UCLA Samueli Computer Science

CS M146 Discussion: Week 4

Logistic Regression & Linear Regression

Junheng Hao

Friday, 01/29/2021

UCLA
Engineer Change.

Roadmap

- Announcement
- Logistic Regression
- Linear Regression

UCLA
Engineer Change.

Announcements

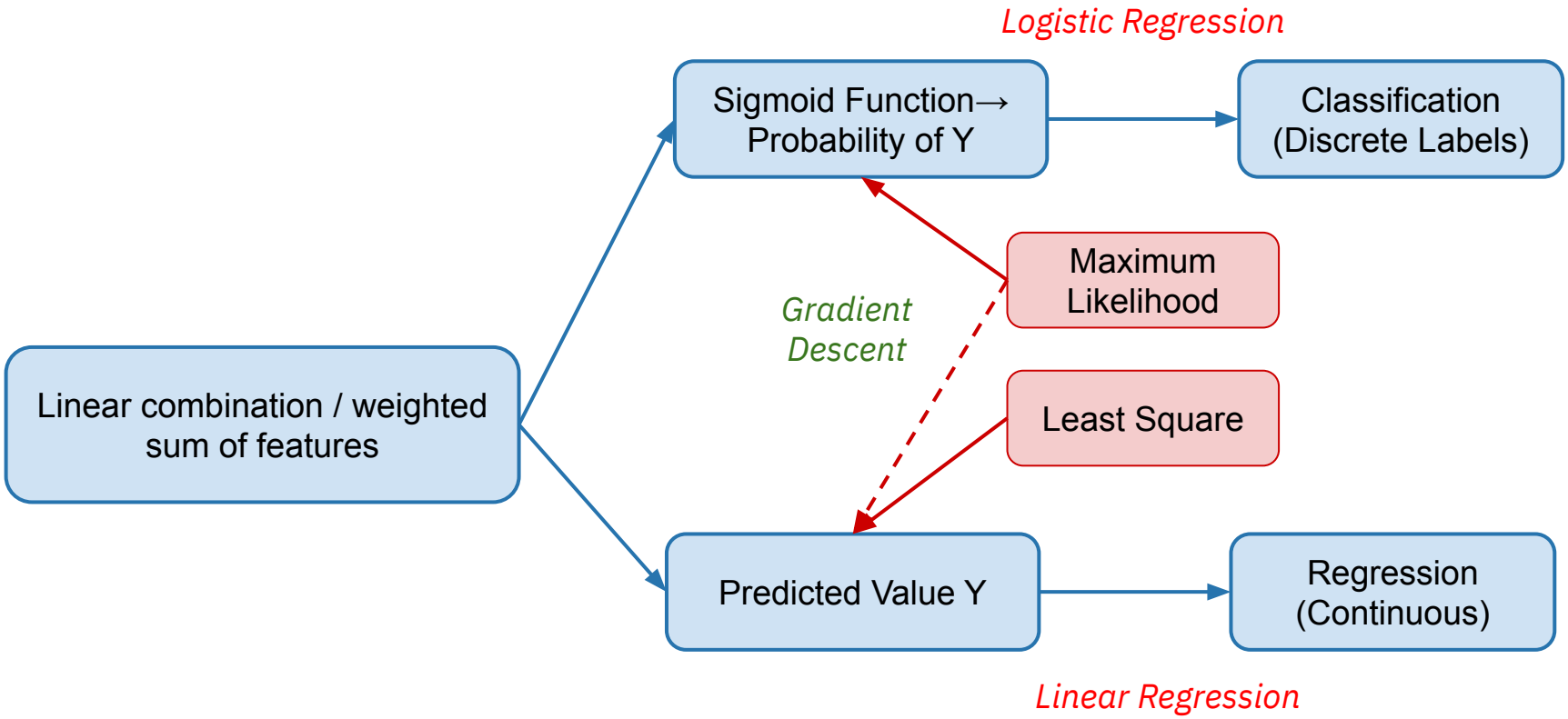
- **5:00 pm PST, Jan. 29:** Weekly Quiz 4 released on Gradescope.
- **11:59 pm PST, Jan. 31 (Sunday):** Weekly quiz 4 closed on Gradescope!
  - Start the quiz before **11:00 pm PST, Jan. 31** to have the full 60-minute time
- Problem set 1 released on CCLE, submission on Gradescope.
  - Please assign pages of your submission with corresponding problem set outline items on GradeScope.
  - You do not need to submit code, only the results required by the problem set
  - Due on **TODAY 11:59pm PST, Jan. 29 (Friday)**
- Problem set 2 expected to be released on CCLE, submission on Gradescope.
  - Due on two week later, **11:59pm PST, Feb. 12 (Friday)**

UCLA
Engineer Change.

About Quiz 4

- Quiz release date and time: **Jan 29, 2021 (Friday) 05:00 PM PST**
- Quiz due/close date and time: **Jan 31, 2021 (Sunday) 11:59 PM PST**
- You will have up to **60 minutes** to take this exam. → Start before **11:00 PM** Sunday
- You can find the exam entry named "Week 4 Quiz" on GradeScope.
- Topics: **Logistic Regression, Linear Regression, Gradient Descent**
- Question Types
  - True/false, multiple choices
- Some light calculations are expected. Some scratch paper and one scientific calculator (physical or online) are recommended for preparation.

Today's topic



UCLA
Engineer Change.

Logistic Regression: Example Question

We are given a data set consisting of the following experiment. Well, the dataset is a little bit small. (O_o)

The height and weight of 3 people were recorded at the beginning of each person's 65th birthday. At exactly one year after each person's 65th birthday the vital status was recorded to be either alive or deceased.

Our end goal is to use logistic regression to predict the probability that a person's life expectancy is at least 66 years given their age of 65, initial vital status of alive, height, and weight (but we won't go that far here).

The data is given in the following table on the right.

Height (inches)	Weight (lbs)	Vital Status
60	155	Deceased
64	135	Alive
73	170	Alive

UCLA Engineer Change.

Logistic Regression: Example Question

Step 1: State the log-likelihood function.

Height (inches)	Weight (lbs)	Vital Status
60	155	Deceased
64	135	Alive
73	170	Alive

UCLA
Engineer Change.

Logistic Regression: Example Question

Step 1: State the log-likelihood function.

Answer:

$$\alpha_1 = -b - 155w_1 - 60w_2$$

$$\alpha_2 = -b - 135w_1 - 64w_2$$

$$\alpha_3 = -b - 170w_1 - 73w_2$$

Height (inches)	Weight (lbs)	Vital Status
60	155	Deceased
64	135	Alive
73	170	Alive

$$L = log\left(1 - \frac{1}{1 + e^{\alpha_1}}\right) + log\left(\frac{1}{1 + e^{\alpha_2}}\right) + log\left(\frac{1}{1 + e^{\alpha_3}}\right)$$

UCLA
Engineer Change.

Logistic Regression: Example Question

Step 2: State the gradients for each parameter.

Height (inches)	Weight (lbs)	Vital Status
60	155	Deceased
64	135	Alive
73	170	Alive

Logistic Regression: Example Question

Step 2: State the gradients for each parameter.

Answer:

$$L = log\left(1 - \frac{1}{1+e^{\alpha_1}}\right) + log\left(\frac{1}{1+e^{\alpha_2}}\right) + log\left(\frac{1}{1+e^{\alpha_3}}\right)$$

$$\nabla_b = -1.0 \cdot \frac{1}{1+e^{\alpha_1}} + -1.0 \cdot -\frac{e^{\alpha_2}}{1+e^{\alpha_2}} + -1.0 \cdot -\frac{e^{\alpha_3}}{1+e^{\alpha_3}}$$

$$\nabla_{w_1} = -155.0 \cdot \frac{1}{1+e^{\alpha_1}} + -135.0 \cdot -\frac{e^{\alpha_2}}{1+e^{\alpha_2}} + -170.0 \cdot -\frac{e^{\alpha_3}}{1+e^{\alpha_3}}$$

$$\nabla_{w_2} = -60 \cdot \frac{1}{1+e^{\alpha_1}} + -64.0 \cdot -\frac{e^{\alpha_2}}{1+e^{\alpha_2}} + -73.0 \cdot -\frac{e^{\alpha_3}}{1+e^{\alpha_3}}$$

Height (inches)	Weight (lbs)	Vital Status
60	155	Deceased
64	135	Alive
73	170	Alive

$$\alpha_1 = -b - 155w_1 - 60w_2$$

$$\alpha_2 = -b - 135w_1 - 64w_2$$

$$\alpha_3 = -b - 170w_1 - 73w_2$$

UCLA
Engineer Change.

Logistic Regression: Example Question

Step 3: Give the Hessian Matrix (Optional)

Height (inches)	Weight (lbs)	Vital Status
60	155	Deceased
64	135	Alive
73	170	Alive

UCLA Engineer Change.

Logistic Regression: Example Question

Step 3: Give the Hessian Matrix

$$H_b^T = \begin{bmatrix} -1.0 \cdot -1.0 \cdot -\frac{e^{\alpha_1}}{(1+e^{\alpha_1})^2} + -1.0 \cdot -1.0 \cdot -\frac{e^{\alpha_2}}{(1+e^{\alpha_2})^2} + -1.0 \cdot -1.0 \cdot -\frac{e^{\alpha_3}}{(1+e^{\alpha_3})^2} \\ -155.0 \cdot -1.0 \cdot -\frac{e^{\alpha_1}}{(1+e^{\alpha_1})^2} + -135.0 \cdot -1.0 \cdot -\frac{e^{\alpha_2}}{(1+e^{\alpha_2})^2} + -170.0 \cdot -1.0 \cdot -\frac{e^{\alpha_3}}{(1+e^{\alpha_3})^2} \\ -60 \cdot -1.0 \cdot -\frac{e^{\alpha_1}}{(1+e^{\alpha_1})^2} + -64.0 \cdot -1.0 \cdot -\frac{e^{\alpha_2}}{(1+e^{\alpha_2})^2} + -73.0 \cdot -1.0 \cdot -\frac{e^{\alpha_3}}{(1+e^{\alpha_3})^2} \end{bmatrix}$$

$$\nabla_b = -1.0 \cdot \frac{1}{1+e^{\alpha_1}} + -1.0 \cdot -\frac{e^{\alpha_2}}{1+e^{\alpha_2}} + -1.0 \cdot -\frac{e^{\alpha_3}}{1+e^{\alpha_3}}$$

$$\nabla_{w_1} = -155.0 \cdot \frac{1}{1+e^{\alpha_1}} + -135.0 \cdot -\frac{e^{\alpha_2}}{1+e^{\alpha_2}} + -170.0 \cdot -\frac{e^{\alpha_3}}{1+e^{\alpha_3}}$$

$$\nabla_{w_2} = -60 \cdot \frac{1}{1+e^{\alpha_1}} + -64.0 \cdot -\frac{e^{\alpha_2}}{1+e^{\alpha_2}} + -73.0 \cdot -\frac{e^{\alpha_3}}{1+e^{\alpha_3}}$$

$$H_{w_1}^T = \begin{bmatrix} -1.0 \cdot -155.0 \cdot -\frac{e^{\alpha_1}}{(1+e^{\alpha_1})^2} + -1.0 \cdot -135.0 \cdot -\frac{e^{\alpha_2}}{(1+e^{\alpha_2})^2} + -1.0 \cdot -170.0 \cdot -\frac{e^{\alpha_3}}{(1+e^{\alpha_3})^2} \\ -155.0 \cdot -155.0 \cdot -\frac{e^{\alpha_1}}{(1+e^{\alpha_1})^2} + -135.0 \cdot -135.0 \cdot -\frac{e^{\alpha_2}}{(1+e^{\alpha_2})^2} + -170.0 \cdot -170.0 \cdot -\frac{e^{\alpha_3}}{(1+e^{\alpha_3})^2} \\ -60 \cdot -155.0 \cdot -\frac{e^{\alpha_1}}{(1+e^{\alpha_1})^2} + -64.0 \cdot -135.0 \cdot -\frac{e^{\alpha_2}}{(1+e^{\alpha_2})^2} + -73.0 \cdot -170.0 \cdot -\frac{e^{\alpha_3}}{(1+e^{\alpha_3})^2} \end{bmatrix}$$

$$H_{w_2}^T = \begin{bmatrix} -1.0 \cdot -60 \cdot -\frac{e^{\alpha_1}}{(1+e^{\alpha_1})^2} + -1.0 \cdot -64.0 \cdot -\frac{e^{\alpha_2}}{(1+e^{\alpha_2})^2} + -1.0 \cdot -73.0 \cdot -\frac{e^{\alpha_3}}{(1+e^{\alpha_3})^2} \\ -155.0 \cdot -60 \cdot -\frac{e^{\alpha_1}}{(1+e^{\alpha_1})^2} + -135.0 \cdot -64.0 \cdot -\frac{e^{\alpha_2}}{(1+e^{\alpha_2})^2} + -170.0 \cdot -73.0 \cdot -\frac{e^{\alpha_3}}{(1+e^{\alpha_3})^2} \\ -60 \cdot -60 \cdot -\frac{e^{\alpha_1}}{(1+e^{\alpha_1})^2} + -64.0 \cdot -64.0 \cdot -\frac{e^{\alpha_2}}{(1+e^{\alpha_2})^2} + -73.0 \cdot -73.0 \cdot -\frac{e^{\alpha_3}}{(1+e^{\alpha_3})^2} \end{bmatrix}$$

UCLA
Engineer Change.

Logistic Regression: Example Question

Step 4: Assuming an initial guess of 0.25 for each parameter, write python code for finding the values of the parameters after 2 iterations using the Newton Raphson method.

$$b = 1.1346728128592689$$

$$w_1 = -2.4878423877892759$$

$$w_2 = 3.8192554544178936$$

Height (inches)	Weight (lbs)	Vital Status
60	155	Deceased
64	135	Alive
73	170	Alive

UCLA
Engineer Change.

Closed form: LR + Regularization

- Model

$$y = \sigma(X) = \frac{1}{1 + e^{-X^T \beta}}$$

- Original Objective

$$J(\beta) = -\frac{1}{n} \sum_i \left(y_i x_i^T \beta - \log\left(1 + \exp\{x_i^T \beta\}\right)\right)$$

- L2-Regularized Objective

$$J(\beta) = -\frac{1}{n} \sum_i \left(y_i x_i^T \beta - \log\left(1 + \exp\{x_i^T \beta\}\right)\right) + \lambda \sum_j \beta_j^2$$

UCLA
Engineer Change.

Sigmoid: Calculus Cheatsheet

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}},$$

where

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

$$\begin{aligned} g'(z) &= \frac{d}{dz} \frac{1}{1 + e^{-z}} \\ &= \frac{1}{(1 + e^{-z})^2} \left(e^{-z}\right) \\ &= \frac{1}{(1 + e^{-z})} \cdot \left(1 - \frac{1}{(1 + e^{-z})}\right) \\ &= g(z)(1 - g(z)). \end{aligned}$$

UCLA
Engineer Change.

Logistic Regression: Likelihood

$$
\begin{aligned}
P(y = 1 \mid x; \theta) &= h_\theta(x) \\
P(y = 0 \mid x; \theta) &= 1 - h_\theta(x)
\end{aligned}
$$

$$p(y \mid x; \theta) = (h_\theta(x))^y \, (1 - h_\theta(x))^{1-y}$$

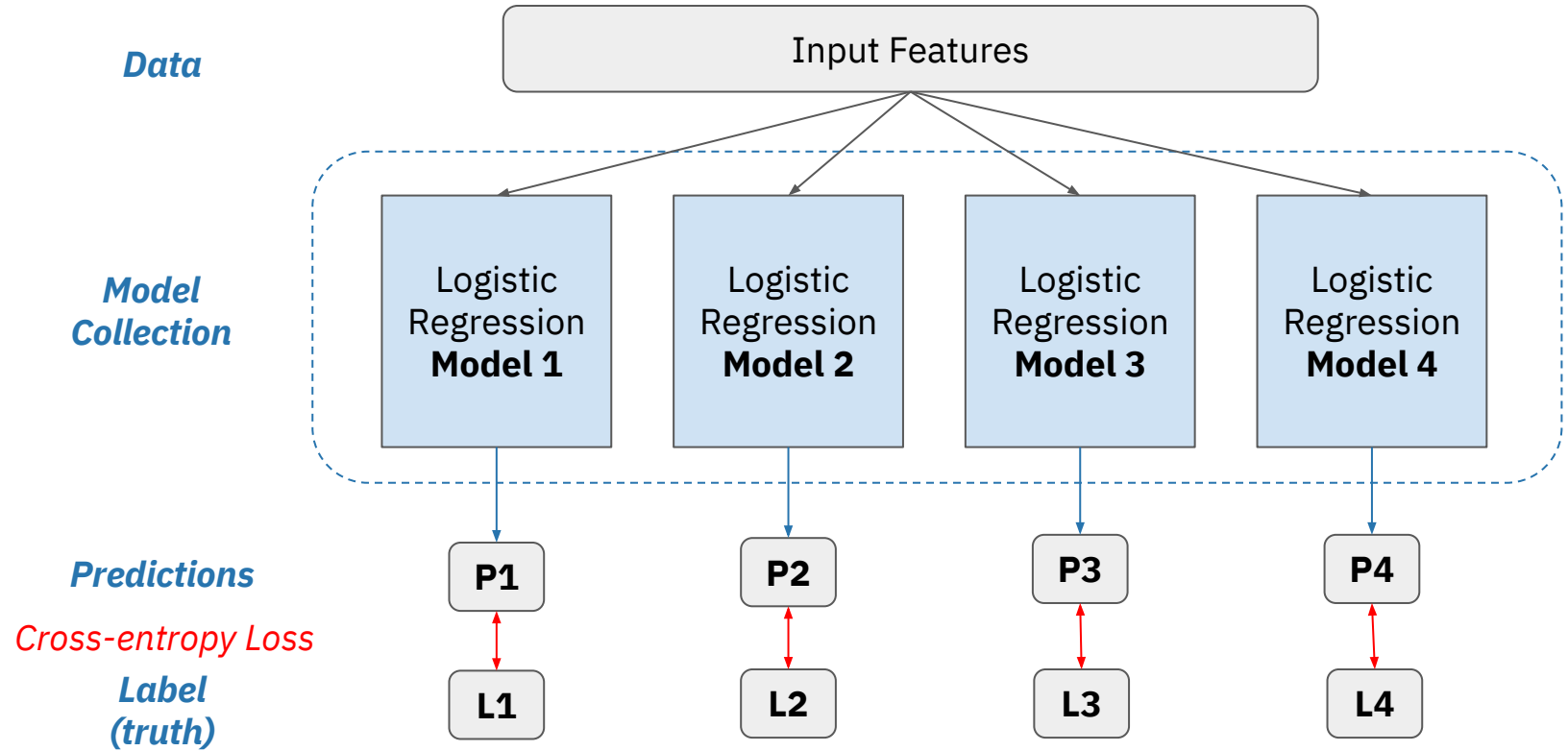
Assuming that the m training examples were generated independently, we can then write down the likelihood of the parameters as

$$
\begin{aligned}
L(\theta) &= p(\vec{y} \mid X; \theta) \\
&= \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta) \\
&= \prod_{i=1}^{m} \left(h_\theta(x^{(i)})\right)^{y^{(i)}} \left(1 - h_\theta(x^{(i)})\right)^{1-y^{(i)}}
\end{aligned}
$$

As before, it will be easier to maximize the log likelihood:

$$
\begin{aligned}
\ell(\theta) &= \log L(\theta) \\
&= \sum_{i=1}^{m} y^{(i)} \log h(x^{(i)}) + (1 - y^{(i)}) \log(1 - h(x^{(i)}))
\end{aligned}
$$

Logistic Regression: Multiclass Case*



UCLA
Engineer Change.

Linear Regression: Model

- Linear model to predict value of a variable *y* using features ***x***

$$y = \boldsymbol{x}^T \boldsymbol{\beta} = x_1\beta_1 + x_2\beta_2 + \cdots + x_p\beta_p$$

- Least Square Estimation

$$J(\boldsymbol{\beta}) = \frac{1}{2n}(X\boldsymbol{\beta} - y)^T(X\boldsymbol{\beta} - y)$$

- Closed form solution

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T y$$

UCLA
Engineer Change.

Linear Regression: Close-Form

Least Square Estimation

$$J(\boldsymbol{\beta}) = \frac{1}{2n}(X\boldsymbol{\beta} - y)^T(X\boldsymbol{\beta} - y)$$

Closed form solution

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T y$$

UCLA
Engineer Change.

Linear Regression: Example

- A ball is rolled down a hallway and its position is recorded at five different times. Use the table shown below to calculate
  - Weights
  - Predicted position at each given time and at time 12 seconds

Time (s)	Position (m)
1	9
2	12
4	17
6	21
8	26

UCLA
Engineer Change.

Linear Regression: Example

Step 1: Question

- What are X and Y variables?
- What are the parameters for our problem?
- Calculating parameters

Time (s)	Position (m)
1	9
2	12
4	17
6	21
8	26

UCLA
Engineer Change.

Linear Regression: Example

Step 1: Calculate Weights

- What are X and Y variables?
  - Time (X) and Position(Y)
- What are the parameters for our problem?
  - $\hat{\beta}_1$:Time $\hat{\beta}_0$:Intercept
- Calculating parameters
  - $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T y$

Time (s)	Position (m)
1	9
2	12
4	17
6	21
8	26

UCLA
Engineer Change.

Linear Regression: Example

Let's calculate on BOARD!

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \end{bmatrix} \qquad y = \begin{bmatrix} 9 \\ 12 \\ 17 \\ 21 \\ 26 \end{bmatrix}$$

$$X^T X = ? \qquad (X^T X)^{-1} = ?$$

$$X^T y = ? \qquad \hat{\beta} = (X^T X)^{-1} X^T y = ?$$

Time (s)	Position (m)
1	9
2	12
4	17
6	21
8	26

UCLA
Engineer Change.

Linear Regression: Example

Step 2: Apply your model and predict

- Plug time values into linear regression equation

$$\hat{y} = 2.378x + 7.012$$

- Predicted value at time = 12 secs

$$\hat{y}(x = 12) = 2.378 \times 12 + 7.012 = 35.548$$

- Matrix form to predict all other positions

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta}$$

Time (s)	Position (m)
1	9
2	12
4	17
6	21
8	26
12	35.55

Linear Regression: Example

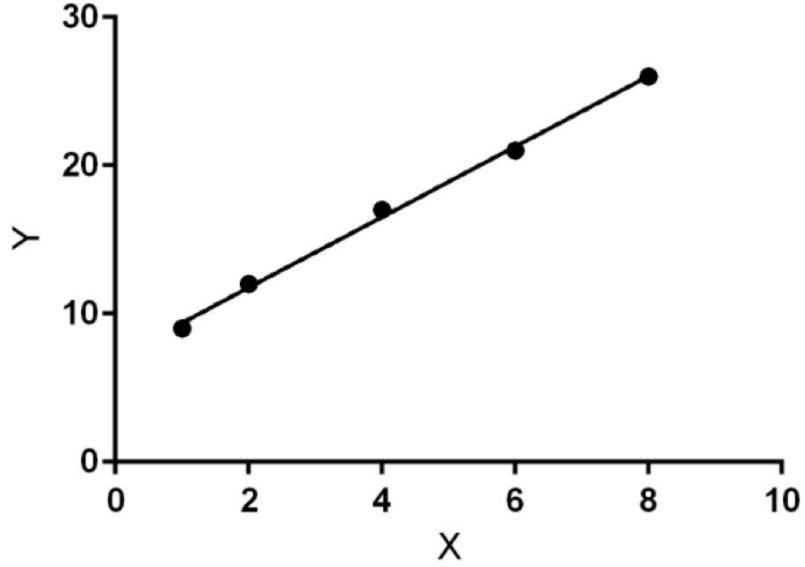
Plot: Check your model

$$\hat{y} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \end{bmatrix} \begin{bmatrix} 7.012 \\ 2.378 \end{bmatrix} = \begin{bmatrix} 9.390 \\ 11.768 \\ 16.524 \\ 21.280 \\ 26.036 \end{bmatrix}$$

Time (s)	Position (m)
1	9
2	12
4	17
6	21
8	26

Linear Regression: Example

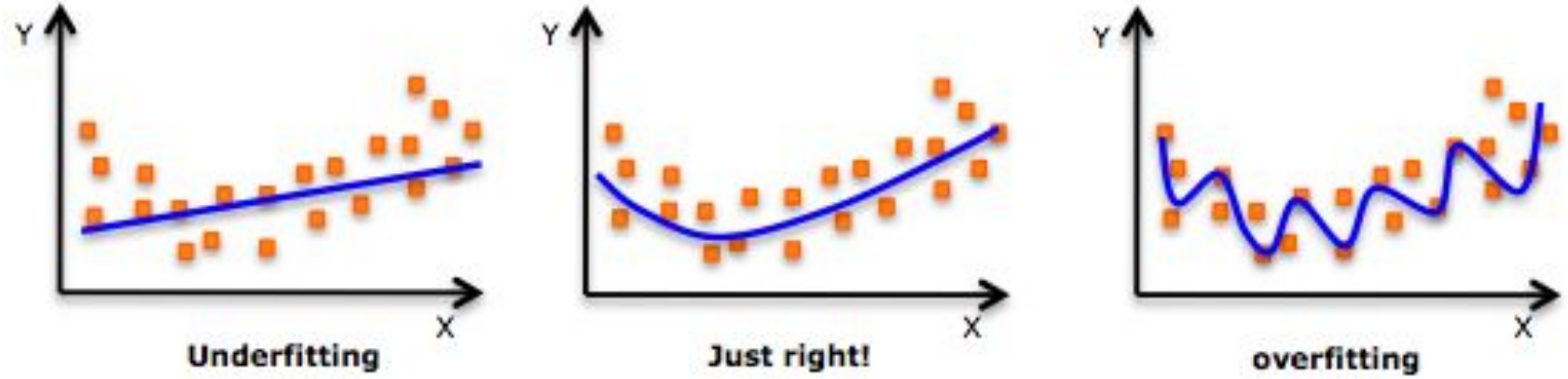
Plot: Check your model



Time (s)	Position (m)
1	9
2	12
4	17
6	21
8	26

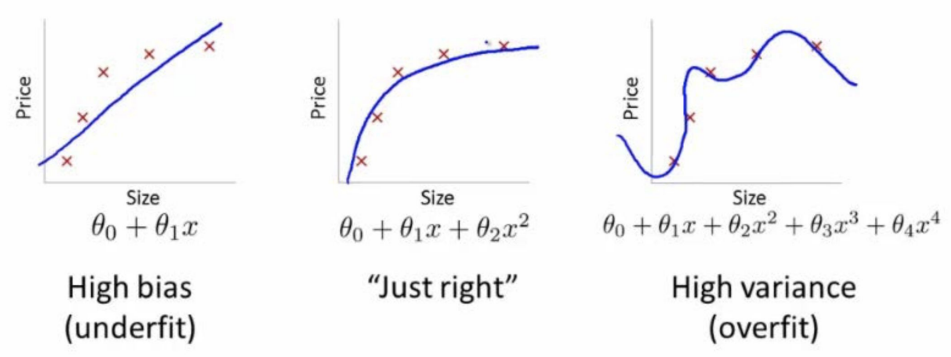
Linear Regression: Underfit & Overfit

- What is overfitting and underfitting in linear regression? → *This topic will be discussed later.*
  - How to avoid overfitting?

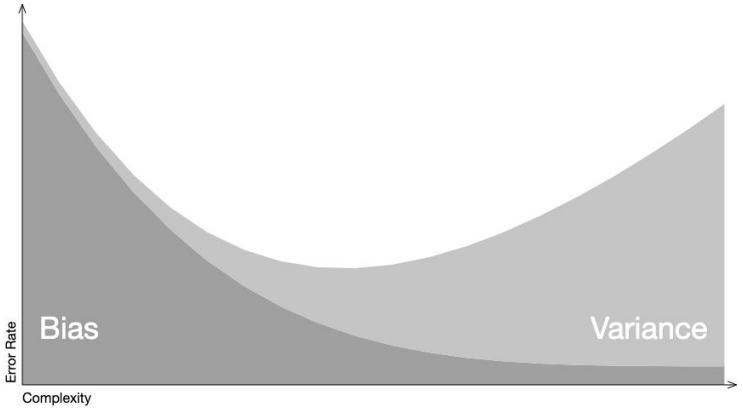


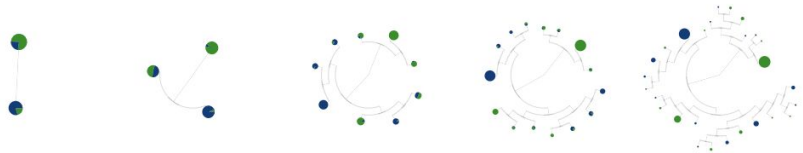
Bias vs Variance: Example

Linear Regression



Decision Tree (Week 2)



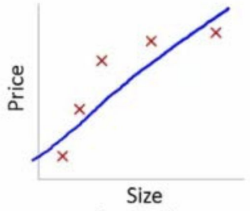


Bias vs Variance: Example

1. $h_\theta(x) = \theta_0 + \theta_1 x$
2. $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$
3. $h_\theta(x) = \theta_0 + \theta_1 x + \cdots + \theta_3 x^3$

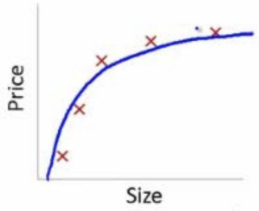
$\vdots$

10. $h_\theta(x) = \theta_0 + \theta_1 x + \cdots + \theta_{10} x^{10}$



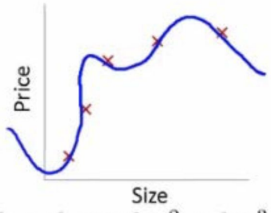
$\theta_0 + \theta_1 x$

High bias
(underfit)



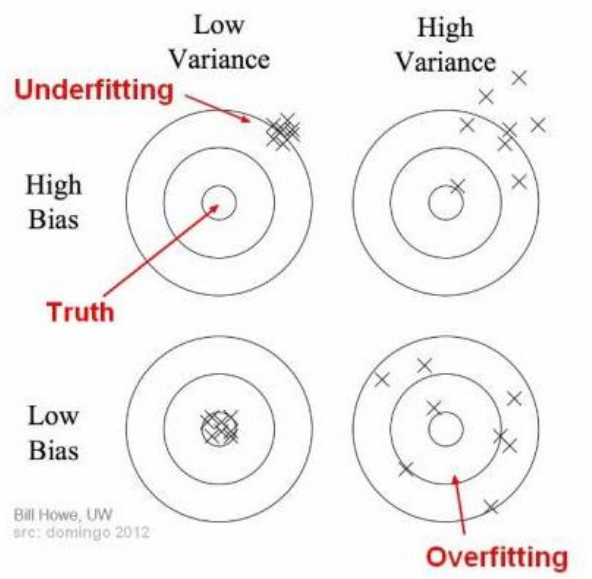
$\theta_0 + \theta_1 x + \theta_2 x^2$

"Just right"



$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

High variance
(overfit)



UCLA
Engineer Change.

Closed form: LR + Regularization

- Model

$$\hat{y} = \boldsymbol{x}^T\boldsymbol{\beta} = x_1\beta_1 + x_2\beta_2 + \cdots + x_p\beta_p$$

- Original Objective

$$\min_{\boldsymbol{\beta}} \; J(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}^T\boldsymbol{\beta} - y)^2$$

- L2-Regularized Objective

$$\min_{\boldsymbol{\beta}} \; J(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}^T\boldsymbol{\beta} - y)^2 + \frac{\lambda}{2}||\boldsymbol{\beta}||^2$$

UCLA
Engineer Change.

Closed form: LR + Regularization

$$\min_{\boldsymbol{\beta}} \; J(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}^T\boldsymbol{\beta} - y)^2 + \frac{\lambda}{2}||\boldsymbol{\beta}||^2$$

$$\frac{\partial J(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N}\boldsymbol{x}(\boldsymbol{x}^T\boldsymbol{\beta} - y) + \frac{\partial \lambda \; ||\boldsymbol{\beta}||^2}{\partial \boldsymbol{\beta}}$$

$$\frac{\partial J(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N}\boldsymbol{x}(\boldsymbol{x}^T\boldsymbol{\beta} - y) + \lambda\boldsymbol{\beta}$$

Linear Regression: Probabilistic Interpretation

Likelihood of one training sample (x_n, y_n)

$$p(y_n|x_n;\boldsymbol{\theta}) = \mathcal{N}(\theta_0 + \theta_1 x_n, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{[y_n-(\theta_0+\theta_1 x_n)]^2}{2\sigma^2}}$$

$$\begin{aligned}
\mathcal{LL}(\boldsymbol{\theta}) &= \log P(\mathcal{D}) \\
&= \log \prod_{n=1}^{\mathsf{N}} p(y_n|x_n) = \sum_n \log p(y_n|x_n) \\
&= \sum_n \left\{ -\frac{[y_n - (\theta_0 + \theta_1 x_n)]^2}{2\sigma^2} - \log \sqrt{2\pi}\sigma \right\} \\
&= -\frac{1}{2\sigma^2} \sum_n [y_n - (\theta_0 + \theta_1 x_n)]^2 - \frac{\mathsf{N}}{2} \log \sigma^2 - \mathsf{N} \log \sqrt{2\pi} \\
&= -\frac{1}{2} \left\{ \frac{1}{\sigma^2} \sum_n [y_n - (\theta_0 + \theta_1 x_n)]^2 + \mathsf{N} \log \sigma^2 \right\} + \text{const}
\end{aligned}$$

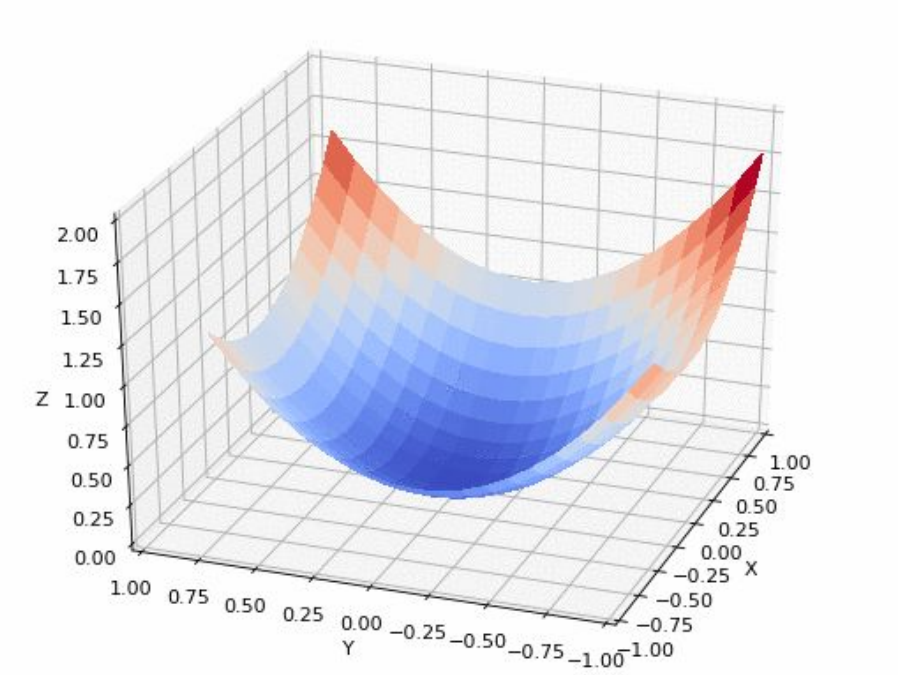
Maximize over θ_0 and θ_1

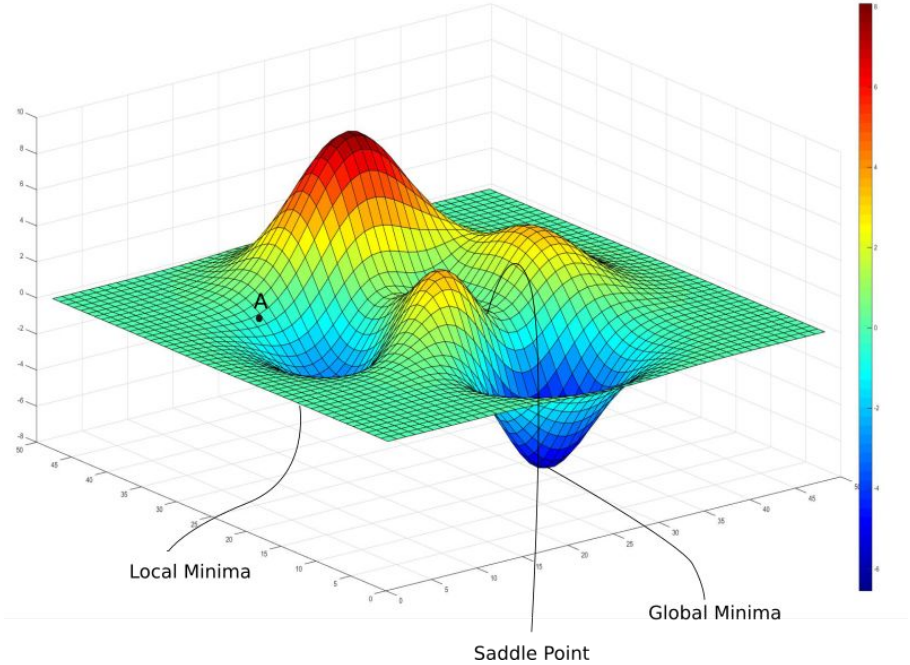
$$\max \ \log P(\mathcal{D}) \Leftrightarrow \min \ \sum_n [y_n - (\theta_0 + \theta_1 x_n)]^2$$

MLE = Least Square Error!

UCLA
Engineer Change.

Gradient Descent





Batch vs Stochastic Gradient Descent

Algorithm 1 Gradient descent

1: $\boldsymbol{\theta} \leftarrow \mathbf{0}$.
2: **for** $epoch = 1 \dots T$ **do**
3: $\quad \boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla J(\boldsymbol{\theta})$
4: **end for**
5: **return** $\boldsymbol{\theta}$

Algorithm 2 Stochastic Gradient descent

1: $\boldsymbol{\theta} \leftarrow \mathbf{0}$.
2: **for** $epoch = 1 \dots T$ **do**
3: $\quad$ **for** $(\boldsymbol{x}, y) \in \mathcal{D}$ **do**
4: $\quad\quad \boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla J_{(\boldsymbol{x},y)}(\boldsymbol{\theta})$
5: $\quad$ **end for**
6: **end for**
7: **return** $\boldsymbol{\theta}$

Randomly choosing a training sample

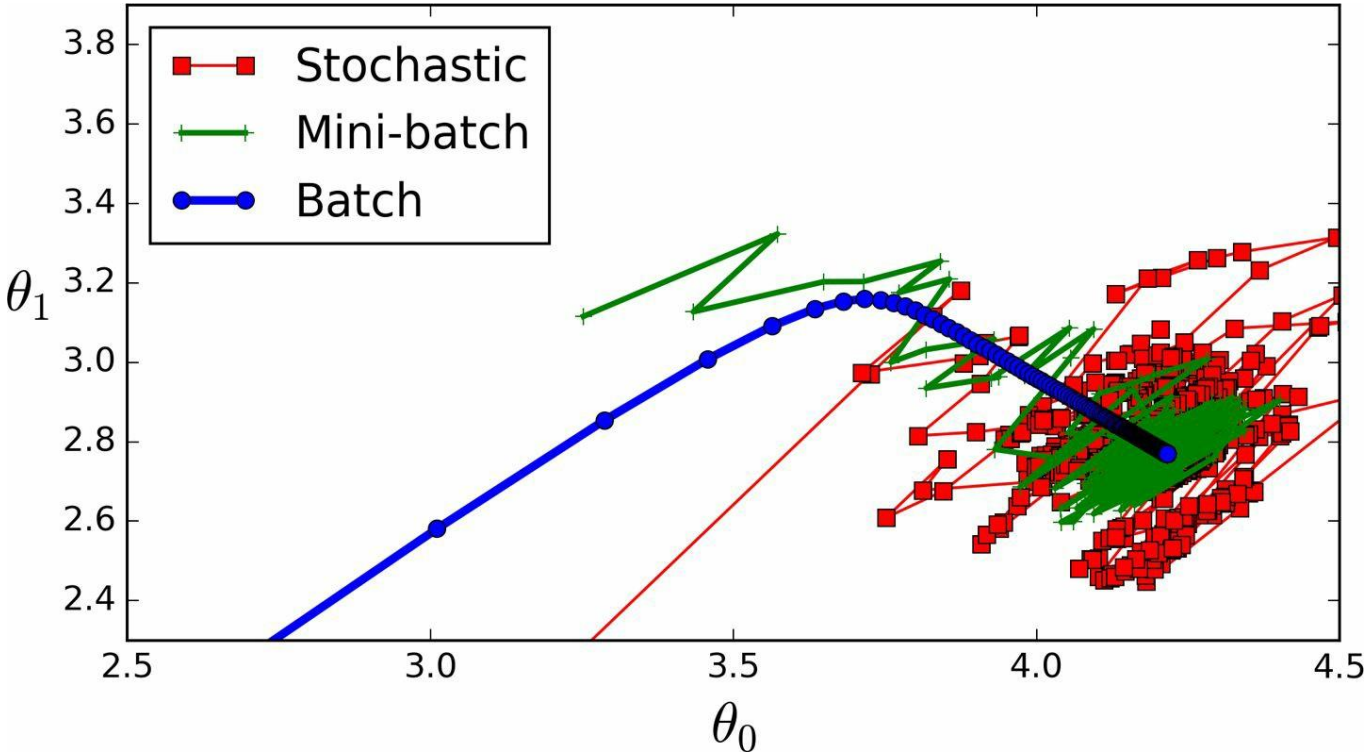
Algorithm 2 Gradient Descent (J)

1: $t \leftarrow 0$
2: Initialize $\theta^{(0)}$
3: **repeat**
4: $\quad \nabla J(\boldsymbol{\theta}^{(t)}) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\theta}^{(t)} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = \sum_n (\boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{\theta}^{(t)} - y_n)\boldsymbol{x}_n$
5: $\quad \boldsymbol{\theta}^{(t+1)} \leftarrow \boldsymbol{\theta}^{(t)} - \eta \nabla J(\boldsymbol{\theta}^{(t)})$
6: $\quad t \leftarrow t + 1$
7: **until** convergence
8: Return final value of $\boldsymbol{\theta}$

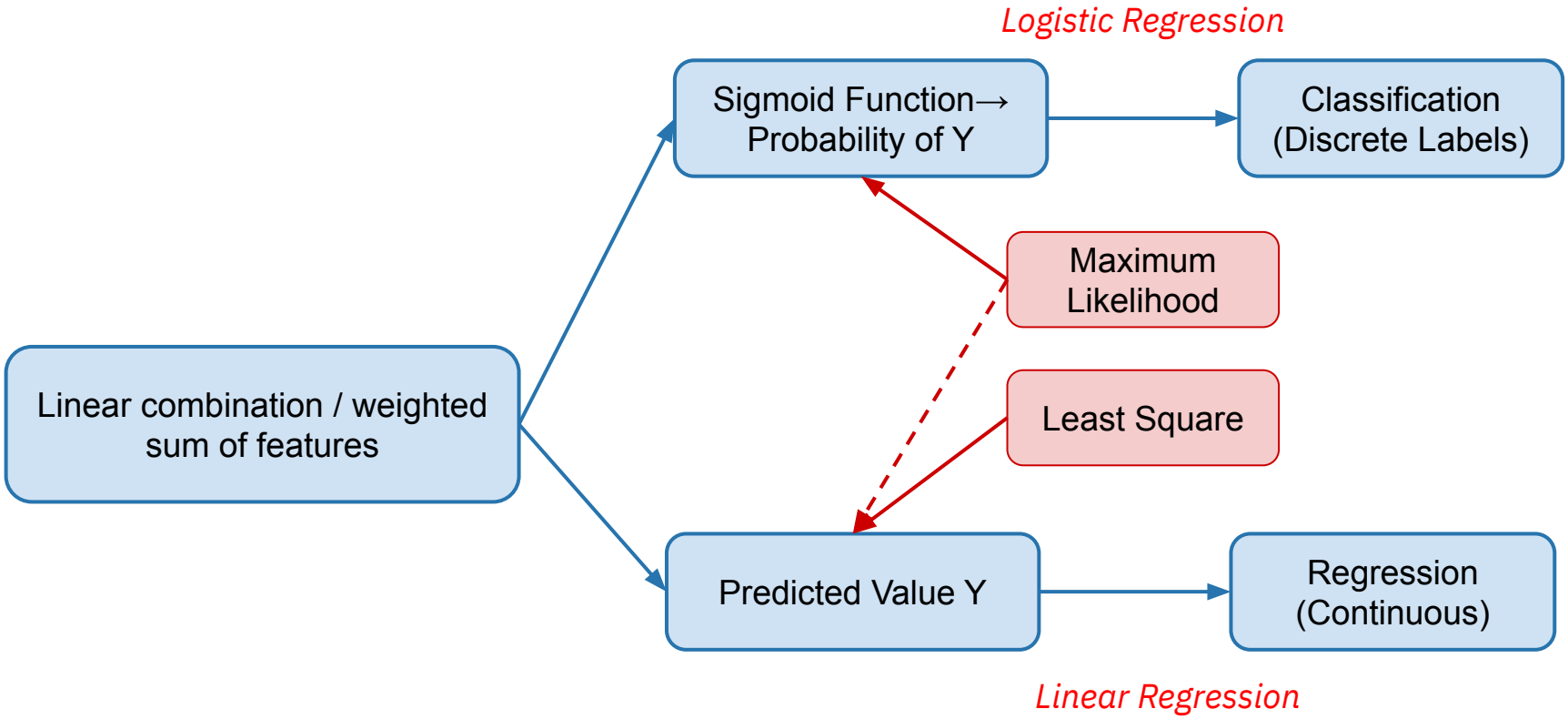
Algorithm 3 Stochastic Gradient Descent (J)

1: $t \leftarrow 0$
2: Initialize $\boldsymbol{\theta}^{(0)}$
3: **repeat**
4: $\quad$ Randomly choose a training a sample $\boldsymbol{x}_t$
5: $\quad$ Compute its contribution to the gradient $\boldsymbol{g}_t = (\boldsymbol{x}_t^{\mathrm{T}}\boldsymbol{\theta}^{(t)} - y_t)\boldsymbol{x}_t$
6: $\quad \boldsymbol{\theta}^{(t+1)} \leftarrow \boldsymbol{\theta}^{(t)} - \eta \boldsymbol{g}_t$
7: $\quad t \leftarrow t + 1$
8: **until** convergence
9: Return final value of $\boldsymbol{\theta}$

Batch / Stochastic / Mini-Batch



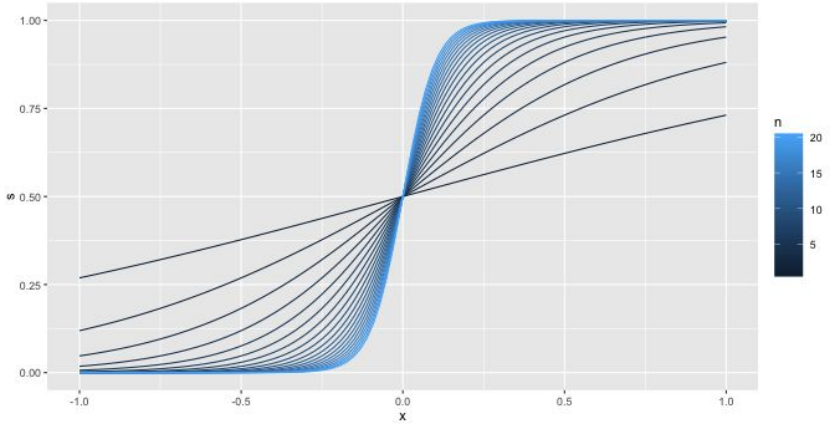
Logistic Regression VS Linear Regression



UCLA
Engineer Change.

* Logistic Regression: Convergence?

- True/False: Logistic regression cannot converge on a linearly separable dataset.



UCLA Samueli Computer Science

Thank you!

Q & A