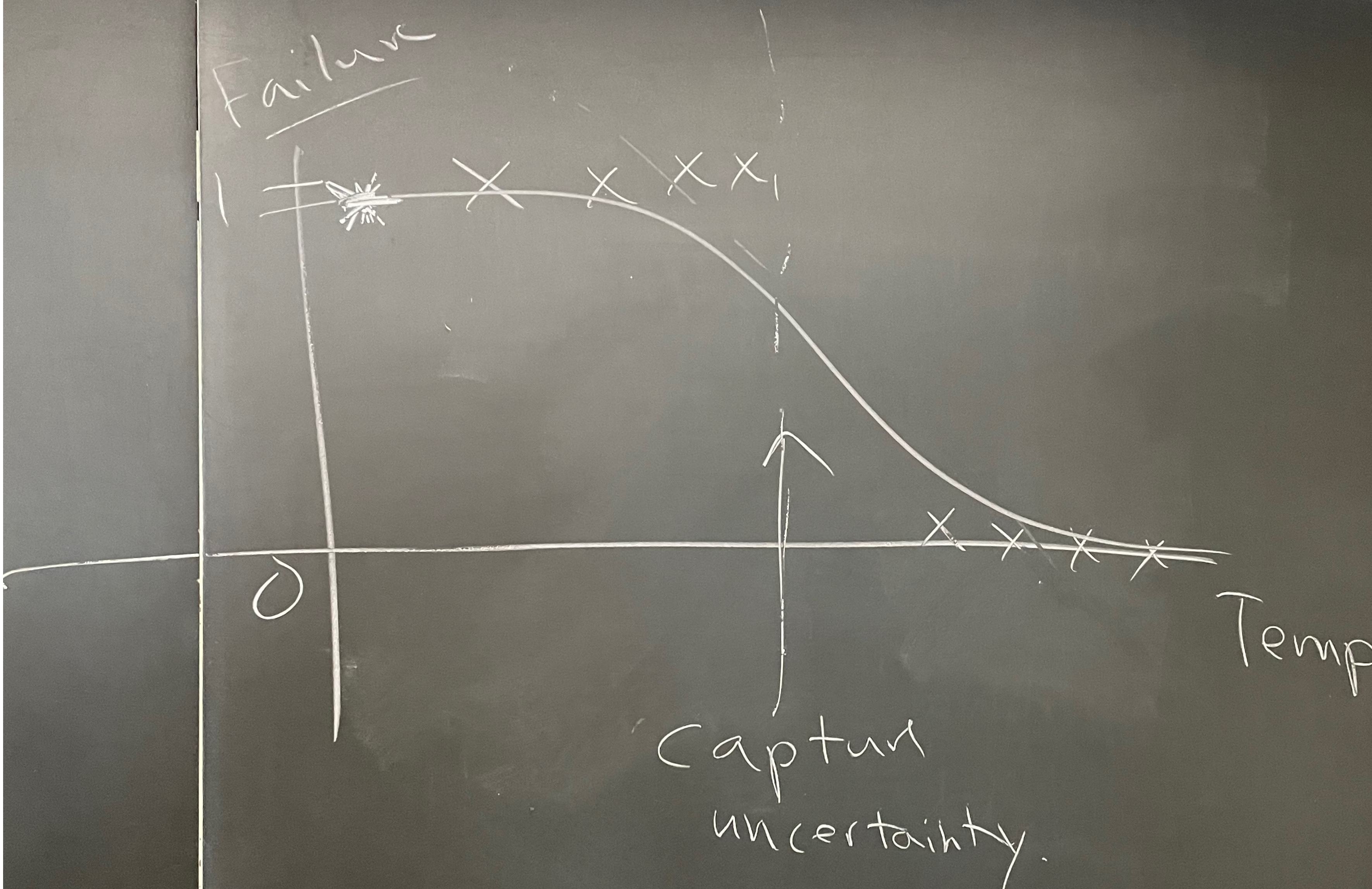
3) What if you just had two outcomes (i.e. stroke and drug overdose) -- why is linear regression still not a good choice?

The range of a linear function (i.e. y values) is $[-\infty, \infty]$, but we want $[0, 1]$

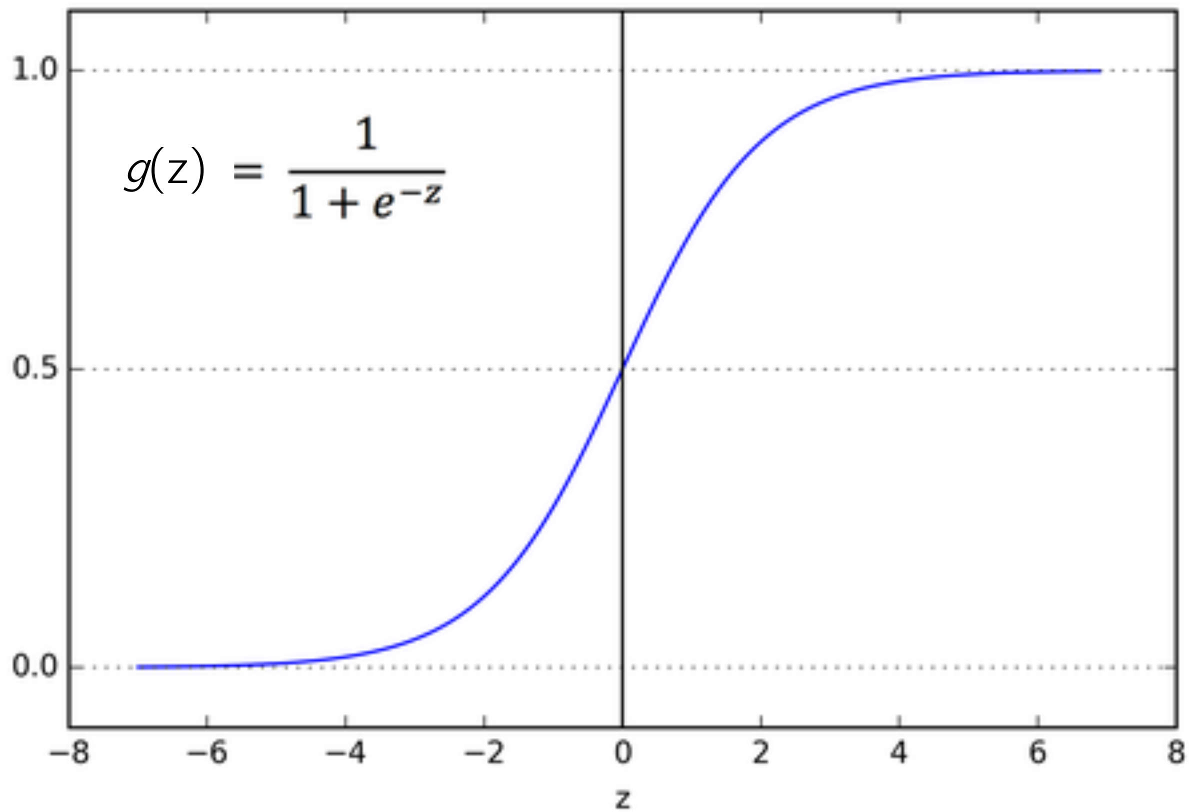
Challenger Explosion Data

Image: NASA

1	Date	Temperature	Damage Incident
2	04/12/1981	66	0
3	11/12/1981	70	1
4	3/22/82	69	0
5	6/27/82	80	NA
6	01/11/1982	68	0
7	04/04/1983	67	0
8	6/18/83	72	0
9	8/30/83	73	0
10	11/28/83	70	0
11	02/03/1984	57	1
⋮			
23	10/30/85	75	1
24	11/26/85	76	0
25	01/12/1986	58	1
26	1/28/86	31	Challenger Accident


Failure
1
0
Temp
Capture
uncertainty.

Logistic (sigmoid) function



Logistic Regression

binary classification $y \in \{0,1\}$

linear regression

$$\begin{array}{ccc} X & & Y \\ [-\infty, \infty] & \rightarrow & [-\infty, \infty] \end{array}$$

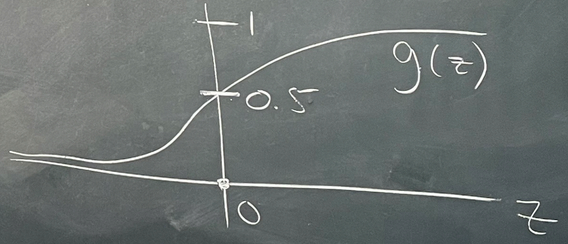
$$[-\infty, \infty] \rightarrow \underbrace{[0, 1]}_{\text{probability}}$$

idea

model will be:

$$h_{\vec{w}}(\vec{x}) = \underbrace{\frac{1}{1+e^{-\underbrace{\vec{w}\cdot\vec{x}}_{\text{linear model}}}}}_{\text{logistic function}}$$

Sigmoid / logistic



$z \rightarrow \infty$, $g(z) \rightarrow 1$

$z \rightarrow -\infty$, $g(z) \rightarrow 0$

$z = 0$, $g(z) = \frac{1}{2}$

$$g(z) = \frac{1}{1+e^{-z}}$$

→ classify 1

← classify 0

istic
function

$(z) \to 1$

$(z) \to 0$

$(z) = \frac{1}{2}$

already have $\vec{w}$ (model) pred

if $\underbrace{\vec{w} \cdot \vec{x}}_{= z} \geq 0 \Rightarrow \hat{y} = 1$

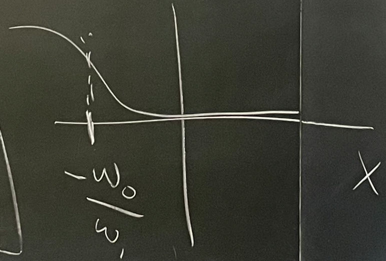
$\vec{w} \cdot \vec{x} < 0 \Rightarrow \hat{y} = 0$

↳ if $p = 1$ (one feature)

$w_0 + w_1 x \geq 0$ solve for x!

$w_1 x \geq -w_0$

$$\boxed{x \geq -\frac{w_0}{w_1}}$$

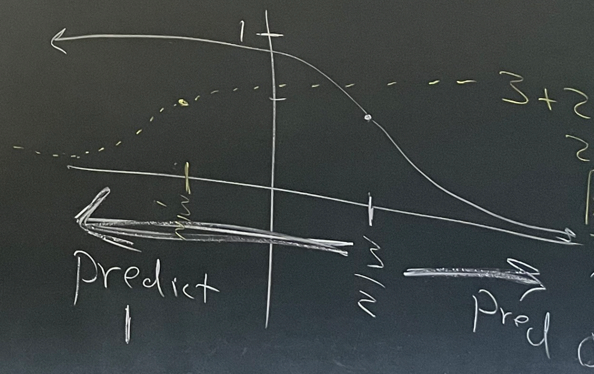


ex $\vec{w} = \begin{bmatrix} 3 \\ -2 \end{bmatrix} \begin{array}{l} \to w_0 \\ \to w_1 \end{array}$ Q? what is the decision boundary?

$$\boxed{3 - 2x \geq 0}$$ predict 1

$-2x \geq -3$

$$\boxed{x \leq \frac{3}{2}}$$ means $\hat{y} = 1$



$\vec{w} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$

$3 + 2x \geq 0$

$2x \geq -3$

$$\boxed{x \geq -\frac{3}{2}}$$

Outline for October 26

- Continuous features
- Introduction to logistic regression
- Cost function and SGD for logistic regression
- Connection to cross entropy

How to find $\vec{w}$?

need a cost function => use $\boxed{\text{SGD}}$!

$\boxed{\text{likelihood}}$

$\vec{y} = [0, 1, 1, 0, 0, 1]^T$

$$L(\vec{w}) = \prod_{i=1}^{n} \underbrace{h_{\vec{w}}(\vec{x}_i)}_{\text{prob of } 1}^{\,y_i = 1} \underbrace{\left(1 - h_{\vec{w}}(\vec{x}_i)\right)}_{\text{prob of } 0}^{\,1 - y_i = 0}$$

$$L(\vec{w}) = (1-h)\cdot h \cdot h \cdot (1-h)(1-h)\,h$$

$$\underbrace{(1 - h_{\vec{w}}(\vec{x}_0))}$$

$$(1 - P(x_0 = 1))$$

$X = [\;\;]$

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

Want high!

$$\log(a^b) = b\log(a)$$

ent

take log

cost want low

$$J(\vec{w}) = -\log(L(\vec{w}))$$

if $y=0$ if $y=1$

$$J(\vec{w}) = -\sum_{i=1}^{n}\left[y_i \log(h_{\vec{w}}(\vec{x}_i)) + (1-y_i)\log(1-h_{\vec{w}}(\vec{x}_i))\right]$$

$\boxed{\text{Single example } \vec{x}, y}$

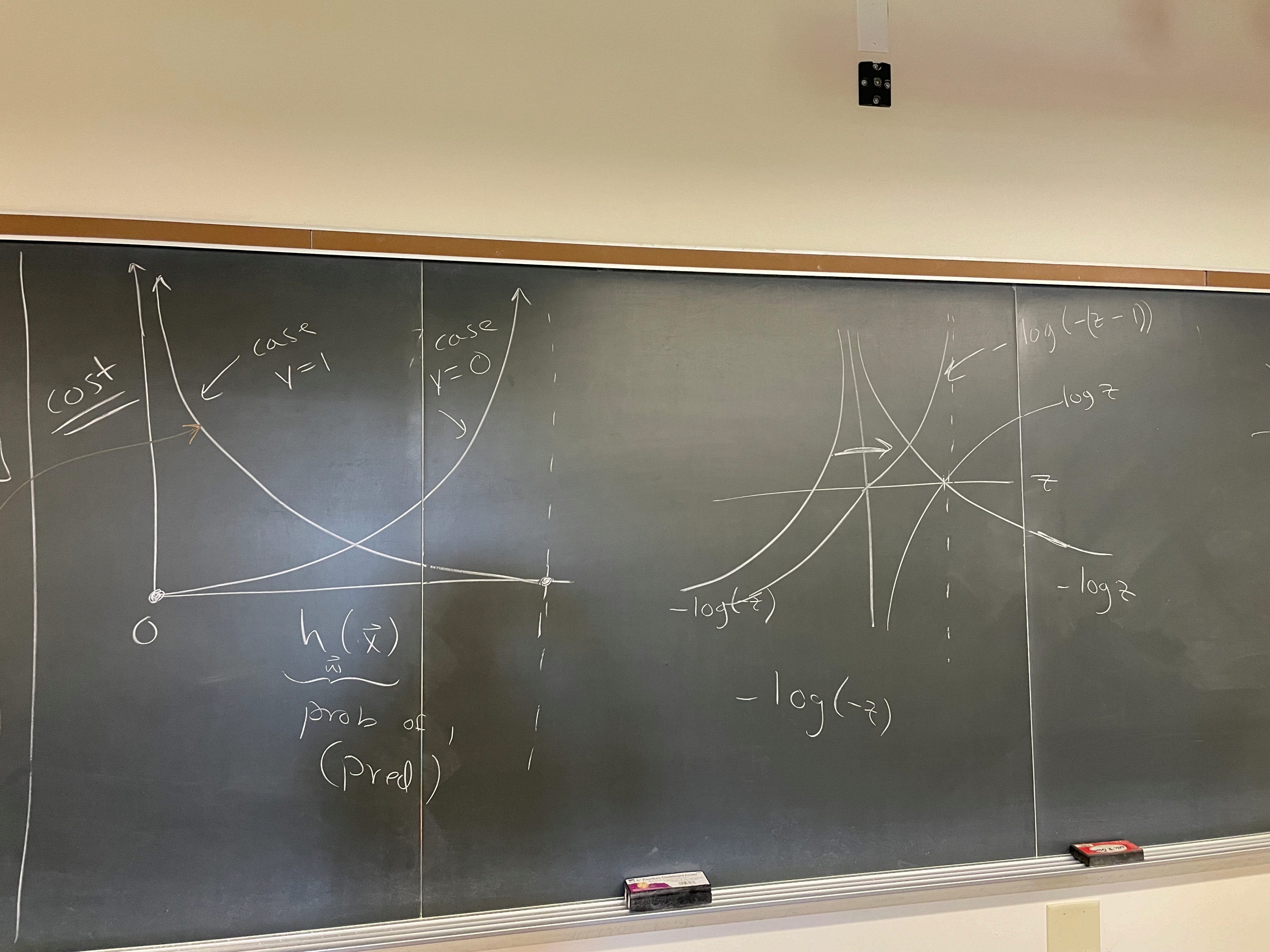
$$J(\vec{w}) = \begin{cases} -y\log h_{\vec{w}}(\vec{x}) = -\log \overbrace{h_{\vec{w}}(\vec{x})}^{z} & \text{if } y=1 \\ -(1-y)\log(1-h_{\vec{w}}(\vec{x})) = -\log(1-h(\vec{x})) & \text{if } y=0 \end{cases}$$

$$\boxed{\sum_x p_x \log p_x}$$

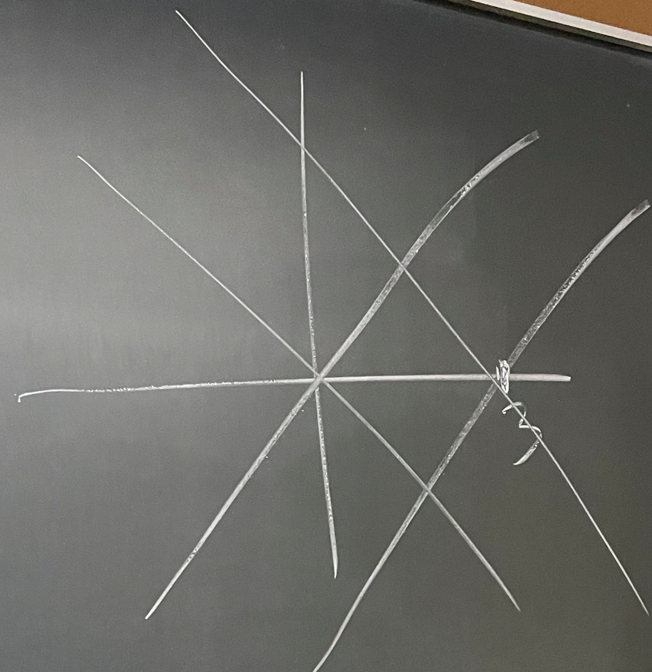
n!

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

cos



cost
case y=1
case y=0
0
$h_{\vec{w}}(\vec{x})$
prob of 1 (pred)
$-\log(-(z-1))$
$\log z$
z
$-\log z$
$-\log(-z)$
$-\log(-z)$



$y = x$

$y = x - 3$

$y = 3 - x$

SGD

for $i = 1, 2, \cdots n$: # shuffle

$\vec{w} \leftarrow \vec{w} - \alpha \nabla J_{\vec{x}_i}(\vec{w})$ → derivative/gradient

many steps! exercise!
(hint: chain rule)

$$\boxed{\vec{w} - \alpha \left(h_{\vec{w}}(\vec{x}_i) - y_i \right) \vec{x}_i}$$

Same as linear regression!

except

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

take

$J($

$J($

Sing

$\vec{J}($

Stochastic Gradient Descent for Logistic Regression (binary classification)

```
set w = 0 vector
while cost J(w) still changing:
    shuffle data points
    for i = 1…n:
        w <- w – alpha(derivative of J(w) wrt x_i)
    store J(w)
```

3 important pieces to SGD

- Hypothesis function (prediction)

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = p(y = 1|\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{w}\cdot\boldsymbol{x}}}$$

3 important pieces to SGD

- Hypothesis function (prediction)

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = p(y=1|\boldsymbol{x}) = \frac{1}{1+e^{-\boldsymbol{w}\cdot\boldsymbol{x}}}$$

- Cost function (want to minimize)

$$J(\boldsymbol{w}) = -\sum_{i=1}^{n} y_i \log h_{\boldsymbol{w}}(\boldsymbol{x_i}) + (1-y_i)\log(1-h_{\boldsymbol{w}}(\boldsymbol{x_i}))$$

3 important pieces to SGD

- Hypothesis function (prediction)

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = p(y = 1|\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{w}\cdot\boldsymbol{x}}}$$

- Cost function (want to minimize)

$$J(\boldsymbol{w}) = -\sum_{i=1}^{n} y_i \log h_{\boldsymbol{w}}(\boldsymbol{x_i}) + (1 - y_i) \log(1 - h_{\boldsymbol{w}}(\boldsymbol{x_i}))$$

- Gradient of cost wrt single data point x_i

$$\nabla J_{\boldsymbol{x}_i}(\boldsymbol{w}) = (h_{\boldsymbol{w}}(\boldsymbol{x_i}) - y_i)\boldsymbol{x_i}$$

Outline for October 26

- Continuous features
- Introduction to logistic regression
- Cost function and SGD for logistic regression
- Connection to cross entropy