

CS 360: Machine Learning

Prof. Sara Mathieson

Fall 2020



HVERFORD
COLLEGE

- **Midterm 1 (take in a 2 hour block)**
 - May use self-created “cheat-sheet” + calculator
 - Can print out in advance, scan the next day if necessary

- Extra office hours **TODAY 1-2pm**

- **Lab 2** grades will be up today

Considerations when using a decision tree in medical applications

Advantages

- Very fast, gives a quick guide
- Could be more objective across different patient circumstances

Potential issues:

- Requires lots of historical data
- Decision trees work with how symptoms and other factors were historically measured
- If a new type of scan, xray, etc becomes available, we won't have historical data to incorporate into the algorithm
- Decision trees do not elegantly handle continuous features
- Patient may have a condition not seen in the training data
- 75% is not an optimal accuracy!

Outline for October 9

- Review
 - Loss functions & bias-variance tradeoff
 - Linear Regression
 - SGD with varying alpha
 - Closed form vs. SGD
- Begin: Logistic Regression

Outline for October 9

- Review
 - Loss functions & bias-variance tradeoff
 - Linear Regression
 - SGD with varying alpha
 - Closed form vs. SGD
- Begin: Logistic Regression

Loss Functions

$$y = \text{true}$$

$$\hat{y} = \text{pred}$$

- Zero-one loss (binary and multi-class classification)

want low

$$l(y, \hat{y}) = \begin{cases} 0 & \text{if } y = \hat{y} \\ 1 & \text{o.w.} \end{cases}$$

- Squared loss (regression)

$$l(y, \hat{y}) = (\hat{y} - y)^2$$

linear reg

$$J(\vec{w}) = \frac{1}{2} \sum_{i=1}^n (\vec{w} \cdot \vec{x}_i - y_i)^2$$

Bias-Variance Tradeoff

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2$$

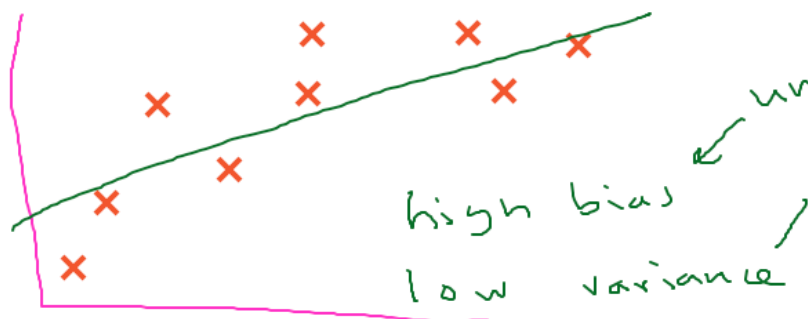
assumption
 $y = f(x) + \varepsilon$
 model noise

$$E[MSE] = E[\underbrace{(\hat{f} - f)^2}_{\text{reducible error}}] + \underbrace{\text{Var}(\varepsilon)}_{\text{irreducible error}}$$

$$\text{Var}(x) = E(x - \mu)^2$$

$$= \underbrace{\text{bias}(\hat{f})^2}_{\text{different training datasets}} + \underbrace{\text{Var}(\hat{f})}_{\text{different training datasets}} + \text{Var}(\varepsilon)$$

$$(E[\hat{f}] - f)^2$$



Outline for October 9

- Review
 - Loss functions & bias-variance tradeoff
 - Linear Regression
 - SGD with varying alpha
 - Closed form vs. SGD
- Begin: Logistic Regression

SGD with varying alpha

* common choice : $\frac{1}{t}$ where t is the iteration

$t = 1$

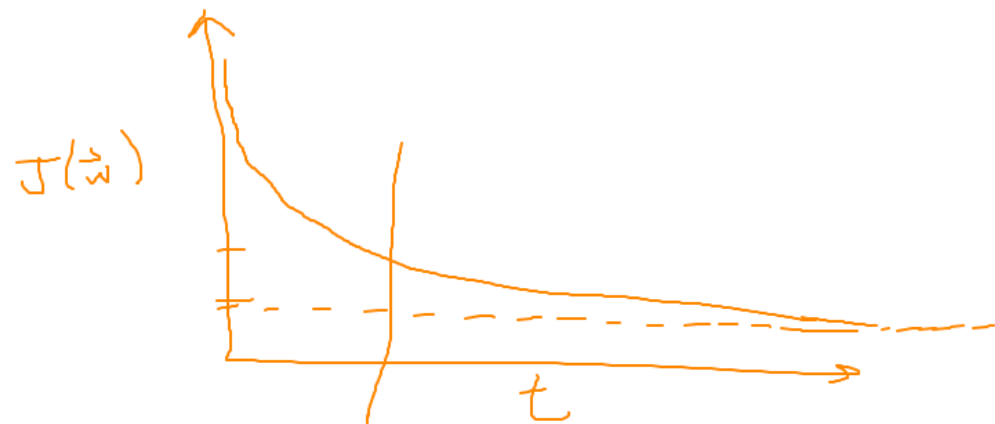
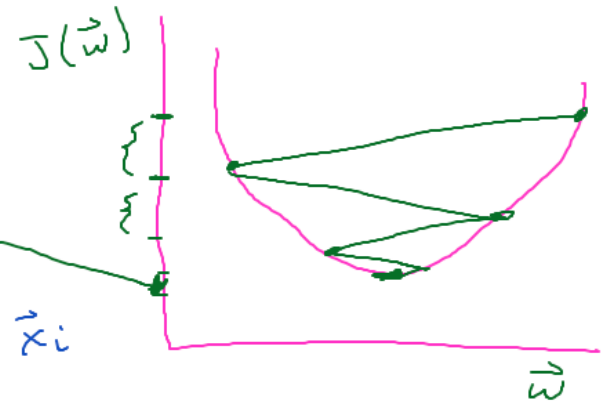
while not converged

$$\alpha = \frac{1}{t}$$

for $i = 1 \dots n$

$$\begin{bmatrix} \vec{w} \end{bmatrix} \leftarrow \begin{bmatrix} \vec{w} \end{bmatrix} - \alpha (\hat{y}_i - y_i) \begin{bmatrix} \vec{x}_i \end{bmatrix}$$

$t += 1$



Pros and Cons

Gradient Descent

- requires multiple iterations
- need to choose α
- works well when p is large
- can support online learning

Normal Equations

- non-iterative
- no need for α
- slow if p is large
 - matrix inversion is $O(p^3)$

Outline for October 9

- Review
 - Loss functions & bias-variance tradeoff
 - Linear Regression
 - SGD with varying alpha
 - Closed form vs. SGD
- **Begin: Logistic Regression**

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?

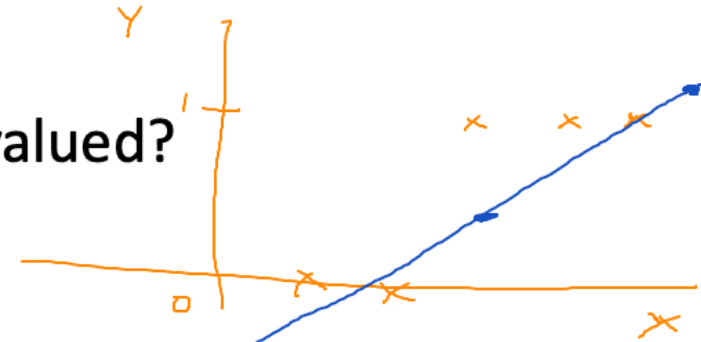
Case Study: you need to identify the medical condition of a patient in the emergency room on the basis of their symptoms.

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?

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

* lin reg gave us notion of feature importance
 $h(\vec{x}) = 0.1x_1 - 7x_2 + 3$

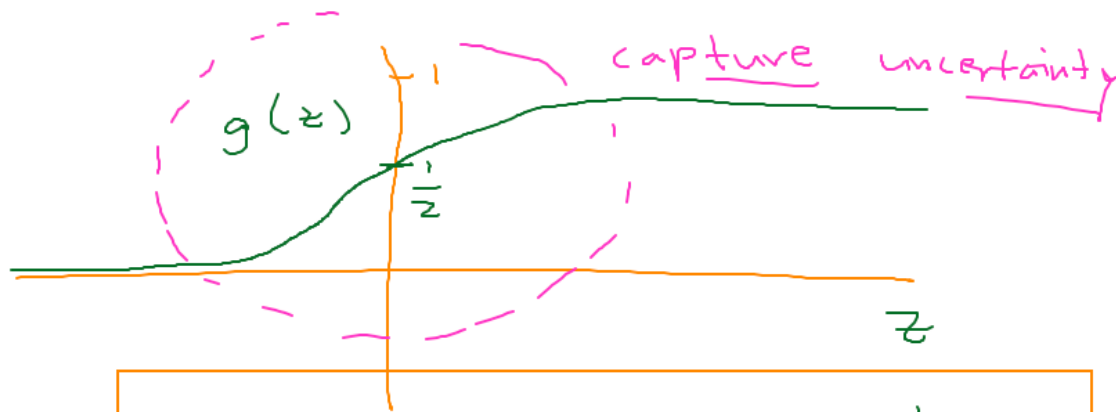
$\Rightarrow [-\infty, \infty] \rightarrow [-\infty, \infty]$

want $[-\infty, \infty] \rightarrow [0, 1]$

model: logistic function

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

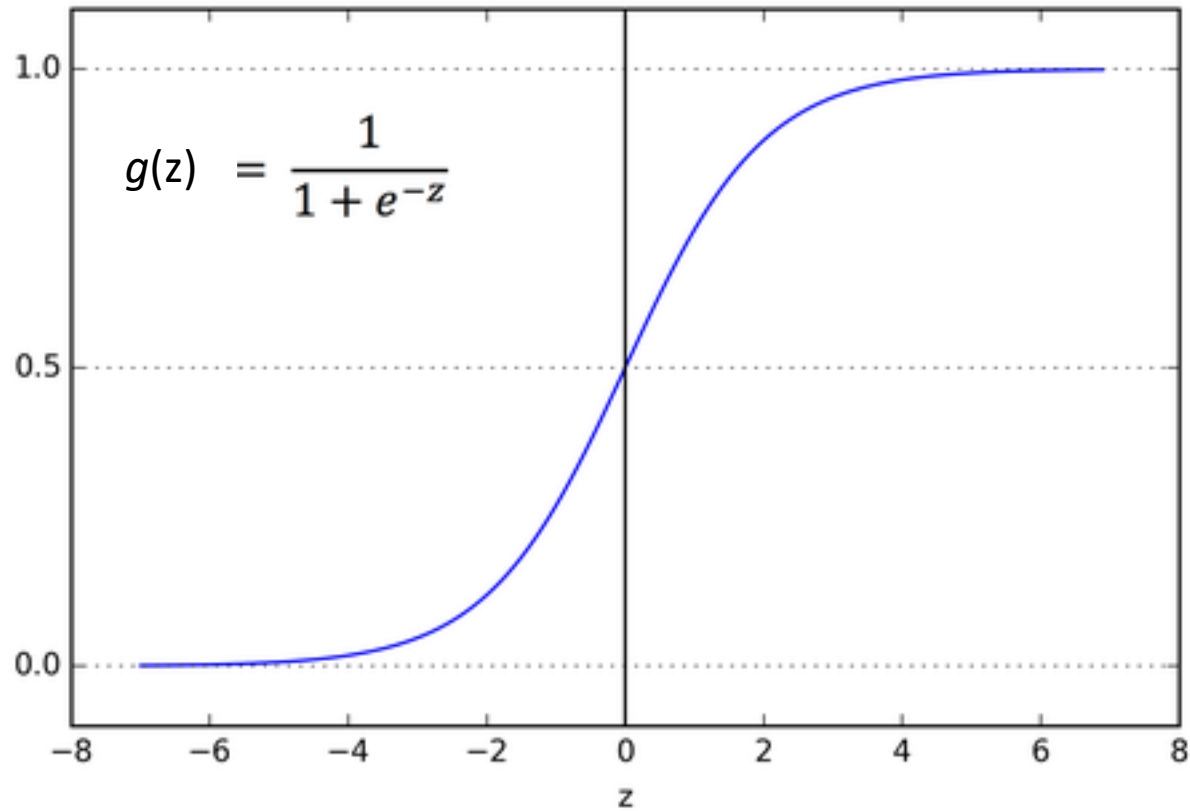
$z \rightarrow \infty, g(z) \rightarrow 1$
 $z \rightarrow -\infty, g(z) \rightarrow 0$



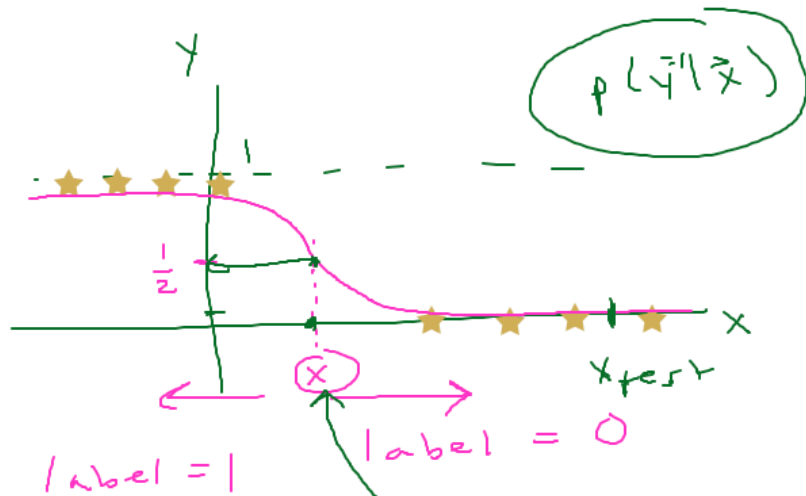
$$h_{\vec{w}}(\vec{x}) = \underbrace{p(y=1|\vec{x})}_{\text{N.B.}} = \frac{1}{1 + e^{-\vec{w} \cdot \vec{x}}}$$

	y	
	cont	disc
cont	lin reg	log reg
disc	?	decision tree
	lin reg	N.B.

Logistic (sigmoid) function



Logistic Regression Intro



$$p(\hat{y}=1|\vec{x}) \rightarrow \frac{g(\vec{w} \cdot \vec{x})}{1 + e^{-\vec{w} \cdot \vec{x}}} = \frac{1}{2}$$

$$\frac{1}{1 + e^{-\vec{w} \cdot \vec{x}}} = \frac{1}{2}$$

$$-\vec{w} \cdot \vec{x} = 0 \quad \text{prediction}$$

$$\vec{w} \cdot \vec{x} > 0 \Rightarrow \hat{y} = 1$$

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

1D
1 feature

$$w_0 + w_1 x = 0$$

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

1-2 pm 01+

Extra: hypotheses for several models

lin reg

$$h_{\vec{w}}(\vec{x}) = \vec{w} \cdot \vec{x}$$

log reg

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

$> \frac{1}{2} \Rightarrow \hat{y} = 1$
 $\leq \frac{1}{2} \Rightarrow \hat{y} = 0$

n B

$$h_{y=k}(\vec{x}) = \frac{p(x_k) \prod p(x_j | x_{j \neq k})}{p(x)}$$

$\arg \max_k$ over k
 $\Rightarrow \hat{y}$

Handout 8

Bernoulli Random Variable. Say we flip a weighted coin n times, and each time the probability of heads (1) is p , so the probability of tails (0) is $(1 - p)$. Let y_i be the outcome of flip i . For example, if $n = 10$, we might observe these values:

$$\mathbf{y} = [0, 0, 1, 1, 0, 1, 0, 1, 0, 0]$$

In this case, the *likelihood* of p given this observed data is $L(p) = p^4(1 - p)^6$, since we observe four 1's and six 0's. In general, we can write the likelihood as

Next time!