CS 260: Foundations of Data Science

Prof. Sara Mathieson

Fall 2021

HAVERFORD
COLLEGE

Admin

- I’m giving a talk “at” a conference during our class on Thursday Oct 21
  - There will still be lab though!
- Watch this lecture video instead and do **Handout 14**
- Come prepared to discuss **Handout 14** on Tuesday Oct 26

Outline for October 21

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

Outline for October 21

- 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 $[0, 1]$

Logistic Regression Intro

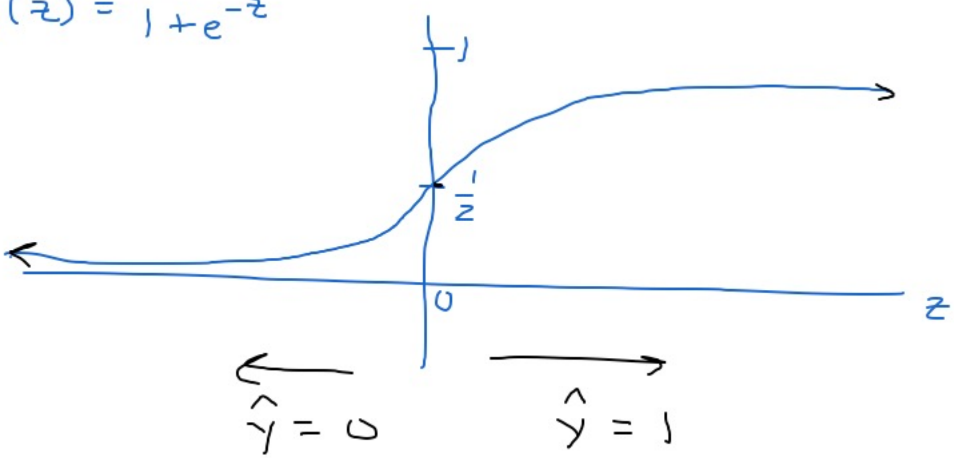
$y \in \{0, 1\}$ binary classification

linear regression: weight on each feature ☆

	x		y	
linear:	$[-\infty, \infty]$	$\longrightarrow$	$[-\infty, \infty]$	
logistic:	$[-\infty, \infty]$	$\longrightarrow$	$\underbrace{[0, 1]}_{\text{prob}}$	} can make discrete

<u>idea</u> model: $h_{\vec{w}}(\vec{x}) = \overbrace{p(y=1 \mid \vec{x})}^{\text{posterior}} = \dfrac{1}{1 + e^{-\vec{w}\cdot\vec{x}}}$ linear function

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

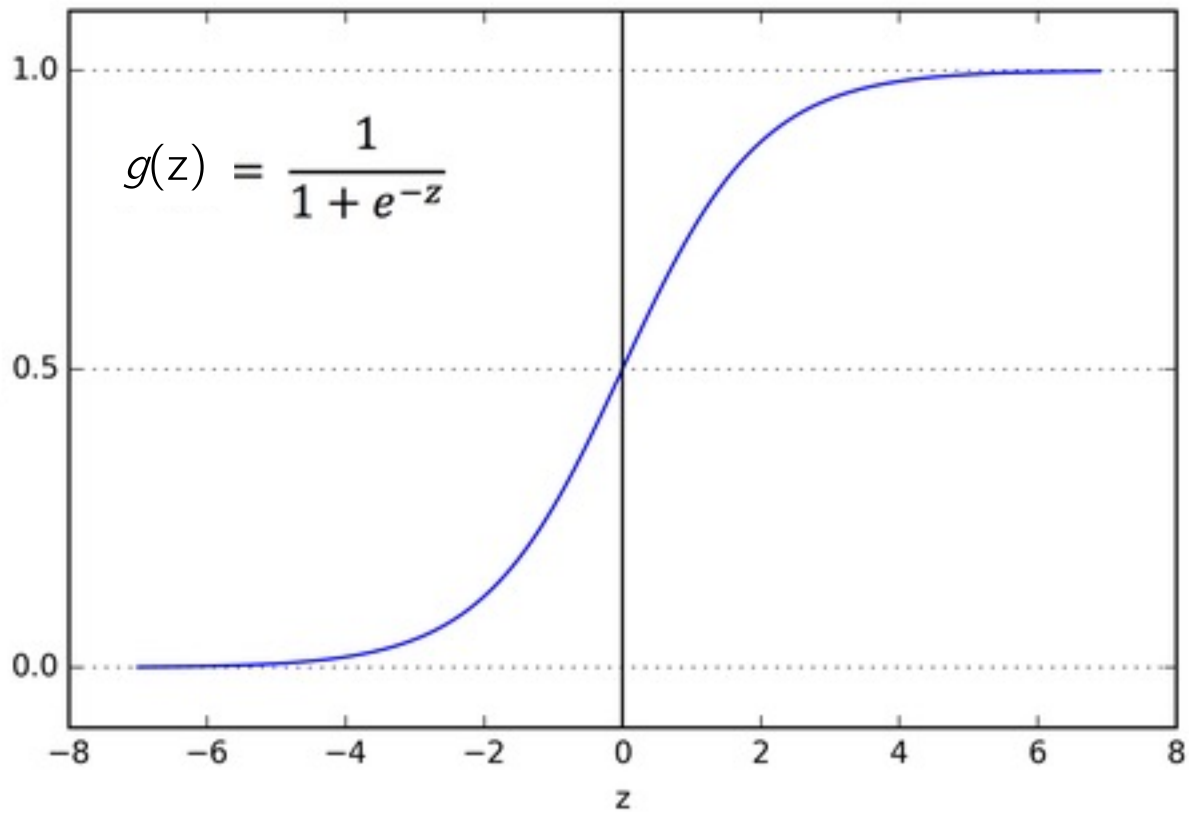


$z \to \infty$, $g(z) \to 1$

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

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

Logistic (sigmoid) function



Logistic Regression Decision Boundaries

if $\boxed{\begin{array}{ll} \vec{w}\cdot\vec{x} > 0 & \Rightarrow \hat{y}=1 \\ \vec{w}\cdot\vec{x} \leq 0 & \Rightarrow \hat{y}=0 \end{array}}$

$\vec{w} = \begin{bmatrix} 3 \\ -2 \end{bmatrix}$, $p=1$

$$h(\vec{x}) = \frac{1}{1+e^{-(3-2x)}} \quad \text{model}$$

$$= p(y=1 \mid \vec{x}) > \frac{1}{2}$$

Q: what is classified as $\hat{y}=1$?

$$\frac{1}{1+e^{-(3-2x)}} > \frac{1}{2}$$

$$2 > 1+e^{-3+2x}$$

$$1 > e^{-3+2x}$$

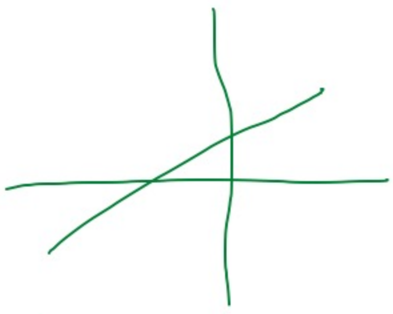
$$0 > -3+2x$$

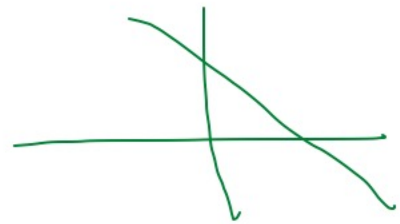
$$3 > 2x$$

$$\boxed{x < \frac{3}{2}} \Rightarrow \hat{y}=1$$

Logistic Regression Decision Boundaries

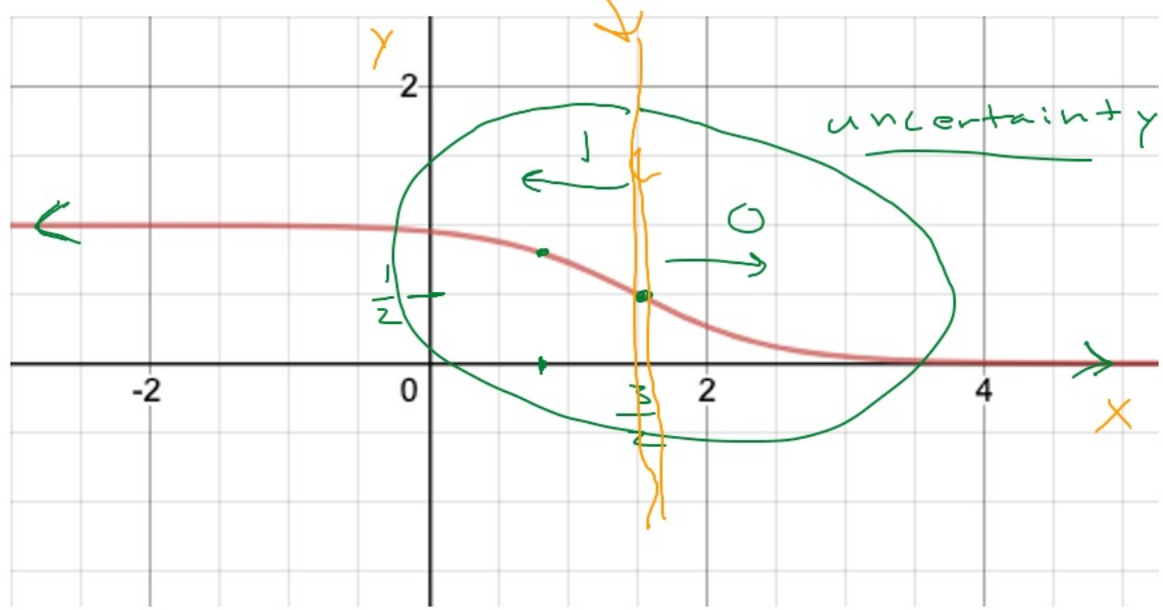
linear





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

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



Outline for October 21

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

Logistic Regression Cost Function

How do we find $\vec{w}$? need cost function! $\Rightarrow$ SGD

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

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

maximize

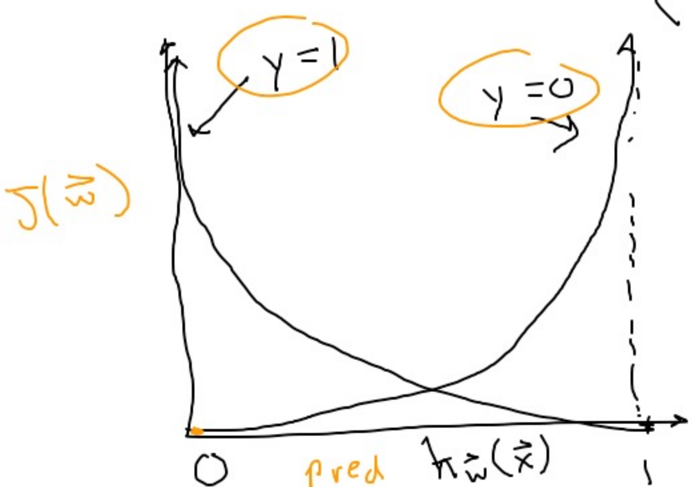
minimize negative log likelihood = cost

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

single example

$(\vec{x}, y)$

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



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)

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

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

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

3 important pieces to SGD

$\vec{w} \leftarrow \vec{w} - \alpha(\underbrace{h_{\vec{w}}(\vec{x}_i)}_{\text{pred}} - \underbrace{y_i}_{\text{truth}})\vec{x}_i$

- 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

same form as linear reg!

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

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

Outline for October 21

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

Cross entropy

$J(\vec{w}) = -\underbrace{y \log h(x) - (1-y)\log(1-h)}$

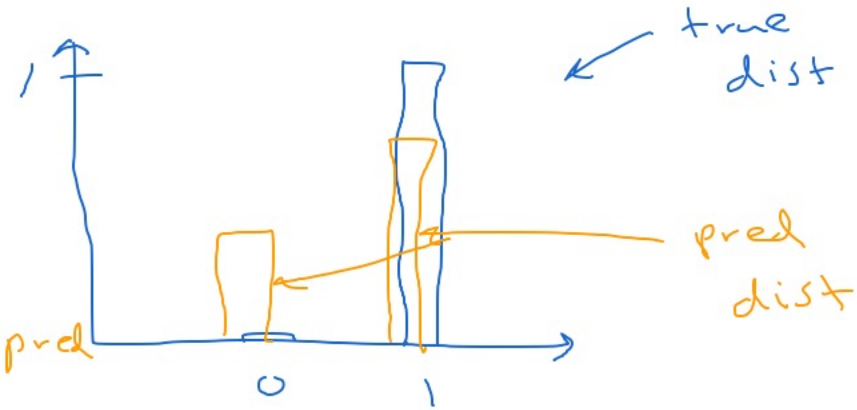
cost function is $\underbrace{\text{Cross entropy}}$

$$H(p, q) = -\sum_x \underline{p(x) \log q(x)}$$

2 probability distributions

if $y = 1$, $1 - y = 0$ } true

$\underbrace{h = 0.6}_{\text{high}}$, $\underbrace{1 - h = 0.4}_{\text{low}}$ } pred



if dist. are exactly the same

$\Rightarrow$ cost / cross-entropy $= 0$

TODO: Handout 14!