CS 260: Foundations of Data Science

Prof. Thao Nguyen

Fall 2024

HAVERFORD COLLEGE

Materials by Sara Mathieson

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!)

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 <= 7.5

X <= 5

X <= 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)

Continuous Features (Handout 14)

3 new cols

x≤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)

40	48	60	80	90
N	Y	N	Y	Y

X <= 44

X <= 54

X <= 70

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 $[0, 1]$

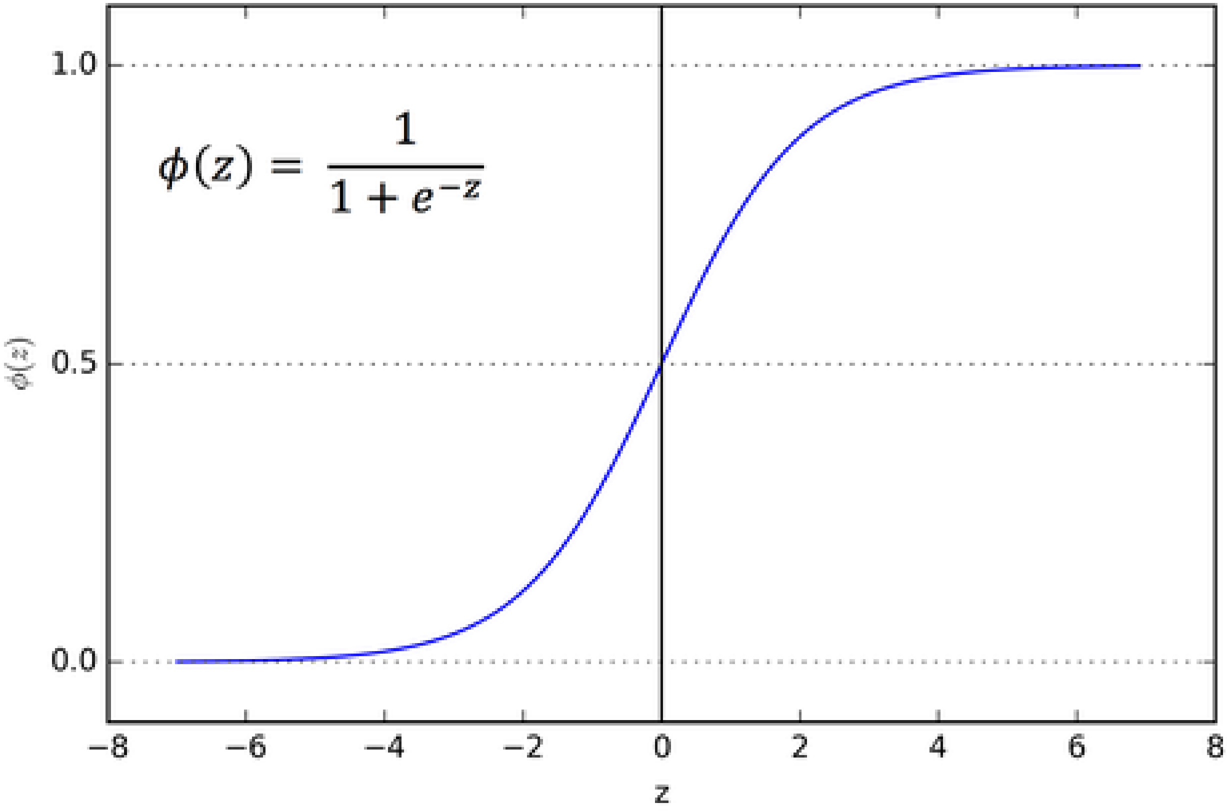
Challenger Explosion Data

Image: NASA

	Date	Temperature	Damage Incident
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 $\vec{w}$)

  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

$$w_0 + w_1 x \geq 0$$
$$w_1 x \geq -w_0$$
$$x \geq -\frac{w_0}{w_1}$$

- <u>Ex:</u> $\vec{w} = \begin{bmatrix} 3 \\ -2 \end{bmatrix}$

  $x \leq \frac{3}{2}$ 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

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

Cost function for logistic regression

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

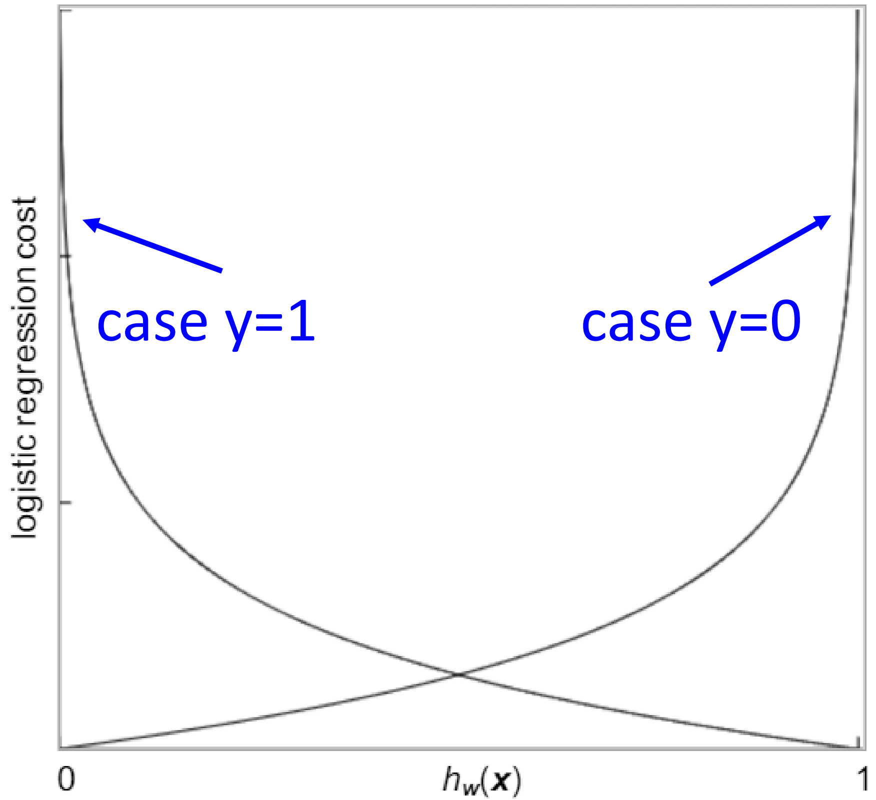
$$J(\vec{w}) = -\sum_{i=1}^{n}[y_i \log(h_{\vec{w}}(\vec{x_i})) + (1 - y_i)\log(1 - h_{\vec{w}}(\vec{x_i}))]$$

- Single example $\vec{x}, y$

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

Single data point

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



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,…,n:

    $$\vec{w} \leftarrow \vec{w} - \underbrace{\alpha \nabla J_{\vec{x_i}}(\vec{w})}_{\text{derivative of } J(\vec{w}) \text{ wrt } x_i}$$

  store $J(\vec{w})$

3 important pieces to SGD

- Hypothesis function (prediction)

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

3 important pieces to SGD

- Hypothesis function (prediction)

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

- Cost function (want to minimize)

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

3 important pieces to SGD

- Hypothesis function (prediction)

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

- Cost function (want to minimize)

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

- Gradient of cost wrt single data point x_i

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

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

probability distribution

$$J(\vec{w}) = -y\log(\overline{h_{\vec{w}}}(x)) + (1-y)\log(\overline{1 - h_{\vec{w}}(x)})$$

probability distribution

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

entropy

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

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