

# CS 260: Foundations of Data Science

Prof. Sara Mathieson

Spring 2025



**HAVERFORD**  
COLLEGE

- **Lab 6** posted (Information Theory)
  - Due next Wednesday

Midterm 1 runtime

# Outline

- Lab 6: binary conversion, continuous features
- Introduction to logistic regression
- Cost function and SGD for logistic regression
- Connection to cross entropy

# Outline

- Lab 6: binary conversion, continuous features
- Introduction to logistic regression
- Cost function and SGD for logistic regression
- Connection to cross entropy

# Conversion to binary

- Handout 14, first question

Convert cumulative probs  $\rightarrow$  binary str

$CP = 0.2$       binary: 0.0011

---

$0.2$   
 $- 0.5 \rightarrow \text{no}$

$0.2$   
 $- 0.25 \rightarrow \text{no}$

$0.2$   
 $- 0.125 \rightarrow \text{yes!}$

$0.075$   
 $- 0.0625 \leftarrow 2^{-4} \rightarrow \text{yes!}$

finite #  
digits  
for the  
code

# Continuous Features

(do this for the TRAIN only!)

1) Sort examples based on given feature

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

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

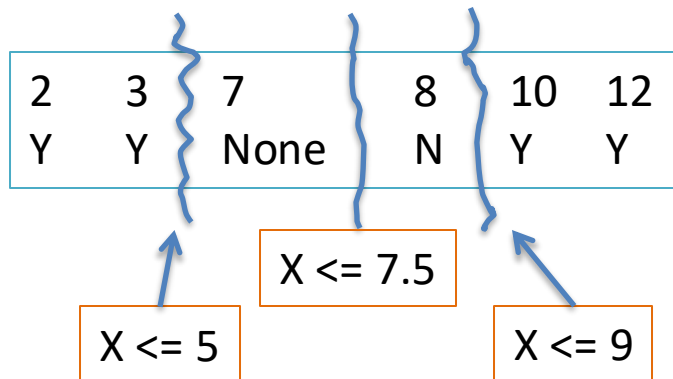
- 1) Sort examples based on given feature

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)



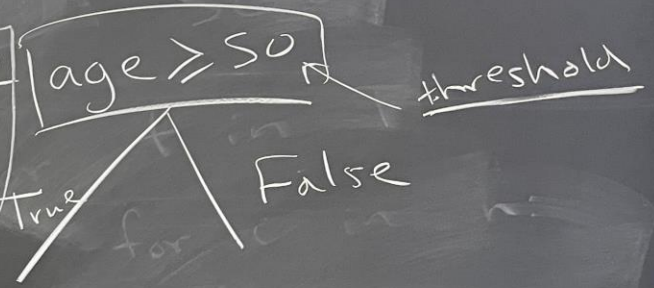


$x \geq 5$   $x \geq 7.5$   $x \geq 9$

feature x	label y
10	Y
7	Y
8	N
3	Y
7	N
12	Y
2	Y

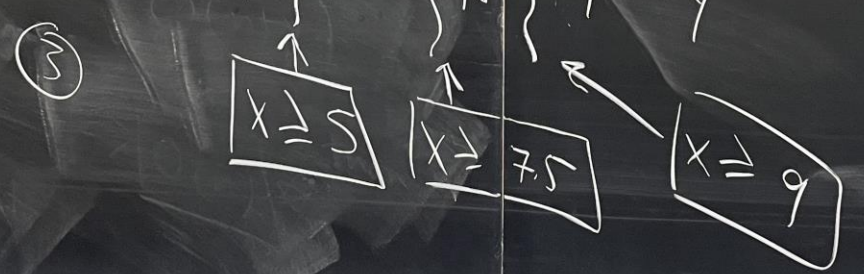
① Sort by x

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



② collapse feature values with more than one label

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



③ change in label, create a threshold / feature

# 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

- Lab 6: binary conversion, 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?

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  $[0, 1]$

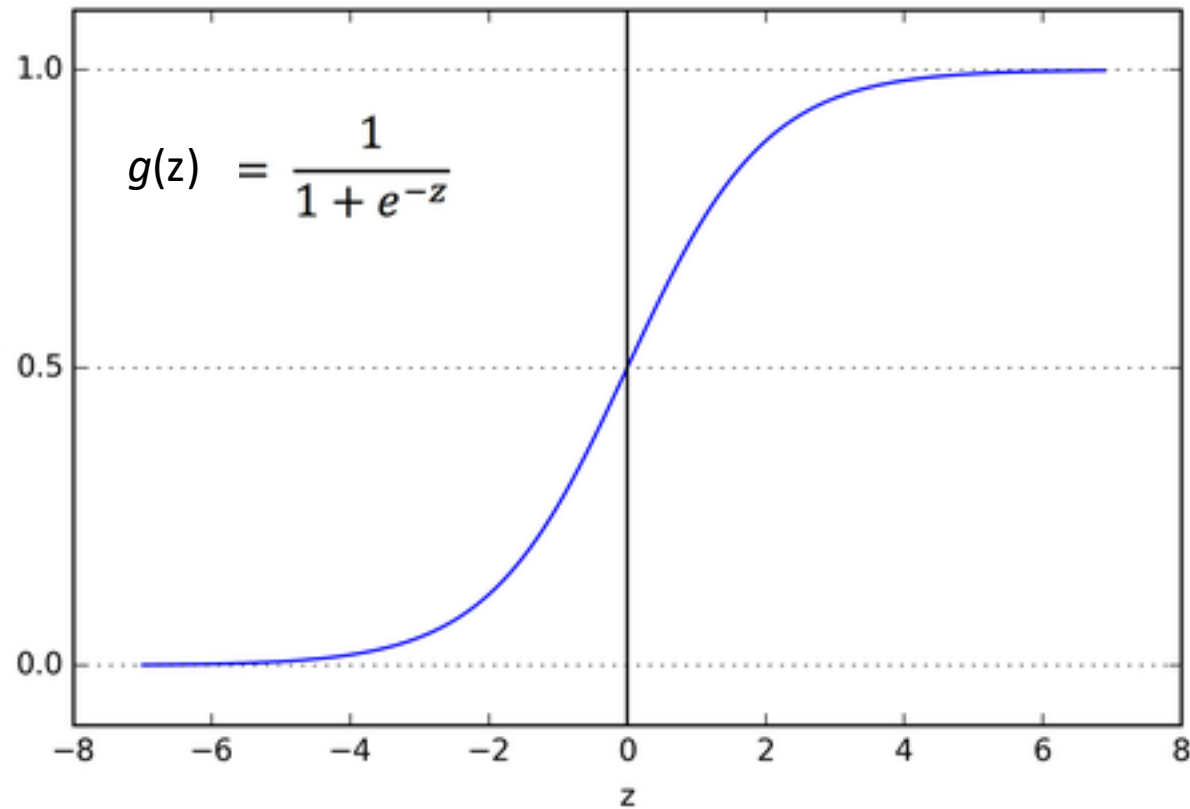
# Challenger Explosion Data



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

# Logistic (sigmoid) function





# Logistic Regression

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

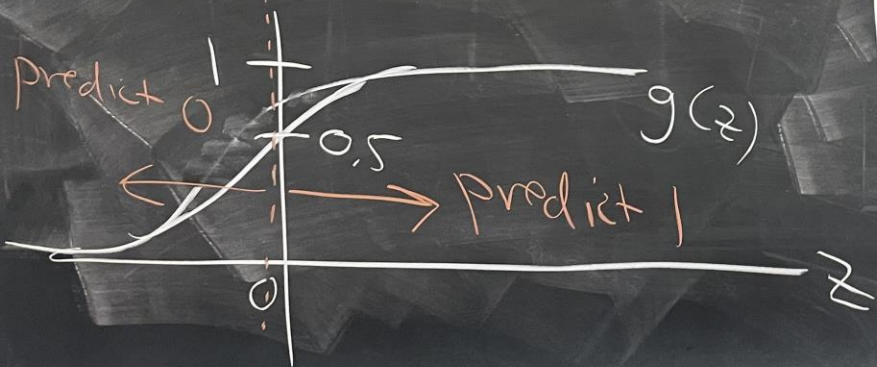
linear model:  $\overset{x}{[-\infty, \infty]} \rightarrow \overset{y}{[-\infty, \infty]}$

logistic model:  $[-\infty, \infty] \rightarrow [0, 1]$

probability

idea

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



linear regression

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

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

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

linear

say we know  $\vec{w}$

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

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

$$\rightarrow 1$$

$$\rightarrow 0$$

$$1 = \frac{1}{2} \left. \vphantom{\begin{matrix} 1 \\ 1 \end{matrix}} \right\} \text{linear decision boundary}$$

$$\text{know } \vec{w}$$

$$\geq 0 \Rightarrow \hat{y} = 1$$

$$< 0 \Rightarrow \hat{y} = 0$$

$$\frac{1}{1+e^0} = \frac{1}{2}$$

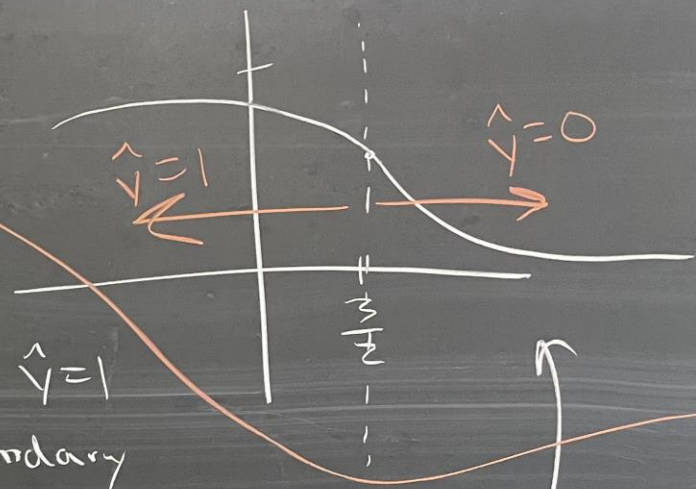
$$\text{ex } \vec{w} = \begin{bmatrix} 3 \\ -2 \end{bmatrix}$$

$$3 - 2x \geq 0$$

$$-2x \geq -3$$

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

decision boundary



$$h_{\vec{w}}(\vec{x}) = \frac{1}{1 + e^{-(3-2x)}} \geq \frac{1}{2}$$

true  
for  $p=1$

$$w_0 + w_1 x \geq 0$$

$$w_1 x \geq -w_0$$

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



# Outline

- Lab 6: binary conversion, continuous features
- Introduction to logistic regression
- **Cost function and SGD for logistic regression**
- Connection to cross entropy

$$\begin{aligned} &\rightarrow 2 \geq 1 + e^{-(3-2x)} \\ &\log(1) \geq \log(e^{-(3-2x)}) \\ &0 \geq -(3-2x) \\ &\boxed{0 \leq 3-2x} \end{aligned}$$

Goal: how to find  $\vec{w}$ ?

need a cost function!

start with likelihood

$$L(\vec{w}) = \prod_{i=1}^n$$

Want high only one term for each  $(x_i, y_i)$

$$y_i^{1 \text{ if } y_i=1} (1 - h_{\vec{w}}(\vec{x}_i))^{1 \text{ if } y_i=0}$$

prob of class 1      prob of class 0

$y_i$   
 1 if  $y_i = 1$   
 $1 - y_i$  if  $y_i = 0$

$(1 - h_{\vec{w}}(\vec{x}_i))$   
 prob of class 0

negative log likelihood =  $-\log$

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

need to change

$$\log N_k - \log n$$

$$\log \frac{N_k}{n}$$

# Stochastic Gradient Descent for Logistic Regression (binary classification)

```
set  $w = 0$  vector
```

```
while cost  $J(w)$  still changing:
```

```
    shuffle data points
```

```
    for  $i = 1 \dots n$ :
```

```
         $w \leftarrow w - \alpha(\text{derivative of } J(w) \text{ wrt } x_i)$ 
```

```
    store  $J(w)$ 
```

# 3 important pieces to SGD

- Hypothesis function (prediction)

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

# 3 important pieces to SGD

- Hypothesis function (prediction)

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

- Cost function (want to minimize)

$$J(\mathbf{w}) = - \sum_{i=1}^n y_i \log h_{\mathbf{w}}(\mathbf{x}_i) + (1 - y_i) \log(1 - h_{\mathbf{w}}(\mathbf{x}_i))$$

# 3 important pieces to SGD

- Hypothesis function (prediction)

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

- Cost function (want to minimize)

$$J(\mathbf{w}) = - \sum_{i=1}^n y_i \log h_{\mathbf{w}}(\mathbf{x}_i) + (1 - y_i) \log(1 - h_{\mathbf{w}}(\mathbf{x}_i))$$

- Gradient of cost wrt single data point  $\mathbf{x}_i$

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