

CS 260: Foundations of Data Science

Prof. Thao Nguyen

Fall 2024



HAVERFORD
COLLEGE

Outline for today

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

Outline for today

- 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!)

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

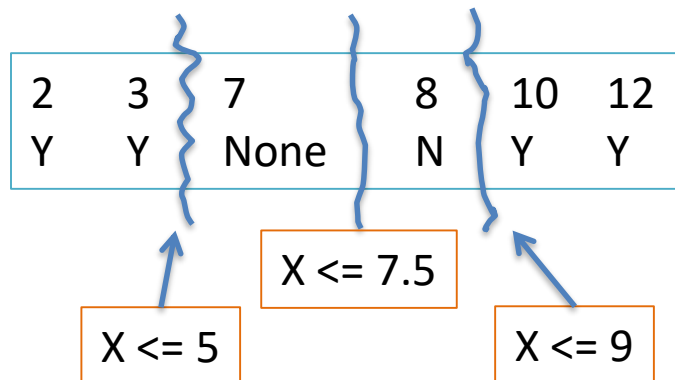
- 1) Sort examples based on given feature

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)



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)

Continuous Features (Handout 14)

3 new cols
 $x \leq 44$

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

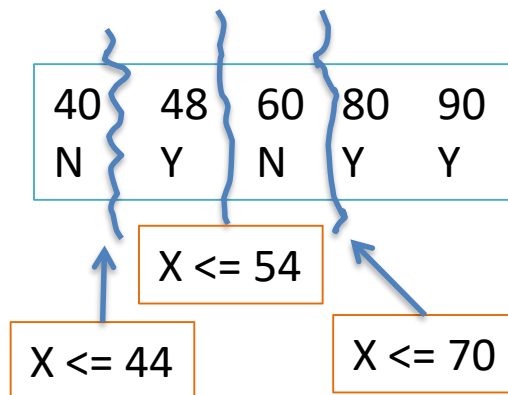
- 1) Sort examples based on feature "temp"

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

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

40	48	60	80	90
N	Y	N	Y	Y

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



Outline for today

- 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

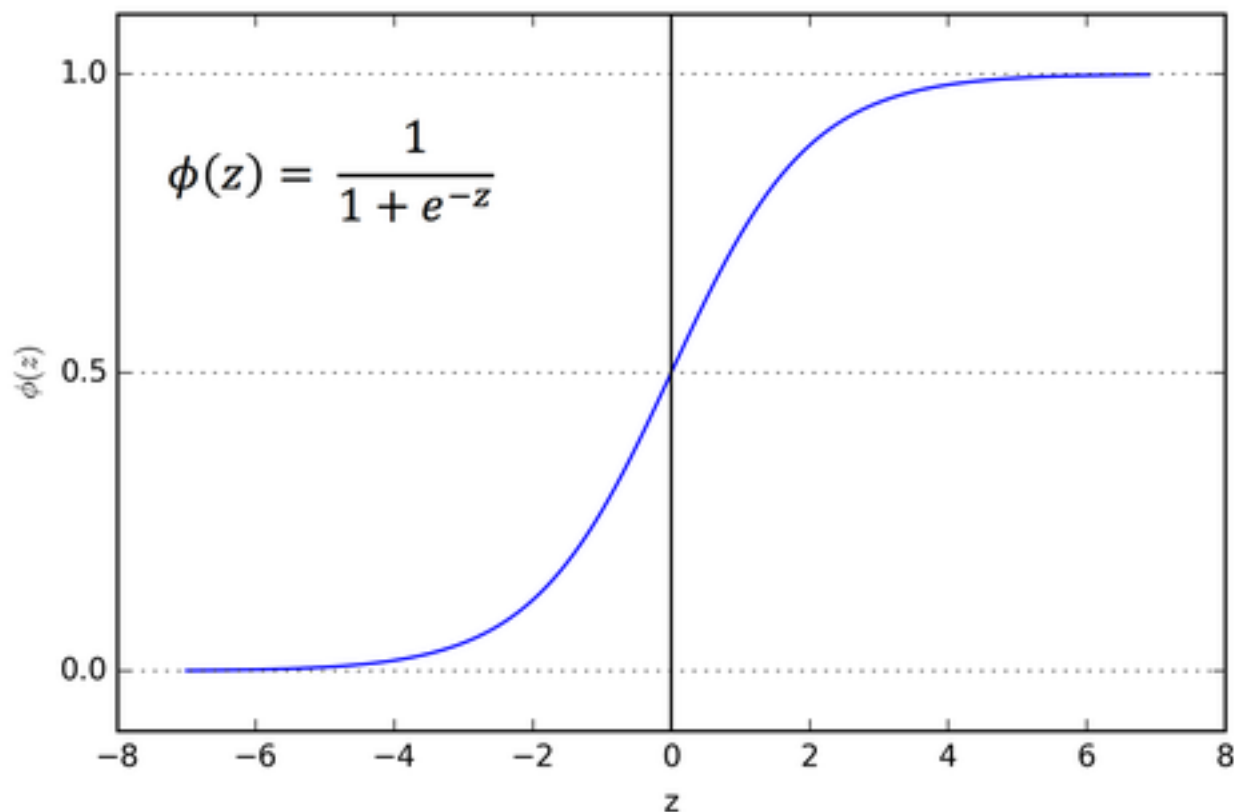


Image: NASA

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

Logistic (sigmoid) function

Transforms a continuous real number into a range of (0, 1)



Logistic Regression

- Binary classification $y \in \{0,1\}$
- Model will be

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

- Classification (already have \vec{w})

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

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

Logistic regression example

- If $p=1$ (one feature), can solve for x

$$w_0 + w_1 x \geq 0$$

$$w_1 x \geq -w_0$$

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

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

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

Outline for today

- 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
- Can measure model performance with likelihood

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

want high

Cost function for logistic regression

$$J(\vec{w}) = \underbrace{-\log(L(\vec{w}))}_{\text{negative log-likelihood}}$$

↑
minimize

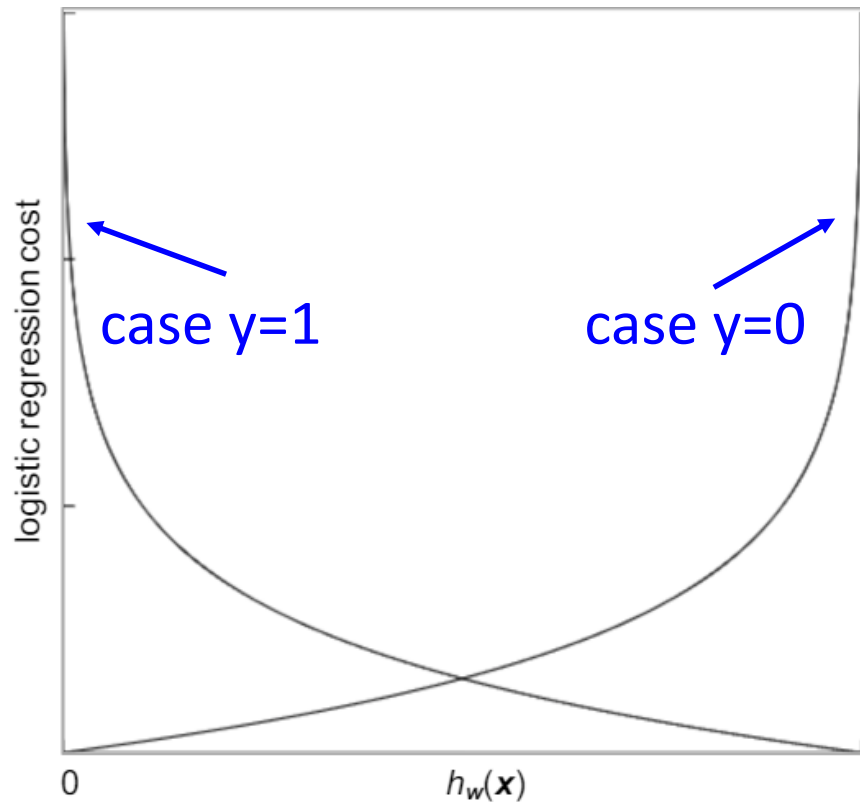
$$J(\vec{w}) = - \sum_{i=1}^n [y_i \log(h_{\vec{w}}(\vec{x}_i)) + (1 - y_i) \log(1 - h_{\vec{w}}(\vec{x}_i))]$$

- Single example \vec{x}, y

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

Single data point

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



Stochastic Gradient Descent for Logistic Regression (binary classification)

set $\vec{w} = \vec{0}$

while cost $J(\vec{w})$ is still changing:

 shuffle data points

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

$$\vec{w} \leftarrow \vec{w} - \alpha \underbrace{\nabla_{\vec{x}_i} J(\vec{w})}_{\text{derivative of } J(\vec{w}) \text{ wrt } x_i}$$

 store $J(\vec{w})$

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

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

Outline for today

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

Cost function as Cross Entropy

$$J(\vec{w}) = -\underbrace{y}_{\text{probability distribution}} \log(h_{\vec{w}}(x)) + \underbrace{(1-y)}_{\text{probability distribution}} \log(1 - h_{\vec{w}}(x))$$

$$\text{entropy } H(Y) = - \sum_{y \in \text{vals}(Y)} p(y) \log p(y)$$

$$\text{Cross entropy } H(p, q) = - \sum_x p(x) \log q(x)$$

Cost function as Cross Entropy

- Example
 - true: $y=0$, $1-y=1$
 - pred: $h=0.4$, $1-h=0.6$

$$H(\text{true}, \text{pred}) = -(0 \log(0.4) + 1 \log(0.6)) = 0.5$$