CS 260 – Foundations of Data Science

Lecture 11 – Logistic Regression

10/19/2026

Adam Poliak

Announcements

HW5 due Tuesday night

Midterm 2 – November 19th

Midterm 1 – go over in lab today

Outline

Introduction to Logistic Regression

Cost function

Challenger Explosion Data

Image: NASA

	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

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

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?

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)

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?

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!

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 (e.g. 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 (e.g. stroke and drug overdose) -- why is linear regression still not a good choice?

The range of a linear function (i.e. *y* values) is [-∞, ∞], but we want [0, 1]

Desirata

Very positive

Zero

Very positive

Desirata

Very positive

Close to 0

Zero

0.5

Very positive

Close to 1

Desirata

Very positive

Close to 0

Zero

0.5

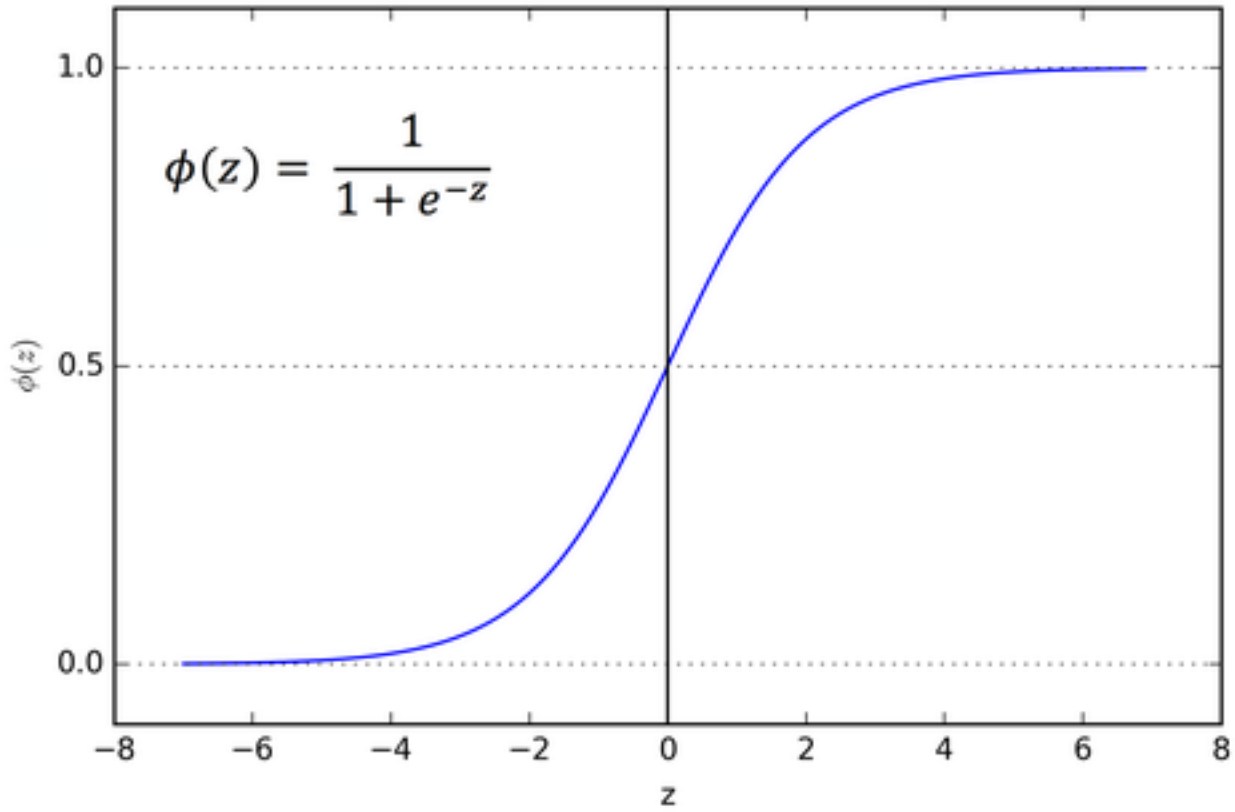
Very positive

Close to 1

Smooth transitions

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

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

$$= p(y = 1|\vec{x})$$

- 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 directly

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

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

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

Outline

Introduction to Logistic Regression

Cost function & SGD

How to find $\vec{w}$?

- Need a cost function
- Can measure model performance with likelihood

$$L(\vec{w}) = \prod_{i=1}^{n} h_{\vec{w}}(\vec{x_i})^{y_i} \left(1 - h_{\vec{w}}(\vec{x_i})\right)^{(1-y_i)}$$

want high

prob of y=1

prob of y=0

Cost function for logistic regression

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

$\uparrow$ minimize

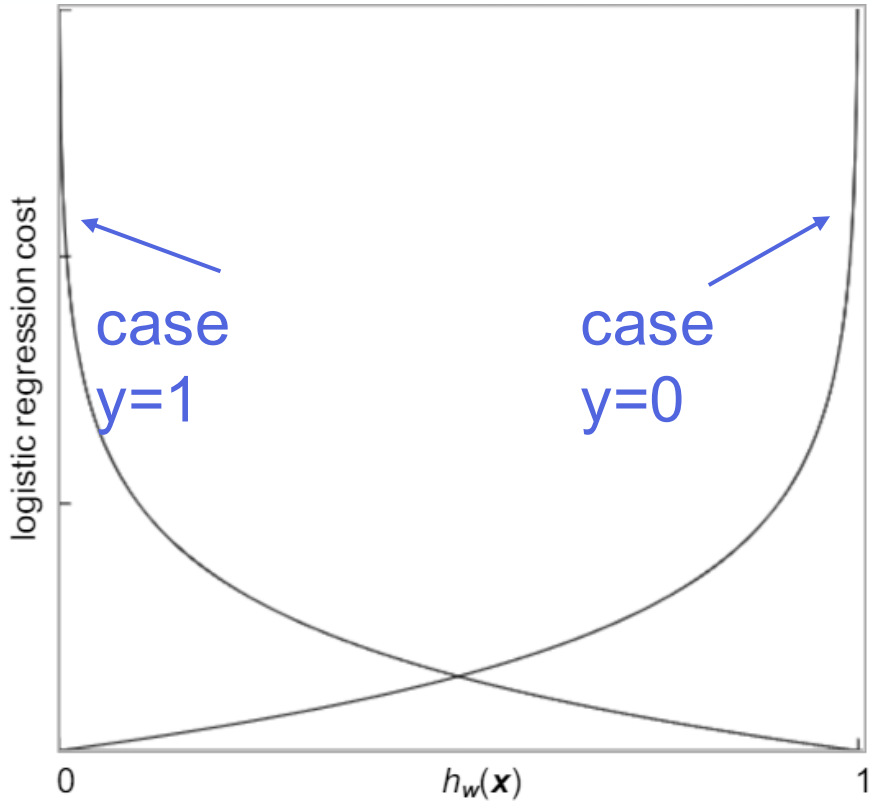
$$J(\vec{w}) = -\sum_{i=1}^{n}[y_i \log(h_{\vec{w}}(\vec{x_i})) + (1 - y_i)\log(1 - h_{\vec{w}}(\vec{x_i}))]$$

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

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

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:

$$\vec{w} \leftarrow \vec{w} - \alpha \nabla J_{\vec{x_i}}(\vec{w})$$

store $J(\vec{w})$

derivative of $J(\vec{w})$ wrt x_i

3 important pieces to SGD

- Hypothesis function (prediction)

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

3 important pieces to SGD

- Hypothesis function (prediction)

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

- Cost function (want to minimize)

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

3 important pieces to SGD

- Hypothesis function (prediction)

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

- Cost function (want to minimize)

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

- Gradient of cost wrt single data point x_i

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

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:

$$\vec{w} \leftarrow \vec{w} - \alpha \nabla J_{\vec{x_i}}(\vec{w})$$

store $J(\vec{w})$

derivative of $J(\vec{w})$ wrt x_i

Cost function

probability distribution

probability distribution

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

entropy

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

Cross entropy

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

Cost function as Cross Entropy

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

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

Handout

Handout

1. The output of logistic regression is a model that creates:

   (a) a linear decision boundary

   (b) a logistic decision boundary

   (c) no decision boundary

Handout

1. The output of logistic regression is a model that creates:

   (a) a linear decision boundary

   (b) a logistic decision boundary

   (c) no decision boundary

2. We use logistic regression for:

   (a) classification

   (b) regression

   (c) both

Handout

1. The output of logistic regression is a model that creates:

   (a) a linear decision boundary

   (b) a logistic decision boundary

   (c) no decision boundary

2. We use logistic regression for:

   (a) classification

   (b) regression

   (c) both

3. Our hypothesis in logistic regression is:

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

   If $\boldsymbol{w}$ is the zero vector (as it would be when starting SGD), what is the probability $y = 1$?

Handout

1. The output of logistic regression is a model that creates:

   (a) a linear decision boundary

   (b) a logistic decision boundary

   (c) no decision boundary

2. We use logistic regression for:

   (a) classification

   (b) regression

   (c) both

3. Our hypothesis in logistic regression is:

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

   If $\boldsymbol{w}$ is the zero vector (as it would be when starting SGD), what is the probability $y = 1$?

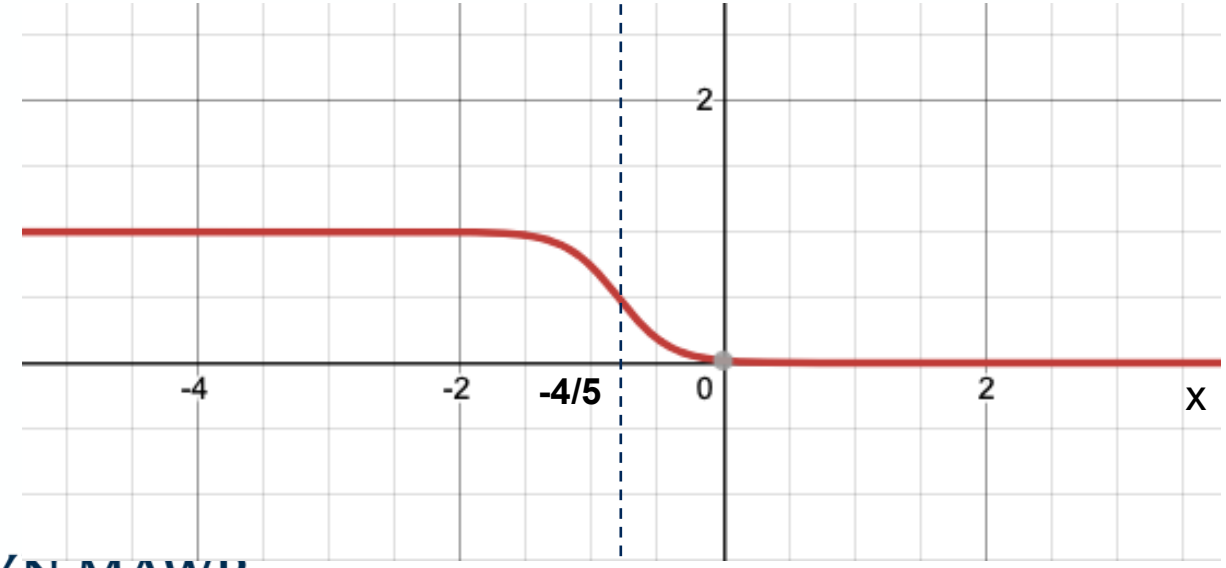
½

Handout

Say I train a binary logistic regression model (i.e. outcomes $\in \{0, 1\}$) and end up with $\hat{\boldsymbol{w}} = [\hat{w}_0, \hat{w}_1]^T = [-4, -5]^T$. What is the decision boundary? Sketch a graph of this logistic model and label the decision boundary. How would you classify a new point $x_{\text{test}} = -2$?

predict y=1

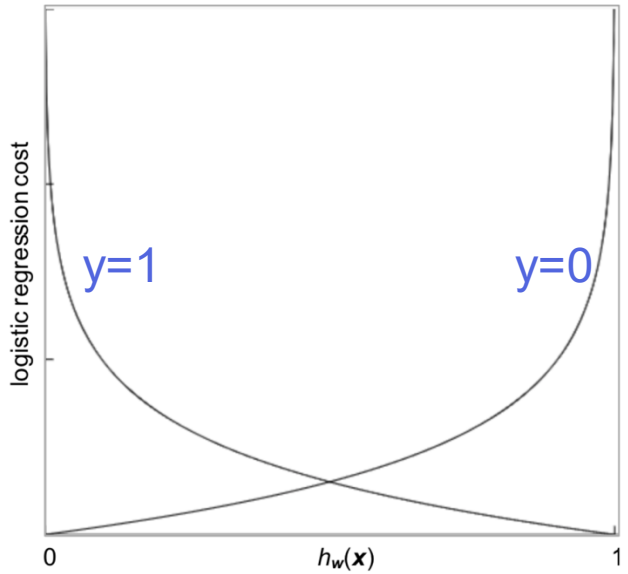
$$\frac{1}{\left(1 + e^{-(-4-5x)}\right)}$$



Decision boundary

Handout

5. The graph below shows the cost for logistic regression as a function of the hypothesis $h_{\boldsymbol{w}}(\boldsymbol{x})$, for one example $\boldsymbol{x}$. Which curve corresponds to the true label $y = 0$ and which corresponds to $y = 1$?



Single example $\vec{x}, y$

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