- Different types of machine learning
- Linear models
- Terminology in linear model problems
- 1-dimensional OLS and minimizing the loss function
- Interpreting parameters in an OLS equation
- Limitations of using OLS through Anscombe's Quartet
- Multivariate OLS models and an example problem
- Using matrix foundation to solve the OLS problem
- The pros and cons of using OLS regression

Learning Objectives

- Understand ordinary least squares (OLS)

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Types of Machine Learning

- Unsupervised Learning
  - Finding structure in data
- Supervised Learning
  - Predict from given data

Introduction

- Objective
  - What is a linear model?
  - Why use linear models?

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Example of Use

- Example of Use
- Building Blocks of Later Techniques

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Terminology

aws training and certification

- Definitions
  - Features (covariates or predictors)
  - Labels (variates or targets)
  - Regression
  - Classification

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definition of 1-Dimension OLS

aws training and certification

- The Problem Statement
- The Loss Function

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Solution of OLS

- Derivation
- Optimal Choices

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Interpretation

- Interpretation of a
- Interpretation of L

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Anscombe's Quartet

- Anscombe's Quartet

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Multivariable OLS

aws training and certification

- Definition of Model
  - Data matrix
  - The loss function

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Optimization

- A Few Facts from Matrix Calculus
- Minimizing the Loss

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

OLS Pros and Cons

- OLS
  - Pros
  - Cons

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Types of Machine Learning

To get us started, the video below explains classifying training data so that humans can ide

the data structures, or make predictions from the data.

Two types of machine learning covered in this video include:

Unsupervised Learning

Supervised Learning

Topics

- Unsupervised Learning
  - Finding structure in data
- Supervised Learning
  - Predict from given data

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Unsupervised Learning

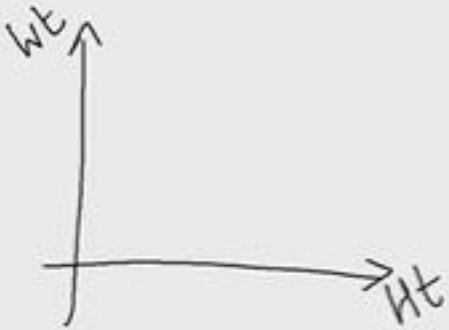
aws training and certification

Finding structure in data

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Supervised Learning

Predict from given data

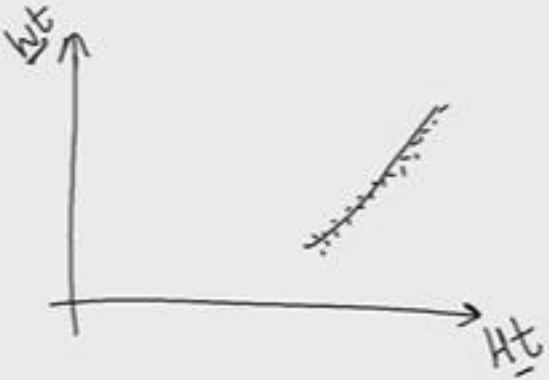


© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Supervised Learning

Predict from given data

Classification



© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Supervised Learning

aws training and certification

Ordinary least squares (OLS) and **logistic regression** are methods of these two types of supervised learning problem.

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Supervised Learning

aws training and certification

Ordinary least squares (OLS) and **logistic regression** are methods of these two types of supervised learning problem.

OLS - Regress

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Supervised Learning

aws training and certification

Ordinary least squares (OLS) and **logistic regression** are methods of these two types of supervised learning problem.

OLS - Regression

Log. Reg - Clo

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Supervised Learning

aws training and certification

Ordinary least squares (OLS) and **logistic regression** are methods of these two types of supervised learning problem.

OLS - Regression

Log. Reg - Classification

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Supervised Learning

aws training and certification

Ordinary least squares (OLS) and **logistic regression** are methods of these two types of supervised learning problem.

OLS - Regression

Log. Reg - Classification

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Introduction to Linear Models

This video explains how linear models look and what they are used for. Take a moment to explore linear models before you begin.

What is a linear model?

In linear modeling, the relationship between each individual input variables and the output is a straight line. Slopes of such lines become the coefficients of the linear equation.

An example of linear notation

$$y_i = a_0 + a_1x_1 + a_2x_2 + \ldots$$

An example of linear notation

$$y_i = a_0 + a_1x_1 + a_2x_2 + \ldots$$

Why use linear models?

- Interpretable
- Low complexity
- Scalable

What is a Linear Model?

aws training and certification

$$\text{Housing price} = k_1(\text{Area}) + k_2(\#\text{bedrooms}) + k_3(\#\text{bathrooms})$$

Housing prices

- Area
- #bedrooms
- #bathrooms

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

What is a Linear Model?

aws training and certification

$$\text{Housing price} = k_1(\text{Area}) + k_2(\#\text{bedrooms}) + k_3(\#\text{bathrooms})$$

Housing prices

- Area
- #bedrooms
- #bathrooms

$$y_i = a_0 + a_1 x_1 + a_2 x_2 + \cdots$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Why Use Linear Models?

aws training and certification

- Interpretable
- Low complexity
- Scalable
- Baseline

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Example of Use

Machines encounter certain problems when trying to identify handwritten digits. Before watching the video below, take a moment to compare the performance of a regression model and a neural network model with regard to this challenge.

Handwriting Digit Recognition

MNIST Example

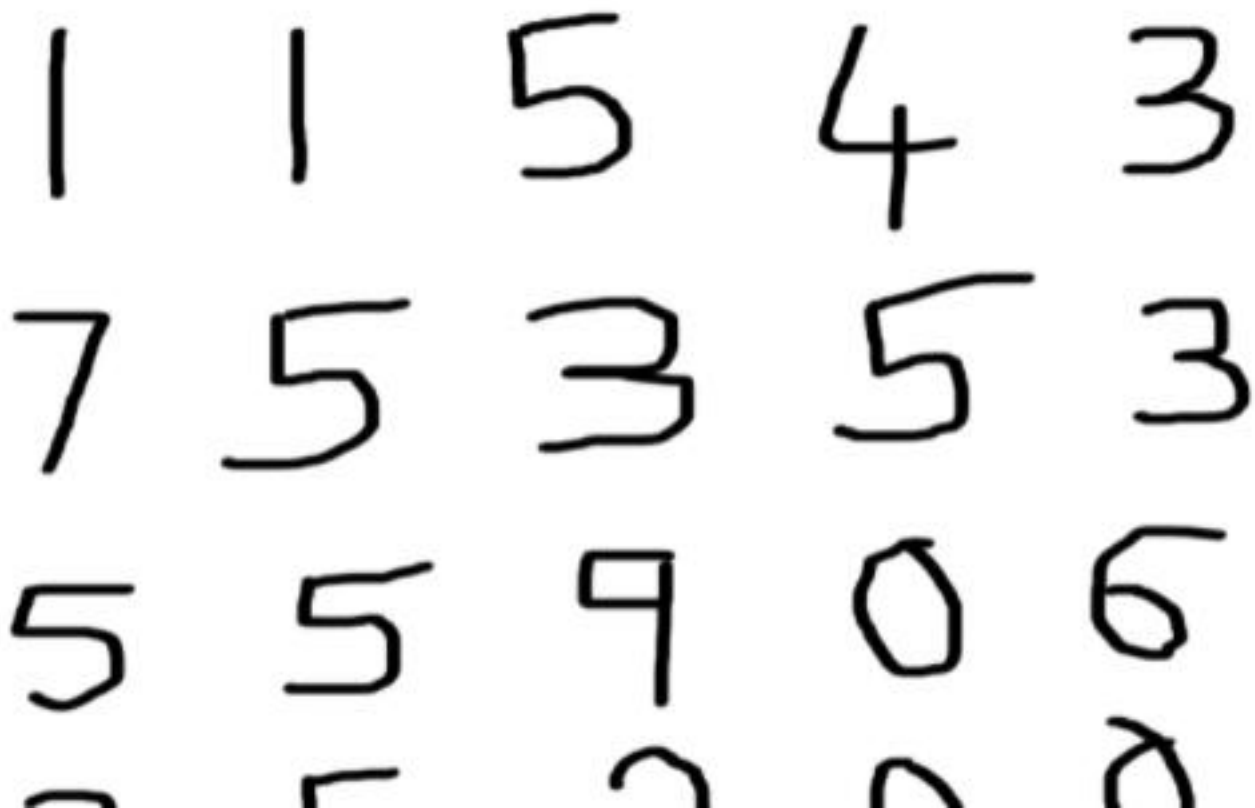
Handwriting Digit Recognition

MNIST Example

THE CHALLENGE | A NEURAL NETWORK SOLUTION | A LOGISTIC REGRESSION SOLUTION

The machine learning task is to identify what the digits are.

**data source: Modified National Institute of Standards and Technology(MNIST) database hand written digits.*



Handwriting Digit Recognition

MNIST Example

THE CHALLENGE	A NEURAL NETWORK SOLUTION	A LOGISTIC REGRESSION SOLUTION

Uses advanced regularization techniques via a neural network

- Is 99.79 % error free
- Takes about a day to train
- Requires significant expertise to achieve

Handwriting Digit Recognition

MNIST Example

THE CHALLENGE	A NEURAL NETWORK SOLUTION	A LOGISTIC REGRESSION SOLUTION

Uses logistic regression

- Is 92.5 % error free
- Takes seconds to train
- Can be done with (comparatively) less expertise

Example of Use

aws training and certification

MNIST is a common data set that contains images of handwritten digits. The machine learning task is to identify what the digits are.

*Data Source: Modified National Institute of Standards and Technology (MNIST) Database of Handwritten Digits

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Example of Use

aws training and certification

Logistic regression

- Is 92.5 percent error free
- Takes seconds to train
- Can be done with (comparatively) less expertise

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Building Blocks for Later Techniques

aws training and certification

More advanced techniques like neural networks are, mathematically, a little more complex than **stacking** logistic regression.

© 2016, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Dataset References

aws training and certification

- Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition." Proceedings of the IEEE, 86(11):2278-2324, November 1998. Dataset available from http://yann.lecun.com/exdb/mnist/

The Problem Statement

$i = 1, \ldots, N$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Problem Statement

$i = 1, \ldots, N$, x = indepent var. y = dep. var.

$$y = ax + b$$

a, b constants.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Problem Statement

$i = 1, \ldots, N$, x = indepent var. y = dep. var.

$$y = ax + b$$

a, b constants

$$\hat{y}_i = ax_i + b$$

$$y_i = ax_i + b + \epsilon$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Loss Function

$$\mathcal{L} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Solutions a and b:

For an optimized model the loss function should be less.

Find the L minimum by equating the first derivatives of L to zero, and solving for the values of a and b for which L is minimum.

To best fit a collection matched input output pairs $\{x_i\}_{i=1}^{N}$ and $\{y_i\}_{i=1}^{N}$ with a function of the form $y = ax + b$ to minimize the mean squared error, you should pick:

$$a = \frac{\sum_i x_i y_i - \frac{1}{N}(\sum_i x_i)(\sum_i y_i)}{\sum_i x_i^2 - \frac{1}{N}(\sum_i x_i)^2} \qquad \left[= \frac{Cov(X,Y)}{Var(X)}\right]$$

$$b = \frac{1}{N}\sum_i y_i - \frac{a}{N}\sum_i x_i \qquad [= E[Y] - aE[X]]$$

These are exact solutions to the minimization problem.

Derivation

$$L = \sum_{i=1}^{N} (y_i - ax_i - b)^2$$

$$\underline{\min L}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Derivation

$$L = \sum_{i=1}^{N} (y_i - ax_i - b)^2$$

$$\underline{\min L}$$

$$\frac{dL}{da} = 0 \quad , \quad \frac{dL}{db} = 0$$

$$\frac{dL}{da} = 0 \Rightarrow \sum_{i=1}^{N} 2(y_i - ax_i - b)($$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Derivation

$$L = \sum_{i=1}^{N} (y_i - ax_i - b)^2$$

$$\underline{\min L}$$

$$\frac{dL}{da} = 0 \ , \ \frac{dL}{db} = 0$$

$$\frac{dL}{da} = 0 \Rightarrow \sum_{i=1}^{N} 2(y_i - ax_i - b)(+x_i) = 0$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Derivation

aws training and certification

$$L = \sum_{i=1}^{N} (y_i - ax_i - b)^2$$

$$\underline{\min L}$$

$$\frac{dL}{da} = 0 \quad , \quad \frac{dL}{db} = 0$$

$$\frac{dL}{da} = 0 \Rightarrow \sum_{i=1}^{N} 2(y_i - ax_i - b)(+x_i) = 0$$

$$\sum_{i=1}^{N} x_i y_i - a \sum_{i=1}^{N} x_i^2 .$$

-4:16

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Derivation

$$\frac{dL}{db} = 0 \Rightarrow \sum_{i=1}^{N} +\cancel{2}(y_i - ax_i - b)(+1) = 0$$

$$\sum y_i - a\sum x_i - bN = 0$$

$$b = \frac{1}{N}\sum$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Derivation

aws training and certification

$$\frac{dL}{db} = 0 \Rightarrow \sum_{i=1}^{N} +\cancel{2}(y_i - ax_i - b)(+1) = 0$$

$$\sum y_i - a\sum x_i - bN = 0$$

$$\boxed{b = \frac{1}{N}\sum y_i - \frac{a}{N}\sum x_i} \quad -②$$

-2:37

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Derivation

$$\frac{dL}{db} = 0 \Rightarrow \sum_{i=1}^{N} +\cancel{2}(y_i - ax_i - b)(+1) = 0$$

$$\sum y_i - a\sum x_i - bN = 0$$

$$\boxed{b = \frac{1}{N}\sum y_i - \frac{a}{N}\sum x_i} \quad -②$$

From ①, ②

$$\sum_{i=1}^{N} x_i y_i - a\sum_{i=1}^{N} x_i^2 - b\sum x_i = 0$$

$$\Rightarrow \sum x_i y_i = a\sum x_i^2 + \sum x_i\left(\frac{1}{N}\sum y_i - \frac{a}{N}\sum x_i\right)$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Derivation

$$\frac{dL}{db} = 0 \Rightarrow \sum_{i=1}^{N} +\cancel{2}(y_i - ax_i - b)(+1) = 0$$

$$\sum y_i - a\sum x_i - bN = 0$$

$$\boxed{b = \frac{1}{N}\sum y_i - \frac{a}{N}\sum x_i} \quad — ②$$

From ①,

$$\sum_{i=1}^{N} x_i y_i - a\sum_{i=1}^{N} x_i^2 - b\sum x_i = 0$$

$$\Rightarrow \sum x_i y_i = a\sum x_i^2 + \sum x_i\left(\frac{1}{N}\sum y_i - \frac{a}{N}\sum x_i\right)$$

$$a = \frac{\sum x_i y_i - \frac{1}{N}\sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{N}\left(\sum x_i\right)^2}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Optimal Choices

aws training and certification

To best fit a collection matched input output pairs $\{x_i\}_{i=1}^N$ and $\{y_i\}_{i=1}^N$ with a function of the form $y = ax + b$ to minimize the mean squared error, you should pick:

$$a = \frac{\sum_i x_i y_i - \frac{1}{N}(\sum_i x_i)(\sum_i y_i)}{\sum_i x_i^2 - \frac{1}{N}(\sum_i x_i)^2} \qquad \left[= \frac{Cov(X,Y)}{Var(X)}\right]$$

$$b = \frac{1}{N}\sum_i y_i - \frac{a}{N}\sum_i x_i \qquad [= E[Y] - aE[X]]$$

These are **exact** solutions to the minimization problem.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Optimal Choices

aws training and certification

To best fit a collection matched input output pairs $\{x_i\}_{i=1}^{N}$ and $\{y_i\}_{i=1}^{N}$ with a function of the form $y = ax + b$ to minimize the mean squared error, you should pick:

$$a = \frac{\sum_i x_i y_i - \frac{1}{N}(\sum_i x_i)(\sum_i y_i)}{\sum_i x_i^2 - \frac{1}{N}(\sum_i x_i)^2} \qquad \left[= \frac{Cov(X,Y)}{Var(X)}\right]$$

$$b = \frac{1}{N}\sum_i y_i - \frac{a}{N}\sum_i x_i \qquad [= E[Y] - aE[X]]$$

These are **exact** solutions to the minimization problem.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Interpretation

In the last video, you found the solutions for a and b. Now you will interpret a and b on a linear graph of x (input parameters) and y (output) to obtain a generalized condition for predicting y. You will also analyze the loss function with a and b values to understand a model's performance.

Interpretation of a and b

a – slope of the line (size of relationship between X and Y)
b – intercept

$$i.e.\ \ b = \begin{cases} \hat{y}, & x = 0 \\ 0, & x = \hat{x} \end{cases}$$

Interpretation of L

In general, $R^2 \in [0,1]$

Let's observe the model performance with the help of R^2 which is defined as. Keep in mind, L = loss function.

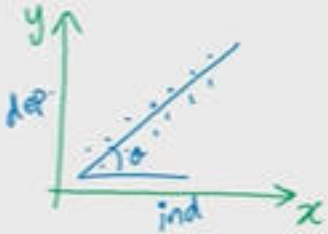
$$R^2 = 1 - \frac{Loss\ Function}{Var\ (y)} = 1 - \frac{MSE}{\bar{y}}$$

Interpretation of *a*, *b*

aws training and certification

a = slope

→ effect



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

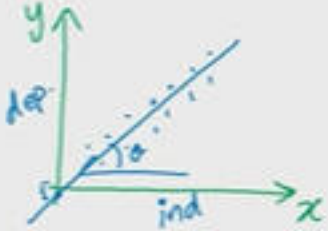
Interpretation of *a*, *b*

aws training and certification

a = slope

→ effect size

b = intercept



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Interpretation of *a*, *b*

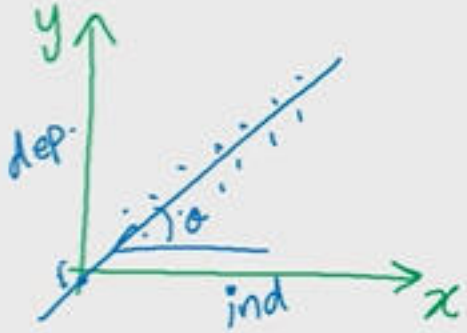
aws training and certification

a = slope

→ effect size.

b = intercept

→ $\hat{y} = b$ when



-3:46

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Interpretation of L

$$L = \sum_{i=1}^{N} (y_i - \hat{y})^2$$

$$\hookrightarrow R^2$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Interpretation of L

aws training and certification

MSE, $L = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$

$\hookrightarrow R^2 = 1 - \frac{MSE}{Var(y)}$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Interpretation of L

aws training and certification

MSE, $L = \sum_{i=1}^{N} (y_i - \hat{y})^2$

$\hookrightarrow R^2 = 1 - \frac{MSE}{Var(Y)}$ $(\hat{Y})$ $(\bar{Y})$

$R^2 \in [0, 1]$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Topics

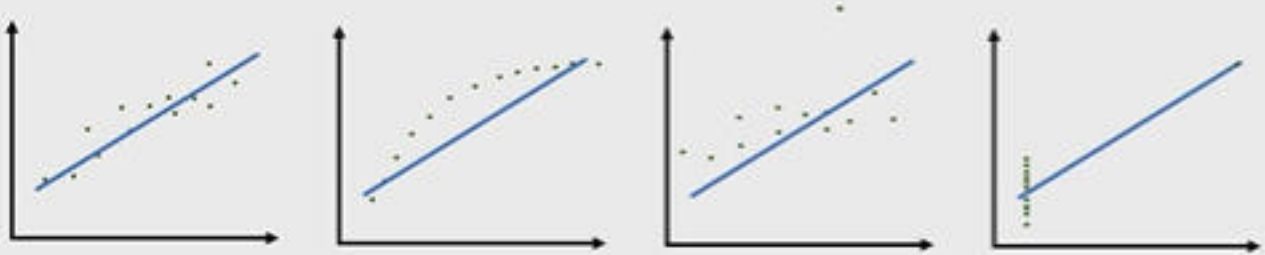
- Anscombe's Quartet

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Anscombe's Quartet

aws training and certification

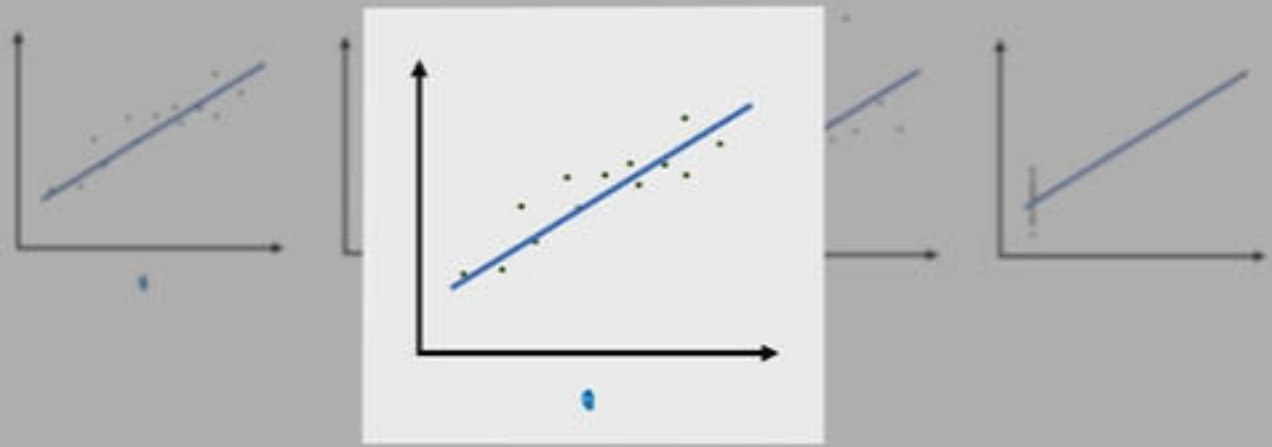
R^2 can be misleading. The following all have the same best fit line:



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Anscombe's Quartet

R^2 can be misleading. The following all have the same best fit line:



Extending to Many Variables

aws training and certification

Data matrix

$i = 1, \ldots, N$

$\vec{x}_i = [x_{i1}, x_{i2}, \ldots, x_{iM}]$ ← M features

indep var.

y_i ← dep variable

data

X

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Extending to Many Variables

aws training and certification

Data matrix $i = 1, \ldots, N$

$\vec{x}_i = [x_{i1}, x_{i2}, \ldots, x_{iM}]$ ← M features

indep var.

y_i ← dep variable

$\underline{\text{Lin}}$

$y = \underline{a}x + \underline{b}(i)$

data matrix, X =

$$\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1M} \\ \vdots & \vdots & \ddots & & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{NM} \end{bmatrix}$$

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_M x_{iM} =$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Extending to Many Variables

The loss function

$$\min L = \sum_{i=1}^{N} \left(y_i - \hat{y}_i \right)^2$$

$$= \sum_{i=1}^{N} \left(y_i - \left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_M x_{iM} \right) \right)^2$$

$$= \sum_{i=1}^{N} \left(y_i - \vec{x}_i \cdot \vec{\beta} \right)$$

Extending to Many Variables

aws training and certification

The loss function

$\underline{\min L}$

$$L = \sum_{i=1}^{N} \left(y_i - \hat{y}_i \right)^2$$

$$= \sum_{i=1}^{N} \left(y_i - \left(\beta_0 + \beta_1 x_{i1} + \dots + \beta_M x_{iM} \right) \right)^2$$

$$= \sum_{i=1}^{N} \left(y_i - \vec{x}_i \cdot \vec{\beta} \right)^2$$

$$= \|$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Extending to Many Variables

aws training and certification

The loss function

min

$$L = \sum_{i=1}^{N} \left(y_i - \hat{y}_i \right)^2$$

$$= \sum_{i=1}^{N} \left(y_i - \left(\beta_0 + \beta_1 x_{i1} + \dots + \beta_M x_{iM} \right) \right)^2$$

$$= \sum_{i=1}^{N} \left(y_i - \vec{x}_i \cdot \vec{\beta} \right)^2$$

$$= \| \vec{y} - X\vec{\beta} \|_2^2$$

$\| \cdot \|_2 \rightarrow$ L2 norm of vec

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Extending to Many Variables

aws training and certification

The loss function

min L

$$L = \sum_{i=1}^{N} \left(y_i - \hat{y}_i \right)^2$$

$$= \sum_{i=1}^{N} \left(y_i - \left(\beta_0 + \beta_1 x_{i1} + \dots + \beta_M x_{iM} \right) \right)^2$$

$$= \sum_{i=1}^{N} \left(y_i - \vec{x}_i \cdot \vec{\beta} \right)^2$$

$$= \| \vec{y} - X\vec{\beta} \|_2^2$$

$\| \cdot \|_2 \rightarrow$ L2 norm of vector.

$$= (y - X\beta)^T (\cdot$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

A Few Facts from Matrix Calculus

aws training and certification

$$\frac{\partial[X\vec{\beta}]}{\partial\vec{\beta}} = X^T \longleftrightarrow \frac{d(ax)}{dx} = a$$

$$\frac{\partial[\vec{\beta}^T A\vec{\beta}]}{\partial\vec{\beta}} = 2A^T\vec{\beta} \qquad \frac{d(ax^2)}{dx}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Minimizing the Loss

aws training and certification

$$L = (y - X\beta)^T (y - X\beta)$$

$$\frac{dL}{d\beta} = 0 \Rightarrow \frac{d}{d\beta}($$

-2:49

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Minimizing the Loss

aws training and certification

$$L = (y - X\beta)^T (y - X\beta)$$

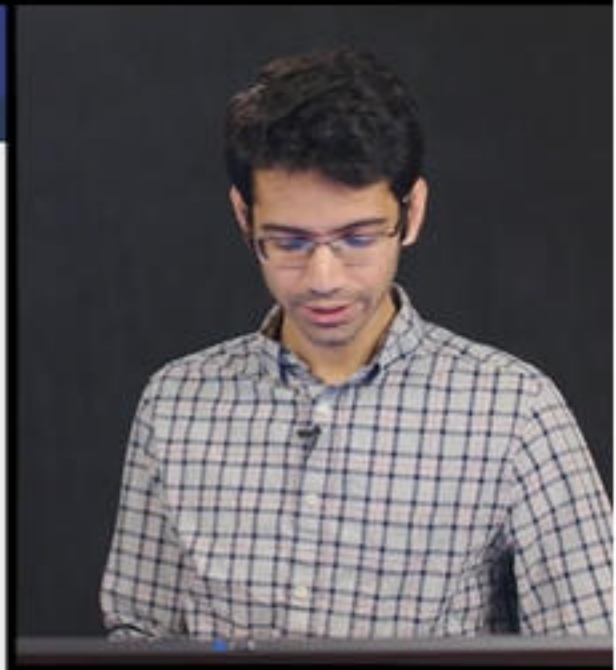
$$\frac{dL}{d\beta} = 0 \Rightarrow \frac{d}{d\beta}\left((y - X\beta)^T (y - X\beta)\right) = 0$$

$$\Rightarrow \frac{d}{d\beta}\left(\underbrace{y^T y}_{0} - \underbrace{y^T X\beta} - \underbrace{\beta^T X^T y} + \underbrace{\beta^T X^T X\beta}\right) = 0$$

$$\Rightarrow -(X^T y) - (X^T y) + 2(X^T X)\beta = 0$$

$$\left((XY)^T = Y^T X^T \right)$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



-1:14

OLS Pros

- Efficient computation ✓
- Unique minimum ✓
- Stable under perturbation of data ✓
- Easy to interpret ✓

© 2018 Amazon Web Services, Inc. or its Affiliates. All rights reserved.

OLS Cons

- Influenced by outliers
- $(X^T X)^{-1}$ need not exist

X cols/features are lin. indep.

© 2018 Amazon Web Services, Inc or its Affiliates. All rights reserved.

Maximum Likelihood Estimate (MLE)

aws training and certification

- Review: Maximum Likelihood Estimates
- A Note on Logarithms
- Definition
- Goal
- A Significant Issue

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Minimize MSE $\Leftrightarrow$ MLE

- Relation to the Previous Problem
- Solution (As Seen Last Time)

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Maximum Likelihood Estimate (MLE)

This video looks at OLS and loss function under a probabilistic light. The video also reviews MLE but first, take a moment to consider the following:

In an OLS model, the choice of least squares is arbitrary.

Why are we only choosing the power of 2, and not the sum of cubes or other degree values?

$$L = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$$

$$\hat{\theta} = \underbrace{arg}_{\theta} \underbrace{max\ L(\theta; Y)}_{\theta \text{ that maximizes L}}$$

Viewing OLS in a Probabilistic Light

$$L = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Viewing OLS in a Probabilistic Light

$$L = \sum_{i=1}^{N} \left(y_i - \hat{y}_i \right)^2$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Viewing OLS in a Probabilistic Light

aws training and certification

$$L = \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2$$

arbitrary?

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Review: MLE

We will review MLE using the example of coin flips.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Review: MLE

We will review MLE using the example of coin flips.

$p \leftarrow$ heads probability

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Review: MLE

We will review MLE using the example of coin flips.

$p \leftarrow$ heads probability

$\{HTHHHT\}$

$$P(Y) = P(\{HTHHHT\})$$
$$= P(H)P(T)P(H)P(H)P(H)P(T)$$
$$= p^4(1-p)^2$$

$$\left(\begin{aligned} P(E_1 E_2) &= P(E_1)P(E_2) \end{aligned} \right)$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Review: MLE

We will review MLE using the example of coin flips.

$p \leftarrow$ heads probability

$\{HTHHHT\}$

$$P(Y) = P(\{HTHHHT\})$$

$$= P(H)P(T)P(H)P(H)P(H)P(T)$$

$$\left(P(E_1 E_2) = P(E_1)P(E_2) \right)$$

$$L = \underbrace{p^4(1-p)^2}$$

$$\frac{dL}{dp} = 0 \Rightarrow 2\cancel{p^3}(1-p)^2 + 2p^{\cancel{4}}(\cancel{1-p})(-1) = 0$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Review: MLE

We will review MLE using the example of coin flips.

$p \leftarrow$ heads probability

$\{HTHHHT\}$

$$P(Y) = P(\{HTHHHT\})$$

$$= P(H)P(T)P(H)P(H)P(H)P(T)$$

$$\left(P(E_1 E_2) = P(E_1)P(E_2) \right)$$

$$L = \underbrace{p^4(1-p)^2}$$

$$\frac{dL}{dp} = 0 \Rightarrow 2\cancel{p^3}(1-p)^2 + 2p^4(\cancel{1-p})(-1) = 0$$

$$2(1-p) - p = 0$$

$$\Rightarrow \boxed{p = 2/3}$$

-4:31

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

A Note on Logarithms

In probability, probabilities of independent events **multiply**. This means likelihoods of data is often a **giant product**.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

A Note on Logarithms

In probability, probabilities of independent events **multiply**. This means likelihoods of data is often a **giant product**.

$$\log(AB) = \log A + \log B$$

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

A Note on Logarithms

In probability, probabilities of independent events **multiply**. This means likelihoods of data is often a **giant product**.

$$\log(AB) = \log A + \log B$$

$$L = p^4(1-p)^2 \rightarrow \log(L) = 4\log p + 2\log$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

A Note on Logarithms

aws training and certification

In probability, probabilities of independent events **multiply**. This means likelihoods of data is often a **giant product**.

$$\log(AB) = \log A + \log B$$

$$L = p^4(1-p)^2 \rightarrow \log(L) = 4\log(p) + 2\log(1-p)$$

$$\frac{d}{dp}\log(L) = 0$$

-2:40

A Note on Logarithms

In probability, probabilities of independent events **multiply**. This means likelihoods of data is often a **giant product**.

$$\log(AB) = \log A + \log B$$

$$L = p^4(1-p)^2 \rightarrow \log(L) = 4\log(p) + 2\log(1-p)$$

$$\frac{d}{dp}\log(L) = 0 \Rightarrow \boxed{p = 2/3}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definition

MLE . $\hat{\theta} = \arg\max$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definition

MLE . $\hat{\theta} = \arg\max_{\theta} L(\theta; Y)$

θ maximize

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definition

aws training and certification

MLE . $\hat{\theta} = \arg\max_{\theta} L(\theta; Y)$

$\theta \rightarrow$ maximize

$= \mathbb{P}(\{Y_i ; i=1 \dots N\} \mid \theta$

-1:15

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definition

aws training and certification

MLE . $\hat{\theta} = \arg\max_{\theta} L(\theta; Y)$

$\theta \rightarrow$ maximize

$= \mathbb{P}_{\theta}(\{Y_i ; i=1 \dots N\})$

-1:05

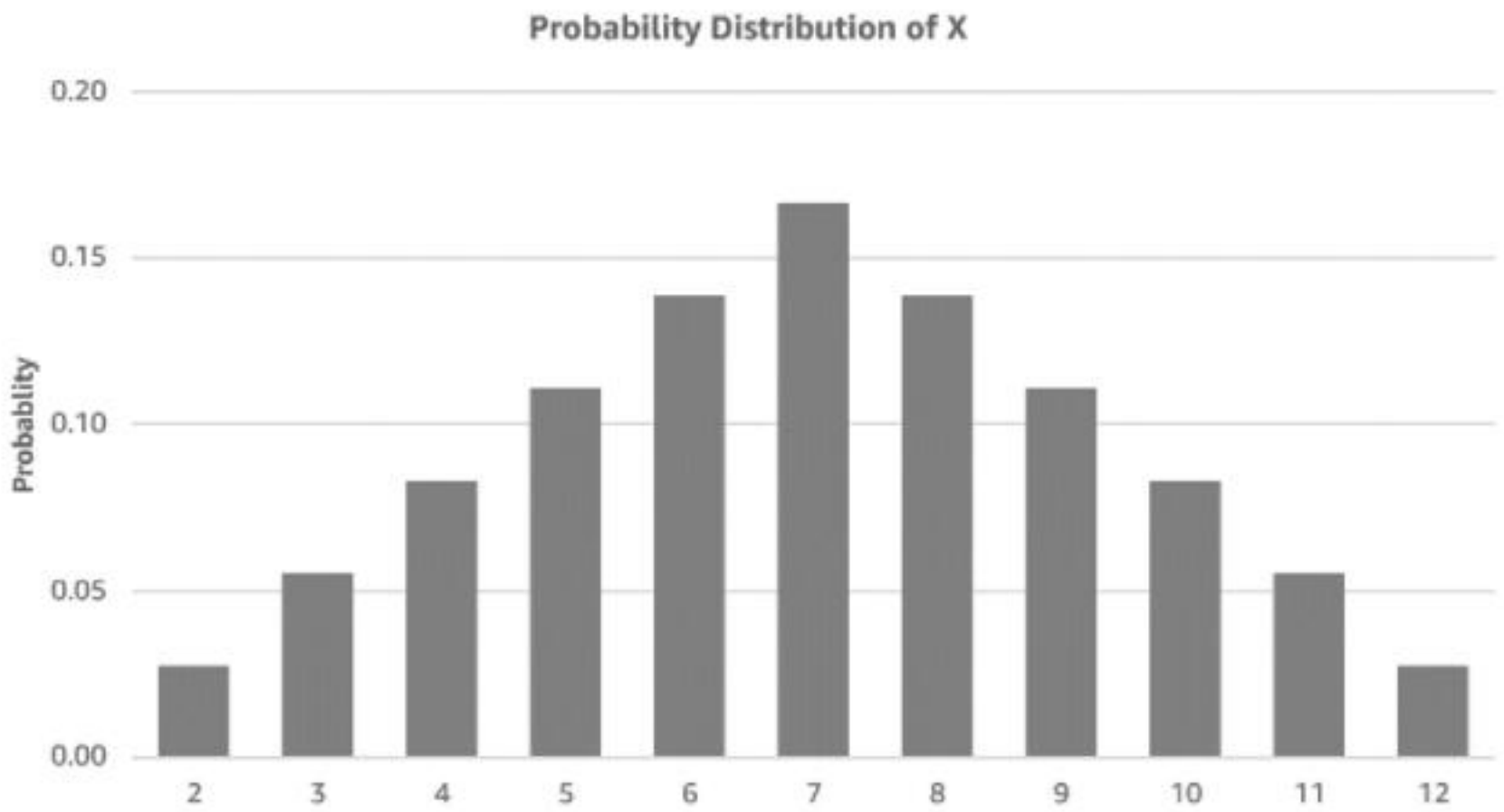
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Continuous Probability

In this video, we'll explore continuous probability and common types of these random variables.

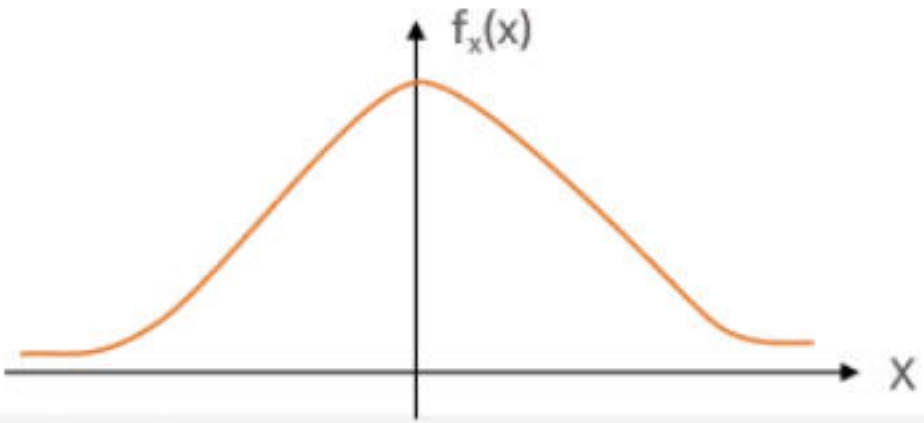
Discrete Random Variables:

Sum of two dice



Continuous Random Variable:

Probability density function



$$X \sim Gaussian\left(\mu, \sigma^2\right) \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

Where
μ - mean of probability density function
Σ – standard deviation

A Significant Issue

aws training and certification

With continuous variables, you can no longer just use probability (the probability a Gaussian is equal to 2.317635618… is 0). Therefore, you need to expand your point of view to **densities**.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definitions

aws training and certification

Probability Density Function

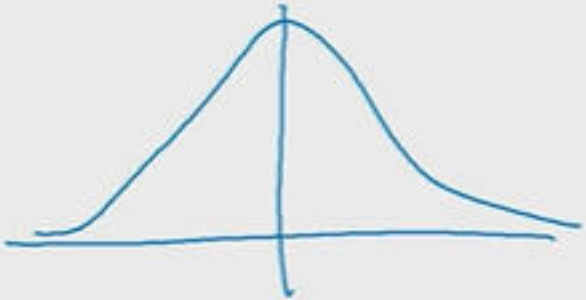
$f_X(x)$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definitions

aws training and certification

Gaussian Random Variables



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

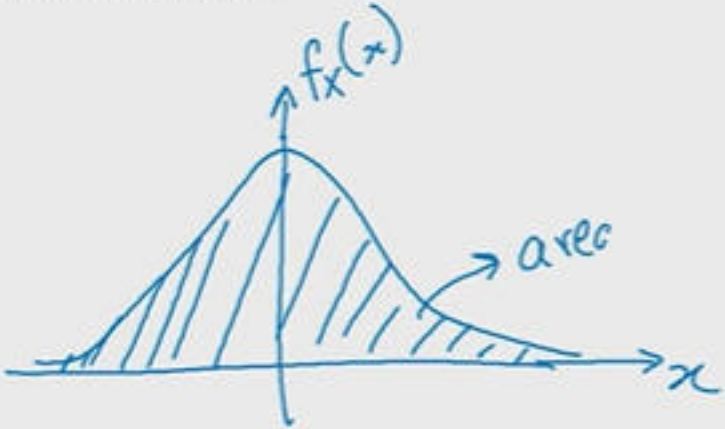
Definitions

Gaussian Random Variables

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definitions

Gaussian Random Variables

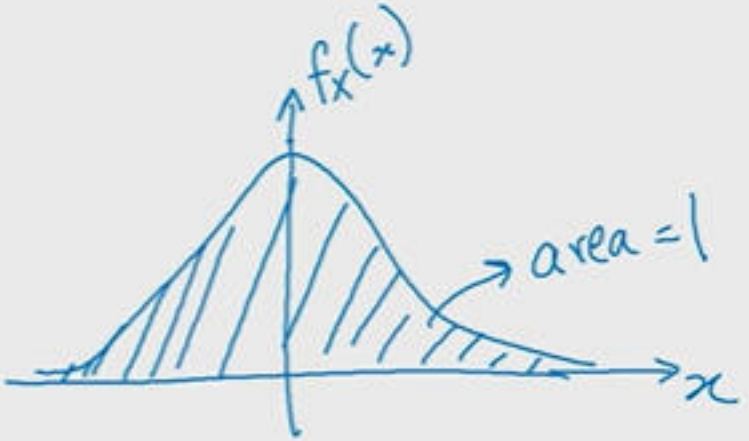


© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definitions

Gaussian Random Variables

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

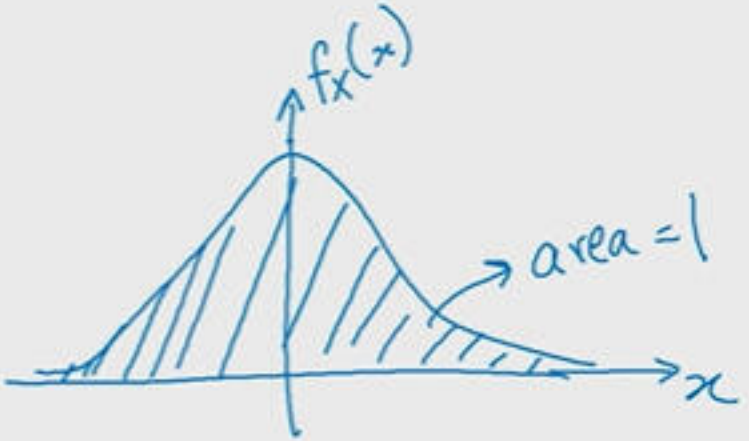


© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definitions

Gaussian Random Variables

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



μ, σ

mean

standard deviation

(σ^2 ← variance)

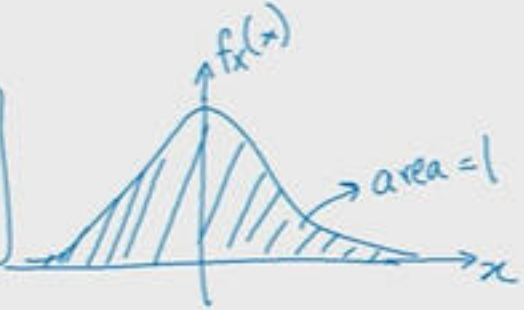
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definitions

aws training and certification

Gaussian Random Variables

$$\boxed{f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$$



μ, σ

mean

standard deviation

($\sigma^2 \leftarrow$ variance)

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Idea

We now extend MLE to **densities** rather than probabilities.

$$L(\theta) = P_\theta(\{Y_i : i=1,\dots,N\})$$

$$\updownarrow$$

cont. RV.

$$L(\theta) = f_X(\{Y_i : i=1,\dots,N\})$$

$$= \prod_{i=1} f_X(Y_i)$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Idea

We now extend MLE to **densities** rather than probabilities.

$$L(\theta) = P_\theta(\{Y_i : i=1,\dots,N\})$$

cont. RV.

$$L(\theta) = f_X(\{Y_i : i=1,\dots,N\})$$

$$= \prod_{i=1}^{N} f_X(Y_i)$$

$$\boxed{\log L(\theta) = \sum_{i=1}^{N} \log f_X(y_i)}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws training and certification

Thank You

© 2018 Amazon Web Services, Inc. or its affiliates. All rights reserved. This work may not be reproduced or redistributed, in whole or in part, without prior written permission from Amazon Web Services, Inc. Commercial copying, lending, or selling is prohibited. Corrections or feedback on the course, please email us at: aws-course-feedback@amazon.com. For all other questions, contact us at: https://aws.amazon.com/contact-us/aws-training/. All trademarks are the property of their owners.

-0:02

aws training and certification

Gaussian Maximum Likelihood Estimate (MLE)

-5:52

Example

Suppose we want to find the MLE for μ when $x \sim \mathbf{N}\ (\mu, 1)$ given observations: $x_1, x_2, x_3 \ldots \ldots x_n$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Example

Suppose we want to find the MLE for μ when $x \sim \mathbf{N}\ (\mu, 1)$ given observations: $x_1, x_2, x_3 \ldots \ldots x_n$

$$f_X(x) = \left(\frac{1}{\sqrt{2\pi}}\right)\exp\left(-\frac{(x-\mu)^2}{2}\right) \Rightarrow \log f_X(x) = -\frac{1}{2}\log(2\pi) - \frac{(x-\mu)^2}{2}$$

$$\log(L) = \sum_{i=1}^{n} \log f_X(x_i) = -\frac{n}{2}\underbrace{\log(2\pi)}_{0} - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\text{MLE}, \hat{\mu} : \frac{d(\log)}{d\mu} = 0 \Rightarrow \sum_{i=1}^{n}(x_i-\hat{\mu})(-1) = 0$$

$$\Rightarrow \boxed{\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

A Note

We see here that **maximum** likelihood estimation is equivalent to **minimizing** the sum of squared errors.

$$\frac{dL}{d\mu} = 0$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

A Note

We see here that **maximum** likelihood estimation is equivalent to **minimizing** the sum of squared errors.

$$\frac{dL}{d\mu} = 0 \Rightarrow \frac{d}{d\mu}\left(-\sum_{i=1}^{n}(x_i-\mu)^2\right) = 0$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

A Note

We see here that **maximum** likelihood estimation is equivalent to **minimizing** the sum of squared errors.

$$\frac{dL}{d\mu} = 0 \Rightarrow \frac{d}{d\mu}\left(-\sum_{i=1}^{n}(x_i-\mu)^2\right) = 0$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

A Note

We see here that **maximum** likelihood estimation is equivalent to **minimizing** the sum of squared errors.

$$\frac{dL}{d\mu} = 0 \Rightarrow \frac{d}{d\mu}\left(-\sum_{i=1}^{n}(x_i - \mu)^2\right) = 0$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

A Note

We see here that **maximum** likelihood estimation is equivalent to **minimizing** the sum of squared errors.

$$\frac{dL}{d\mu} = 0 \Rightarrow \frac{d}{d\mu}\left(-\sum_{i=1}^{n}(x_i-\mu)^2\right) = 0$$

maximize L = minimize sum of squared errors.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Topics

- OLS as MLE
- The Model

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

OLS AS MLE

We can now **rephrase** OLS as a **likelihood maximization problem**.

OLS:

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

OLS AS MLE

We can now **rephrase** OLS as a **likelihood maximization problem**.

OLS:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} = \bar{y} = X\bar{\beta}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

OLS AS MLE

We can now **rephrase** OLS as a **likelihood maximization problem**.

OLS:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} = \bar{y} = X\bar{\beta}$$

output variable. data matrix coefficients

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

OLS AS MLE

We can now **rephrase** OLS as a **likelihood maximization problem**.

OLS:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} = \hat{\bar{y}} = X\bar{\beta}$$

output variable

data matrix

coefficients

$$\bar{y} = X\bar{\beta} + \epsilon$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

OLS AS MLE

We can now **rephrase** OLS as a **likelihood maximization problem**.

OLS:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} = \hat{\bar{y}} = X\bar{\beta}$$

output variable

data matrix

coefficients

$$\bar{y} = X\bar{\beta} + \epsilon$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

OLS AS MLE

We can now **rephrase** OLS as a **likelihood maximization problem**.

OLS:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} = \underbrace{\hat{\bar{y}}}_{\text{output variable}} = \underbrace{X}_{\text{data matrix}} \underbrace{\bar{\beta}}_{\text{coefficients}}$$

$$\bar{y} = X\bar{\beta} + \underline{\epsilon}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Model

$$\epsilon \sim N(0, \sigma^2)$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Likelihood of Given Data

μ

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Likelihood of Given Data

$y_i \sim N(\bar{x}_i \cdot \bar{\beta}, \sigma^2) \longleftrightarrow \bar{y} \sim$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Likelihood of Given Data

$$y_i \sim N(\bar{x}_i \cdot \bar{\beta}, \sigma^2) \longleftrightarrow \bar{y} \sim N(X\bar{\beta}, \Sigma^2)$$

$$f_Y(y_i, \underline{x}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp($$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Likelihood of Given Data

$$y_i \sim N(\bar{x}_i \cdot \bar{\beta}, \sigma^2) \longleftrightarrow \bar{y} \sim N(X\bar{\beta}, \Sigma^2)$$

$$f_Y(y_i, x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \bar{x}_i \cdot \bar{\beta})^2}{2\sigma^2}\right)$$

$$\log f_Y(y_i, x_i) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_i - \bar{x}_i \cdot \bar{\beta})^2}{2\sigma^2}$$

$$L(\beta, \sigma^2) = \sum_{i=1}^{N}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Likelihood of Given Data

$$y_i \sim N(\bar{x}_i \cdot \bar{\beta}, \sigma^2) \longleftrightarrow \bar{y} \sim N(X\bar{\beta}, \Sigma^2)$$

$$f_Y(y_i, x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \bar{x}_i \cdot \bar{\beta})^2}{2\sigma^2}\right)$$

$$\log f_Y(y_i, x_i) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_i - \bar{x}_i \cdot \bar{\beta})^2}{2\sigma^2}$$

$$L(\beta, \sigma^2) = \sum_{i=1}^{N} \left(\log f_Y(y_i | x_i)\right)$$

$$= -\frac{N}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{N} \frac{(y_i - \bar{x}_i \cdot \bar{\beta})^2}{2\sigma^2}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Likelihood in Matrix Form

$$\log L(\beta, \sigma^2) = -\frac{1}{2\sigma^2} \|\bar{y} - X\bar{\beta}\|_2^2 - \frac{1}{2} \log(2\pi\sigma^2)$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Relation to the Least Squares Approach

$$\log L(\beta, \sigma^2) = -\frac{N}{2}\log(\cancel{2\pi}\sigma^2) - \frac{1}{2}\underbrace{(\bar{y} - X\bar{\beta})^T(\bar{y} - X\bar{\beta})}$$

$$\frac{d}{d\beta}\log L = 0$$

$$\Rightarrow \frac{d}{d\beta}\left(-\frac{1}{2}\underbrace{(\bar{y} - X\bar{\beta})^T(\bar{y} - X\bar{\beta})}_{\text{MSE}}\right) = 0$$

MLE

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Solution

Suppose we model our problem as:

$$y \sim \vec{x}\vec{\beta} + \mathcal{N}(0, \sigma^2).$$

Then, the maximum likelihood estimate of $\vec{\beta}$ is

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

same as MSE approach.

the same as the OLS estimator

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

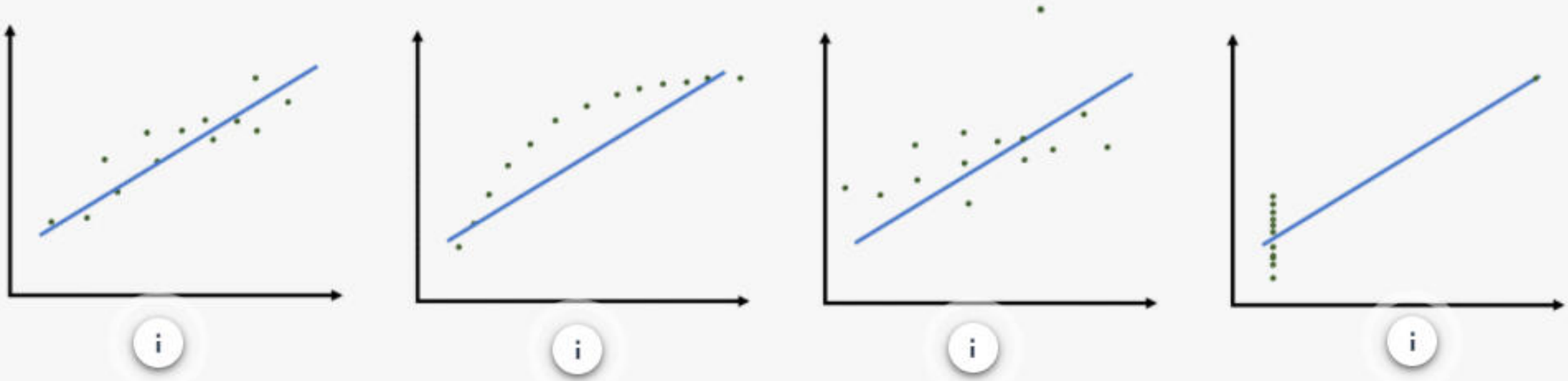
Improvement #1

Event 1 Title

This lets you get the MLE for standard deviation (\sigma).

$$\sigma^2 = \frac{2\pi}{N} \left\| \bar{Y} - X\bar{\beta} \right\|_2^2$$

Improvement in Anscombe's Quartet



First Upshot

We can get a MLE for σ.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

First Upshot

We can get a MLE for σ.

$$\log \mathcal{L} = -\frac{1}{2\sigma^2} \|\bar{y} - X\bar{\beta}\|_2^2 - \frac{N}{2} \log(2\pi\sigma^2)$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

First Upshot

We can get a MLE for σ.

$$\log \mathcal{L} = -\frac{1}{2\sigma^2} \|\bar{y} - X\bar{\beta}\|_2^2 - \frac{N}{2} \log(2\pi\sigma^2)$$

$$\frac{\partial \log \mathcal{L}}{\partial \sigma} = \frac{\|\bar{y} - X\bar{\beta}\|_2^2 (2)}{2\sigma^3} - \frac{N}{2} \frac{1}{2\pi\sigma^2} (2\pi\sigma)(2)$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

First Upshot

We can get a MLE for σ.

$$\log \mathcal{L} = -\frac{1}{2\sigma^2}\|\bar{y} - X\bar{\beta}\|_2^2 - \frac{N}{2}\log(2\pi\sigma^2)$$

$$\frac{\partial \log \mathcal{L}}{\partial \sigma} = \frac{\|\bar{y} - X\bar{\beta}\|_2^2 (\cancel{2})}{\cancel{2}\sigma^3} - \frac{N}{\cancel{2}}\frac{1}{\cancel{2}\cancel{\pi}\sigma^{\cancel{2}}}(\cancel{2}\cancel{\pi}\cancel{\sigma})(\cancel{2}) =$$

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

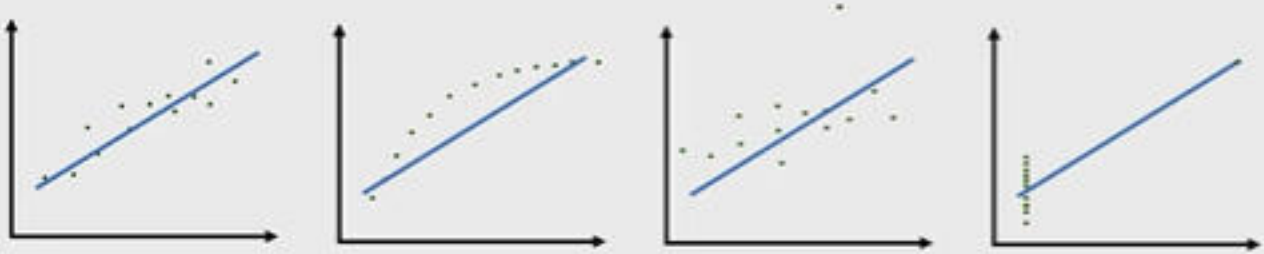
Summary

Thus, we may interpret the MSE loss function as a MLE for the variance of the Gaussian noise.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Second Upshot

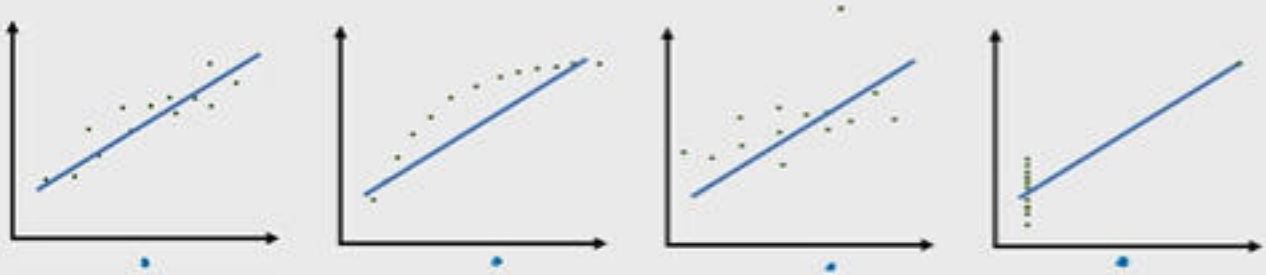
We may gain a deeper understanding of Anscombe's Quartet.



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Second Upshot

We may gain a deeper understanding of Anscombe's Quartet.

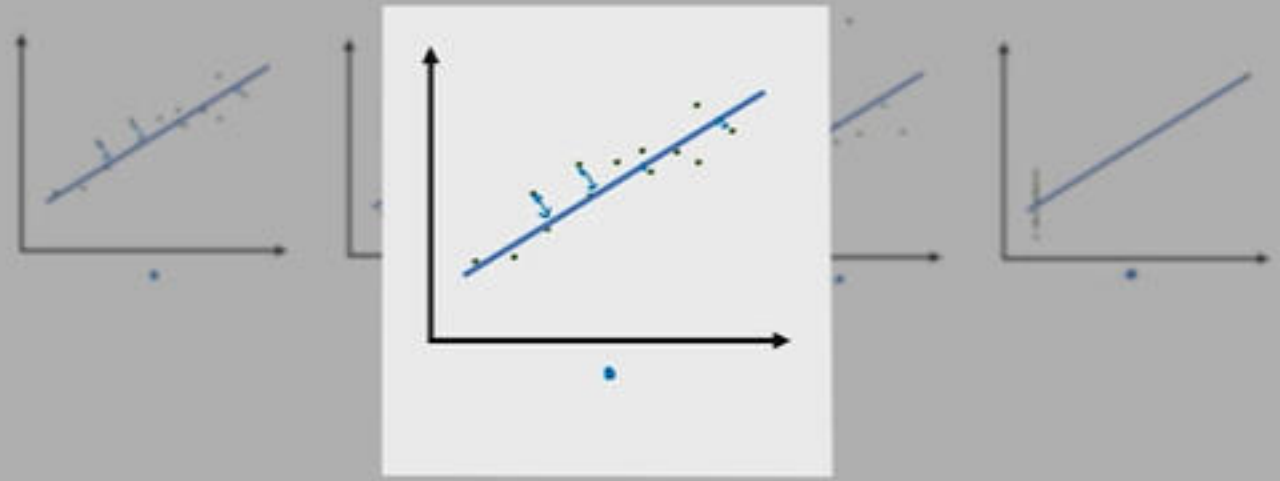


© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Second Upshot

aws training and certification

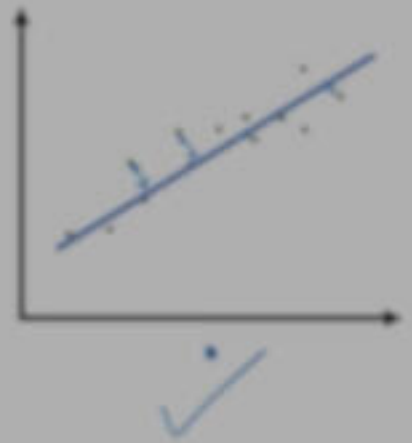
We may gain a deeper understanding of Anscombe's Quartet.

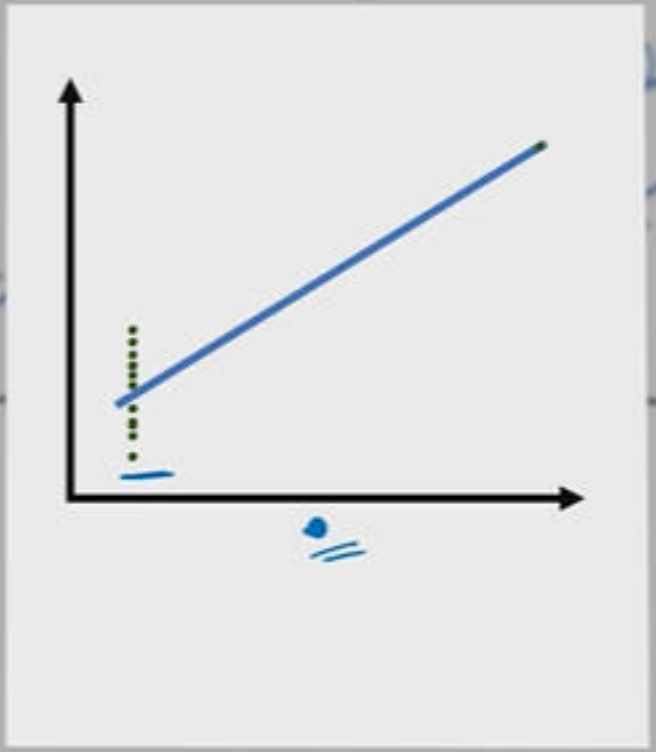


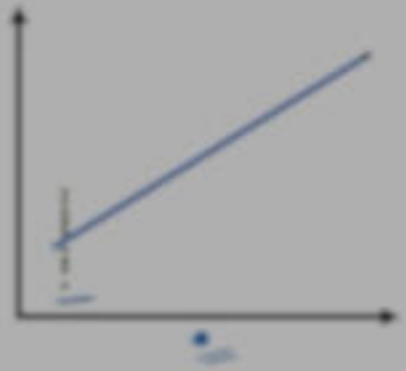
-4:25

Second Upshot

We may gain a deeper understanding of Anscombe's Quartet.







© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Third Upshot

aws training and certification

Prediction Improvement

① Error model is not correct → Different error model.

② Error not indep. w/ x → Tva

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Assumptions of OLS, Nonlinear OLS, Regularization, and Cross-Validation

Introduction

This video contains lessons as they relate to the assumptions of linear regression, overfitting an
regularization.

The following topics will be covered:

- Assumptions of linear regression
- Concept of residuals
- Extensions of linearity
- Extreme learning machines
- Overfitting
- Regularizations
- Other issues

Extensions of Linear Regression

- Extensions of Linear Regression

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Beyond Linearity

- Definition
  - Residuals
- Assumptions on Residuals
- A Good Residual Plot
- A Bad Residual Plot
- OLS Actually Applies Beyond Linear Models
- Basis Functions

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Regularization

- The Bayesian Point of View
- Ridge Regression
- LASSO Regression
- The Geometry of Regularization
- The Geometry of Ridge Regularization
- The Geometry of LASSO Regularization
- Elastic Net Regression

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Assumptions of Linear Regression

In the linear regression model, various assumptions were made. This video covers how the model performs outside of those assumptions. The assumptions made in a linear regression model are:

Model

Click to flip

Errors

X^TX

Assumptions of Linear Regression

aws training and certification

1. Linearity

$$\bar{y} = X\bar{\beta} + N(0, \Sigma)$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Assumptions of Linear Regression

aws training and certification

2. The errors are independent normals

$$\underbrace{N}_{\text{normal}}\left(\underbrace{0}_{\cdot 2}, \Sigma^2\right)$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Assumptions of Linear Regression

aws training and certification

2. The errors are independent normals

$$\underbrace{N}_{\text{normal}}\left(\underbrace{0}_{\text{zero mean}}, \underbrace{\Sigma^2}_{\text{covariance matrix}}\right) \text{ (diagonal)} \quad \begin{bmatrix} \sigma^2 & & 0 \\ & \sigma^2 & \\ 0 & & \ddots & \\ & & & \sigma^2 \end{bmatrix}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Assumptions of Linear Regression

aws training and certification

3. $\boldsymbol{X^T X}$ is invertible

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

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Pushing Beyond Assumptions

We will continue to see how to **push beyond each of these assumptions**, one at a time, and see how **ordinary least squares** is actually far more **flexible** than usually assumed.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Images

- https://thenounproject.com/search/?q=push&i=11715

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws training and certification

Thank You

© 2018 Amazon Web Services, Inc. or its affiliates. All rights reserved. This work may not be reproduced or redistributed, in whole or in part, without prior written permission from Amazon Web Services, Inc. Commercial copying, lending, or selling is prohibited. Corrections or feedback on the course, please email us at: aws-course-feedback@amazon.com. For all other questions, contact us at: https://aws.amazon.com/contact-us/aws-training/. All trademarks are the property of their owners.

Beyond Linearity

In this video, we discuss residuals and their plots, OLS models when the relationship between data matrix (X) and output function (Y) is nonlinear, and different basis functions that are used to expand features. Take a moment to explore residuals.

A residual plot:

Plot of ϵ against x.

Where

$$\epsilon_i = y_i(x) - \hat{y}_i$$

Assumptions:

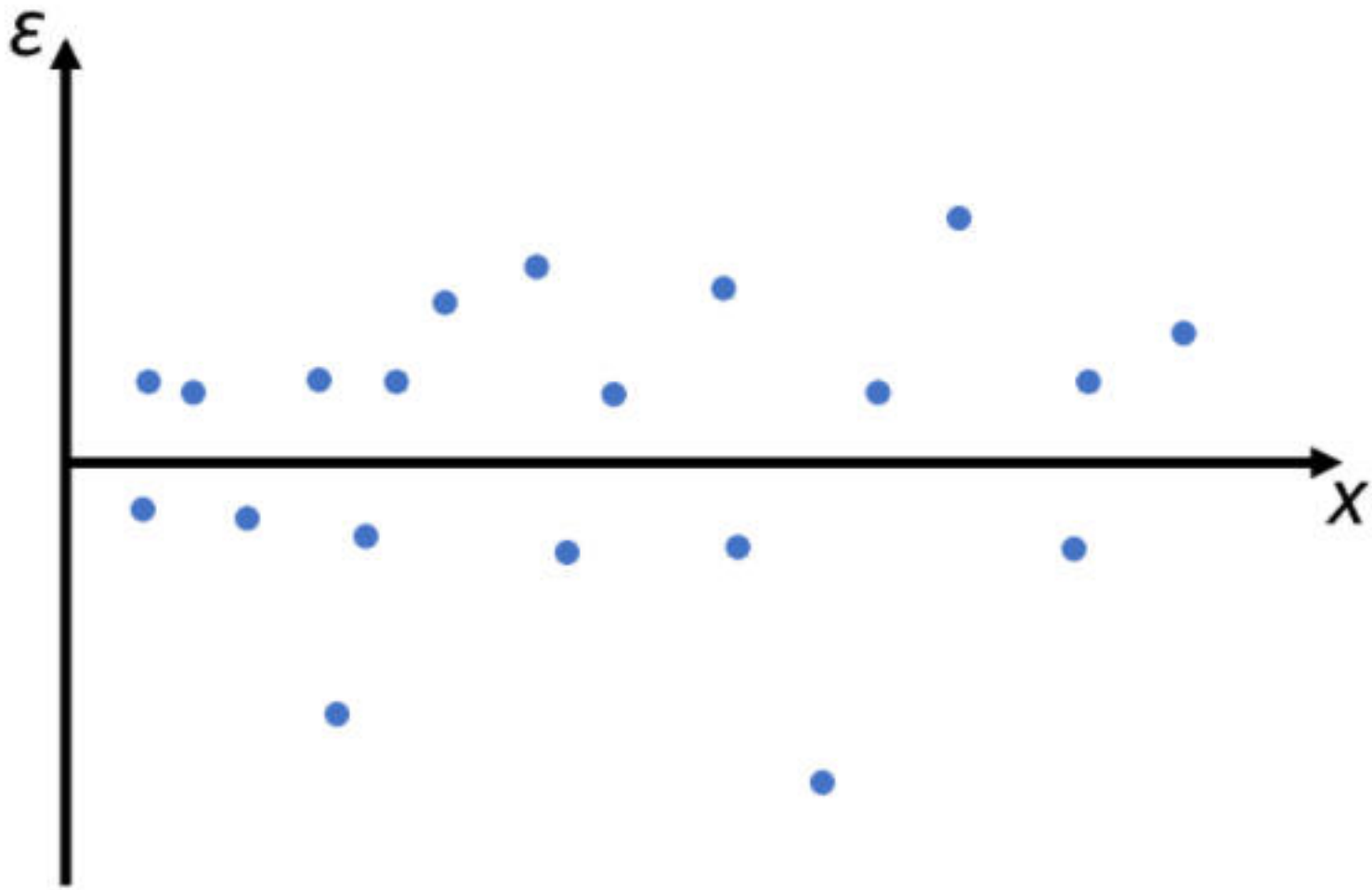
Residuals are independent of x, normally distributed, and uncorrelated.

A good residual plot

Every creative endeavor requires that you take risks. If you try and don't succeed, you've still learned something. You're not failing. You're discovering what doesn't work.

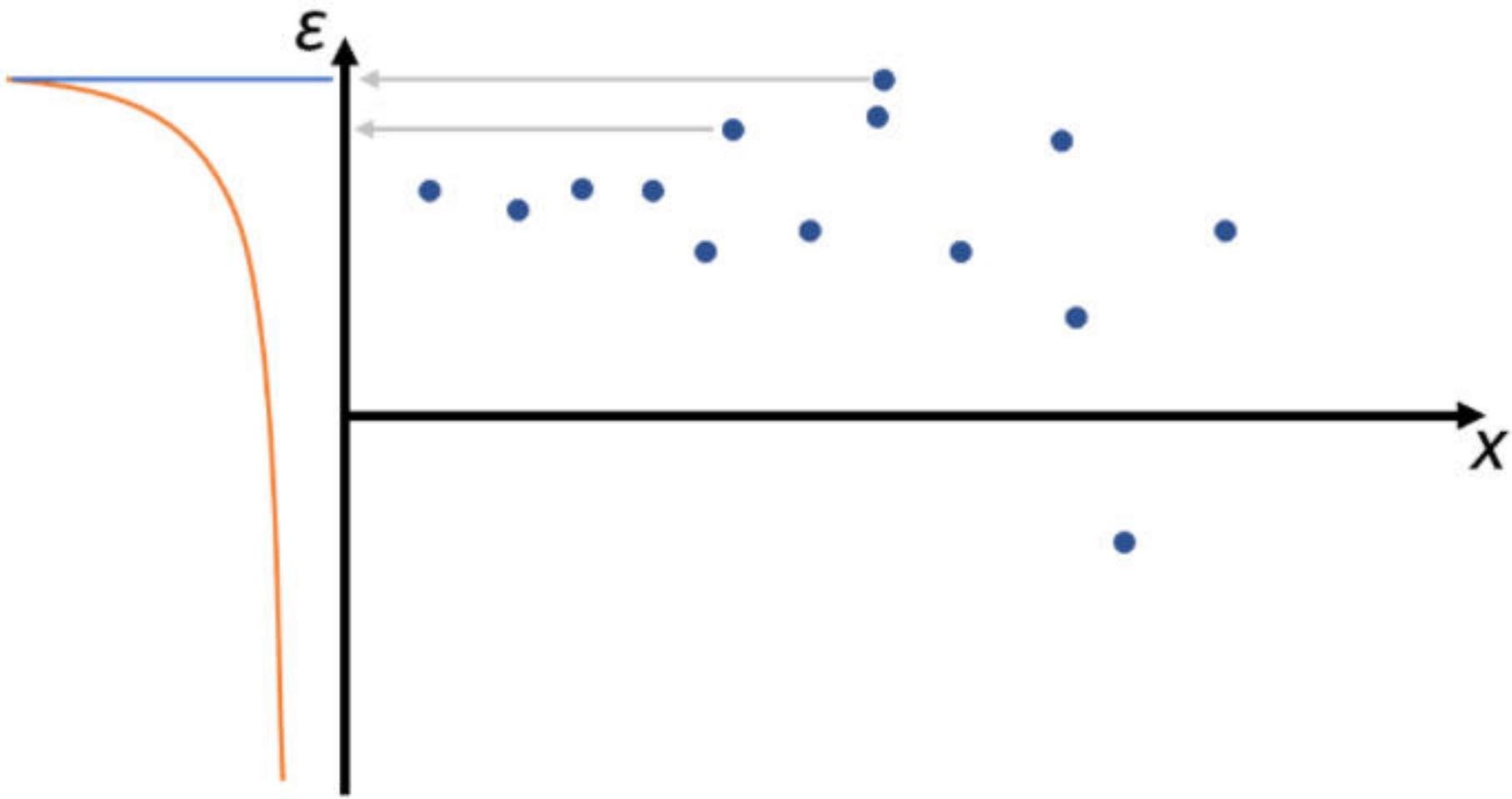
A good residual plot

Every creative endeavor requires that you take risks. If you try and don't succeed, you've still learned something. You're not failing. You're discovering what doesn't work.



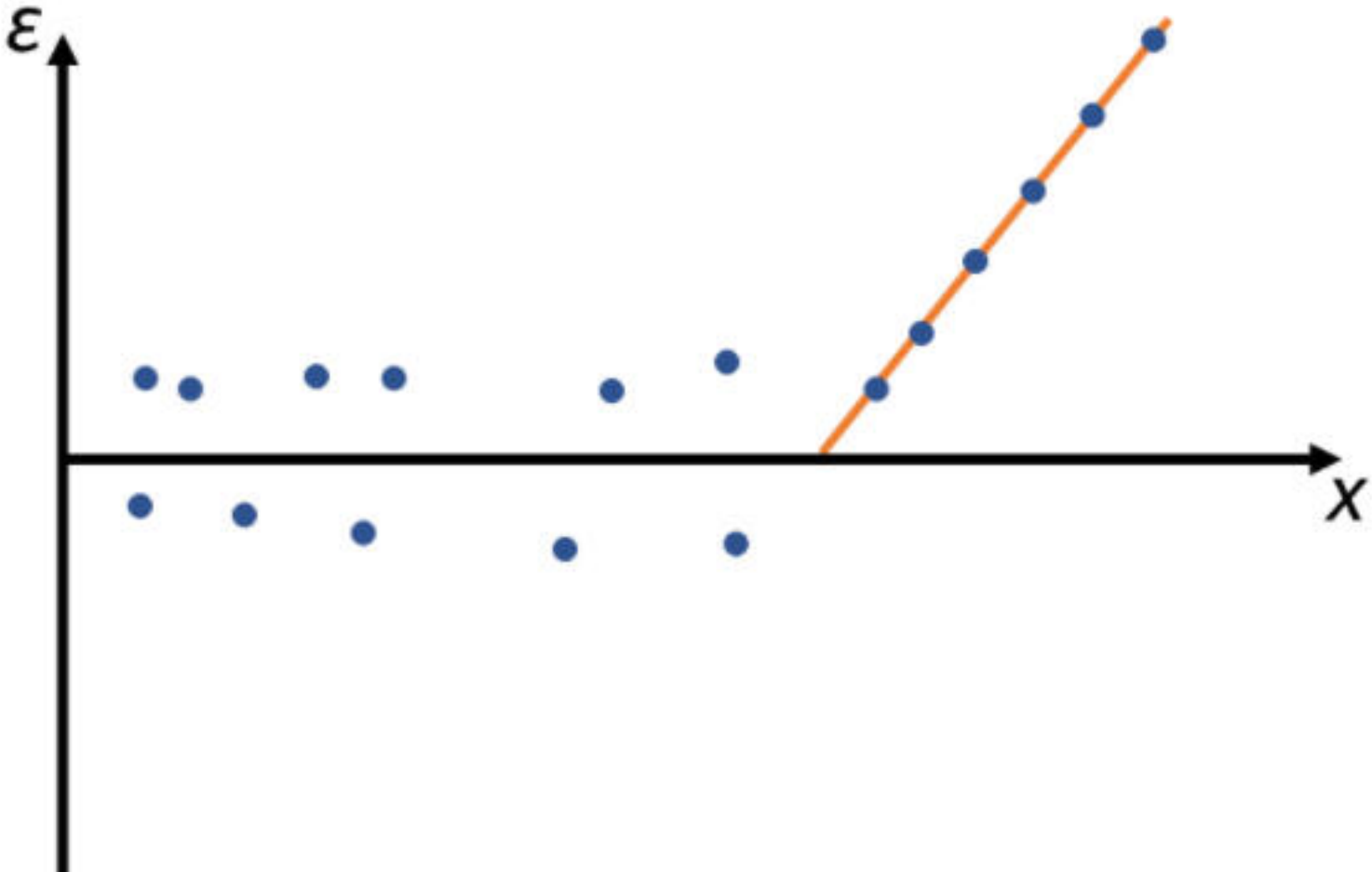
A bad residual plot

The errors are not normally distributed.



A misleading residual plot

Up to certain point the model seems to be a good fit, but after that the errors are dependent of variable x.



Topics

- Definition
  - Residuals
- Assumptions on Residuals
- A Good Residual Plot
- A Bad Residual Plot
- OLS Actually Applies Beyond Linear Models

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definition

Residuals

$$\epsilon_i = y_i - \hat{y}_i$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definition

aws training and certification

Residuals

$$\epsilon_i = y_i - \hat{y}_i$$

Residual plot

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Assumptions on Residuals

Assume the residuals are independent normals uncorrelated with x. If you plot (x_i, ϵ_i) pairs, then you can attempt to diagnose if this is a reasonable assumption.

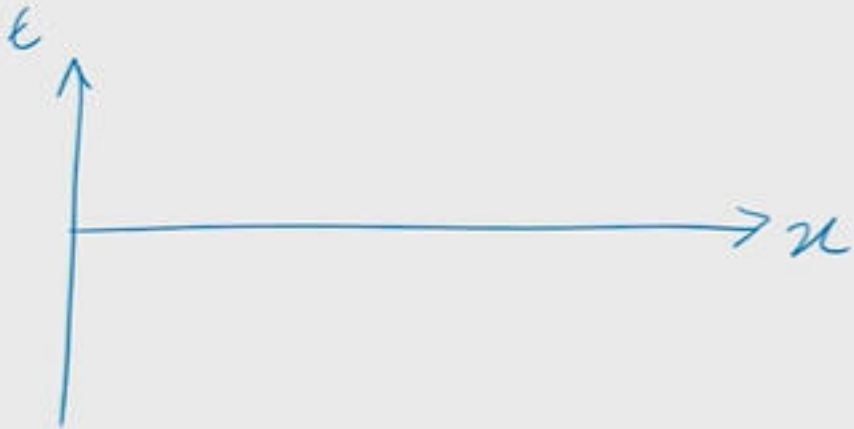
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Assumptions on Residuals

Assume the residuals are independent, normals, uncorrelated with $\boldsymbol{x}$. If you plot (x_i, ϵ_i) pairs, then you can attempt to diagnose if this is a reasonable assumption.

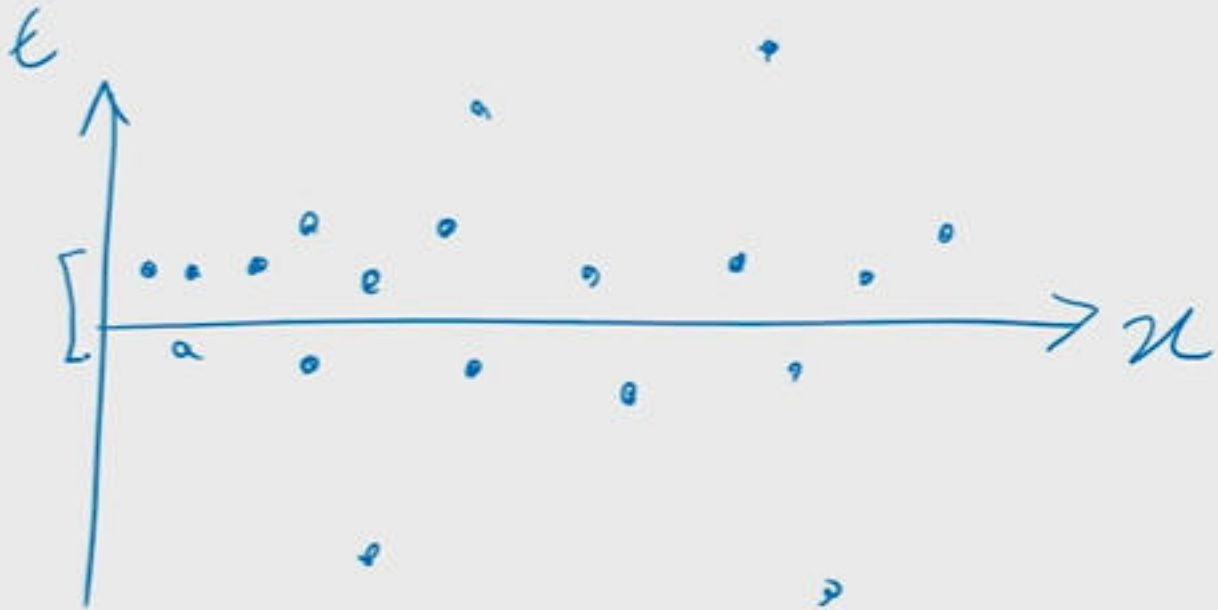
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

A Good Residual Plot



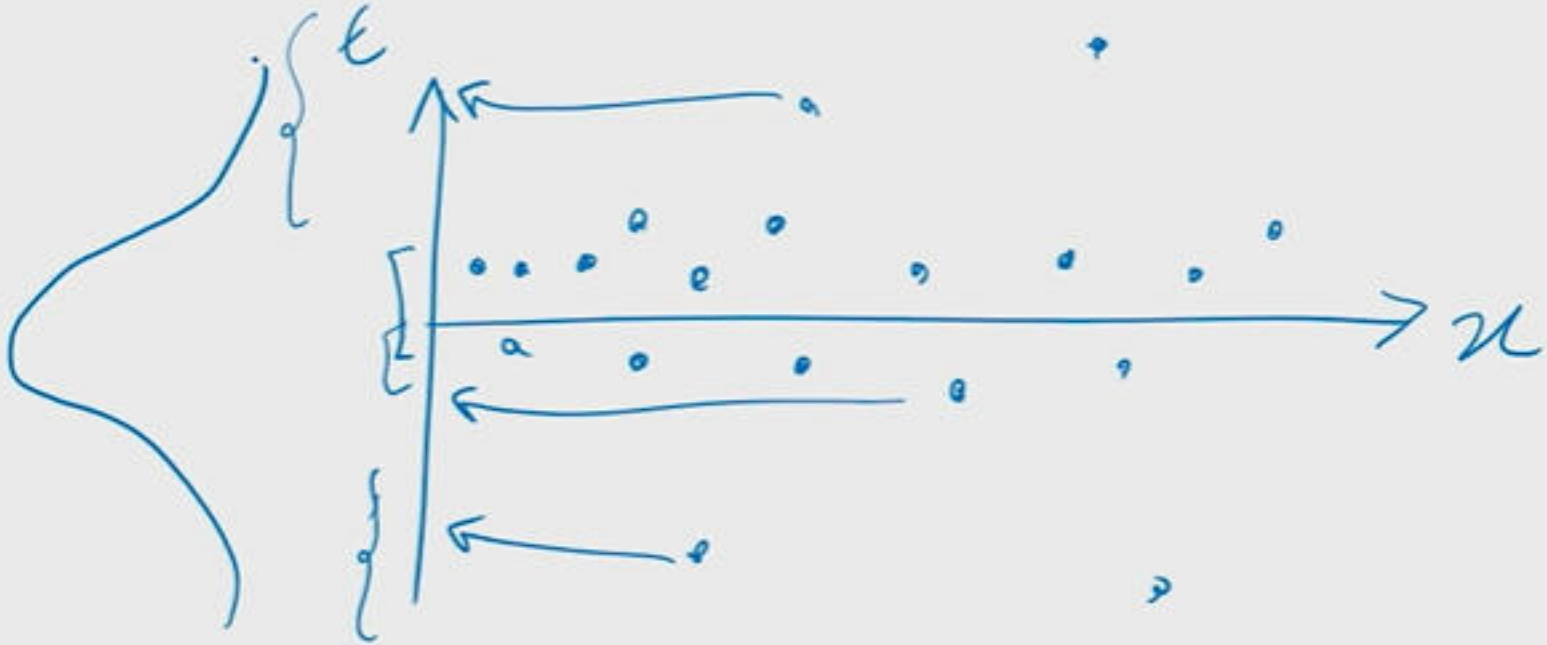
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

A Good Residual Plot



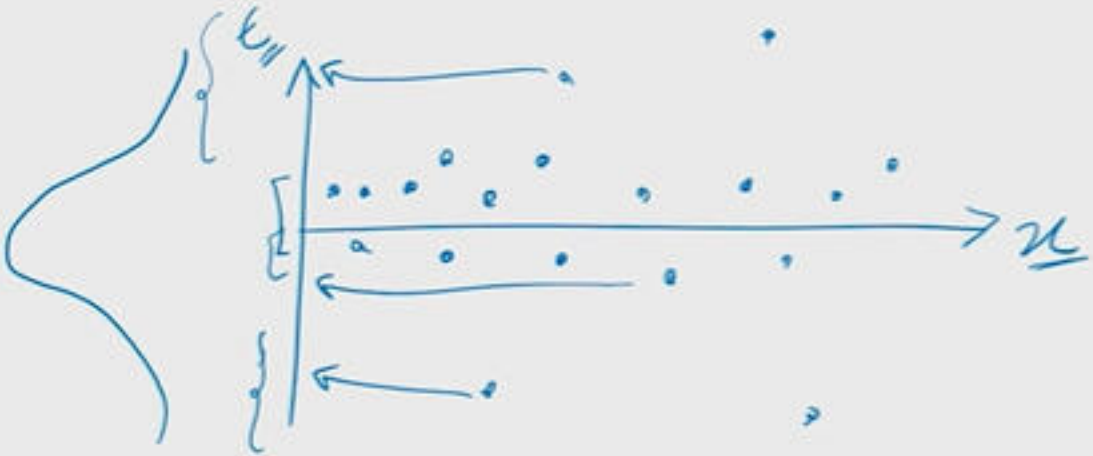
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

A Good Residual Plot



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

A Good Residual Plot



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

A Bad Residual Plot

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

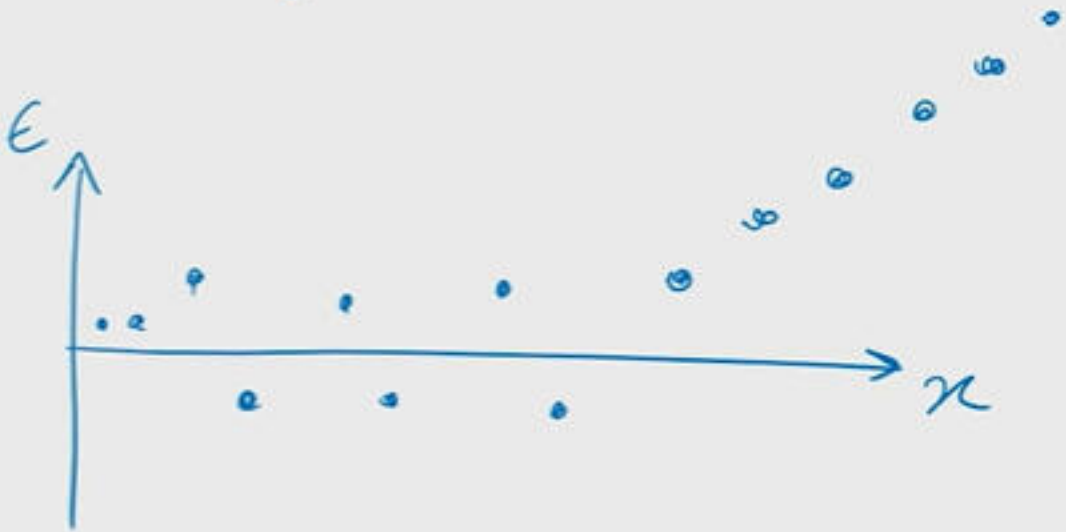
A Bad Residual Plot

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

A Bad Residual Plot

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

A Bad Residual Plot



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Ordinary Least Squares (OLS) Applies Beyond Linear Models

You can augment your data matrix to use OLS to fit non-linear models:

$$y \sim \beta_0 \cdot 1 + \beta_1 x + \beta_2 x^2 + N(0,1)$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Ordinary Least Squares (OLS) Applies Beyond Linear Models

aws training and certification

You can augment your data matrix to use OLS to fit non-linear models:

$$y \sim \beta_0 \cdot 1 + \beta_1 x + \beta_2 x^2 + N(0,1)$$

	x	y
1	1	1·1
2	2	4·1
3	3	e
4	4	
⋮		

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Ordinary Least Squares (OLS) Applies Beyond Linear Models

aws training and certification

You can augment your data matrix to use OLS to fit non-linear models:

$$y \sim \beta_0 \cdot 1 + \beta_1 x + \beta_2 x^2 + N(0,1)$$

	x	y
1	1	1.1
2	2	4.1
3	3	9.2
4	4	15.9
⋮		

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Ordinary Least Squares (OLS) Applies Beyond Linear Models

aws training and certification

You can augment your data matrix to use OLS to fit non-linear models:

$$y \sim \beta_0 \cdot 1 + \beta_1 x + \beta_2 x^2 + N(0,1)$$

	x	y
1	1	1.1
2	2	4.1
3	3	9.2
4	4	15.9
⋮		

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Ordinary Least Squares (OLS) Applies Beyond Linear Models

aws training and certification

You can augment your data matrix to use OLS to fit non-linear models:

$$y \sim \beta_0 \cdot 1 + \beta_1 x + \beta_2 x^2 + N(0,1)$$

	x	y
1	1	1.1
2	2	4.1
3	3	9.2
4	4	15.9
⋮		

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Ordinary Least Squares (OLS) Applies Beyond Linear Models

aws training and certification

You can augment your data matrix to use OLS to fit non-linear models:

$$y \sim \beta_0 \cdot 1 + \beta_1 x + \beta_2 x^2 + N(0,1)$$

	x	y
1	1	1.1
2	2	4.1
3	3	9.2
4	4	15.9
⋮		

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Ordinary Least Squares (OLS) Applies Beyond Linear Models

aws training and certification

You can augment your data matrix to use OLS to fit non-linear models:

$$y \sim \beta_0 \cdot 1 + \beta_1 x + \beta_2 x^2 + N(0,1)$$

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{bmatrix}$$

	x	y
0	1	1.1
1	2	4.1
2	3	9.2
3	4	15.9

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Ordinary Least Squares (OLS) Applies Beyond Linear Models

aws training and certification

You can augment your data matrix to use OLS to fit non-linear models:

$$y \sim \beta_0 \cdot 1 + \beta_1 x + \beta_2 x^2 + N(0,1)$$

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_N \end{bmatrix} \rightarrow \tilde{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 \end{bmatrix}$$

	x	y
①	1	1.1
②	2	4.1
③	3	9.2
④	4	15.9

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Basis Functions

You may use a basis function φ to expand our features with non-linear functions of the measured data in many ways. Here are a few common ones:

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Basis Functions

aws training and certification

You may use a basis function φ to expand our features with non-linear functions of the measured data in many ways. Here are a few common ones:

① polynomial $= (1, x, x^2, x^3, \dots, x^k)$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Basis Functions

aws training and certification

You may use a basis function φ to expand our features with non-linear functions of the measured data in many ways. Here are a few common ones:

① polynomial $= (1, x, x^2, x^3, \dots, x^k)$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Basis Functions

aws training and certification

You may use a basis function φ to expand our features with non-linear functions of the measured data in many ways. Here are a few common ones:

① polynomial $= (1, x, x^2, x^3, \dots, x^k)$

② periodic $=$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Basis Functions

aws training and certification

You may use a basis function φ to expand our features with non-linear functions of the measured data in many ways. Here are a few common ones:

① polynomial $= (1, x, x^2, x^3, \dots, x^k)$

② periodic $= (1, \sin(x), \cos(x), \dots$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Basis Functions

aws training and certification

You may use a basis function φ to expand our features with non-linear functions of the measured data in many ways. Here are a few common ones:

① polynomial $= (1, x, x^2, x^3, \dots, x^k)$

② periodic $= (1, \sin(x), \cos(x), \sin(2x),$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Basis Functions

aws training and certification

You may use a basis function φ to expand our features with non-linear functions of the measured data in many ways. Here are a few common ones:

① polynomial $= (1, x, x^2, x^3, \dots, x^k)$

② periodic $= (1, \sin(x), \cos(x), \sin(2x), \cos(2x), \dots$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Basis Functions

aws training and certification

You may use a basis function φ to expand our features with non-linear functions of the measured data in many ways. Here are a few common ones:

① polynomial $= (1, x, x^2, x^3, \dots, x^k)$

② periodic $= (1, \sin(x), \cos(x), \sin(2x), \cos(2x), \dots)$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Basis Functions

aws training and certification

You may use a basis function φ to expand our features with non-linear functions of the measured data in many ways. Here are a few common ones:

① polynomial $= (1, x, x^2, x^3, \dots, x^k)$

② periodic $= (1, \sin(x), \cos(x), \sin(2x), \cos(2x), \dots, \sin(kx), \cos(k$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Basis Functions

aws training and certification

You may use a basis function φ to expand our features with non-linear functions of the measured data in many ways. Here are a few common ones:

① polynomial $= (1, x, x^2, x^3, \dots, x^k)$

② periodic $= (1, \sin(x), \cos(x), \sin(2x), \cos(2x), \dots, \sin(kx), \cos(kx))$

③ ⊮

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Basis Functions

aws training and certification

You may use a basis function φ to expand our features with non-linear functions of the measured data in many ways. Here are a few common ones:

① polynomial $= (1, x, x^2, x^3, \ldots, x^k)$

② periodic $= (1, \sin(x), \cos(x), \sin(2x), \cos(2x), \ldots, \sin(kx), \cos(kx))$

③ kernel method $= (1, e^{-\|x - x_1\|^2},$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Extreme Learning Machines

A particularly extreme version (circa 2006 or so) is to generate a large number of **random** non-linear features and use those.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Extreme Learning Machines

A particularly extreme version (circa 2006 or so) is to generate a large number of **random** non-linear features and use those.

basis functions = (1, x,

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Extreme Learning Machines

A particularly extreme version (circa 2006 or so) is to generate a large number of **random** non-linear features and use those.

$$\text{basis functions} = \left(1, x, \tan^{-1}(x \cdot w_1), \tan^{-1}(x \cdot w_2), \ldots, \tan^{-1}(\vec{x} \cdot \vec{w}_k)\right)$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Extreme Learning Machines

A particularly extreme version (circa 2006 or so) is to generate a large number of **random** non-linear features and use those.

$$\text{basis functions} = \left(1, \bar{x}, \tan^{-1}(\bar{x}\cdot\bar{\omega}_1), \tan^{-1}(\bar{x}\cdot\bar{\omega}_2), \ldots, \tan^{-1}(\bar{x}\cdot\bar{\omega}_k)\right)$$

rando

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Extreme Learning Machines

A particularly extreme version (circa 2006 or so) is to generate a large number of **random** non-linear features and use those.

$$\text{basis functions} = \left(1, \bar{x}, \tan^{-1}(\bar{x}\cdot\bar{\omega}_1), \tan^{-1}(\bar{x}\cdot\bar{\omega}_2), \ldots, \tan^{-1}(\bar{x}\cdot\bar{\omega}_k)\right)$$

randomly generated vectors $\sim N(0, \xi)$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Extreme Learning Machines

A particularly extreme version (circa 2006 or so) is to generate a large number of **random** non-linear features and use those.

$$\text{basis functions} = \left(1, \bar{x}, \tan^{-1}(\bar{x}\cdot\bar{\omega}_1), \tan^{-1}(\bar{x}\cdot\bar{\omega}_2), \ldots, \tan^{-1}(\bar{x}\cdot\bar{\omega}_k)\right)$$

randomly generated vectors $\sim N(0, \xi)$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Performance

For a rough comparison, recall the MNIST digit recognition problem.

- Logistic regression gives about 92% accuracy.
- Most advanced neural networks give about 99.8% after days of training.

*Data Source: MNIST Database of Handwritten Digits

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Performance

For a rough comparison, recall the MNIST digit recognition problem.

- Logistic regression gives about 92% accuracy.
- Most advanced neural networks give about 99.8% after days of training.
- Extreme learning techniques give about 99.14% after a couple of **minutes**.

*Data Source: MNIST Database of Handwritten Digits

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Best Value For Your Time

aws training and certification

Opinion: Extreme learning machines are one of the best value for your time techniques in machine learning.

*Extreme learning machines: theory and applications (Huang, Shu, Siew) (2006)

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Best Value For Your Time

aws training and certification

Opinion: Extreme learning machines are one of the best value for your time techniques in machine learning.

*Extreme learning machines: theory and applications (Huang, Shu, Siew) (2006)

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Best Value For Your Time

aws training and certification

Opinion: Extreme learning machines are one of the best value for your time techniques in machine learning.

It is no harder than linear regression, but can often get within a handful of percent of the absolute best result.

Many people agree: the paper introducing it in 2006* has been cited nearly 5000 times.

*Extreme learning machines: theory and applications (Huang, Shu, Siew) (2006)

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

References

- Huang, Guang-Bin; Zhu, Qin-Yu; Siew, Chee-Kheong (2006). "Extreme learning machine: theory and applications". Neurocomputing. 70 (1): 489–501. CiteSeerX 10.1.1.217.3692 Freely accessible. doi:10.1016/j.neucom.2005.12.126

-0:11

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Dataset References

- Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition." Proceedings of the IEEE, 86(11):2278-2324, November 1998. Dataset available from http://yann.lecun.com/exdb/mnist/

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

References

- Huang, Guang-Bin; Zhu, Qin-Yu; Siew, Chee-Kheong (2006). "Extreme learning machine: theory and applications". Neurocomputing. 70 (1): 489–501. CiteSeerX 10.1.1.217.3692 Freely accessible. doi:10.1016/j.neucom.2005.12.126

-0:09

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws training and certification

Thank You

© 2018 Amazon Web Services, Inc. or its affiliates. All rights reserved. This work may not be reproduced or redistributed, in whole or in part, without prior written permission from Amazon Web Services, Inc. Commercial copying, lending, or selling is prohibited. Corrections or feedback on the course, please email us at: aws-course-feedback@amazon.com. For all other questions, contact us at: https://aws.amazon.com/contact-us/aws-training/. All trademarks are the property of their owners.

aws training and certification

Thank You

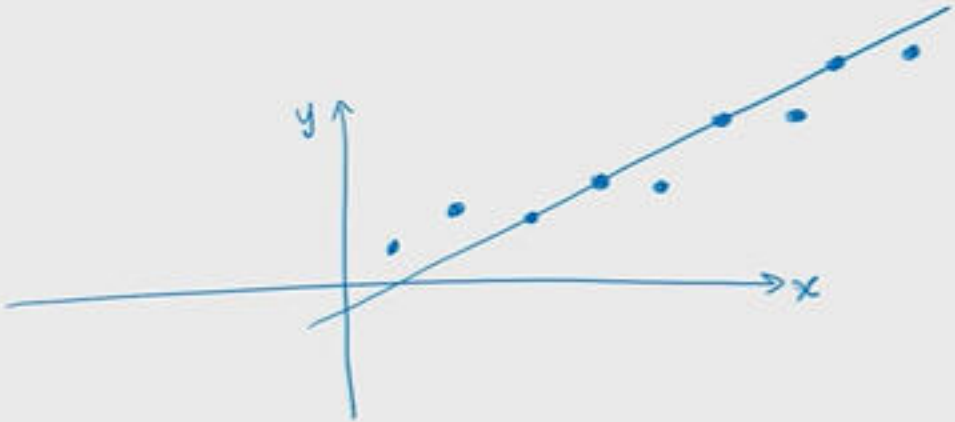
© 2018 Amazon Web Services, Inc. or its affiliates. All rights reserved. This work may not be reproduced or redistributed, in whole or in part, without prior written permission from Amazon Web Services, Inc. Commercial copying, lending, or selling is prohibited. Corrections or feedback on the course, please email us at: aws-course-feedback@amazon.com. For all other questions, contact us at: https://aws.amazon.com/contact-us/aws-training/. All trademarks are the property of their owners.

Polynomial Regression of Different Degrees

aws training and certification

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Polynomial Regression of Different Degrees

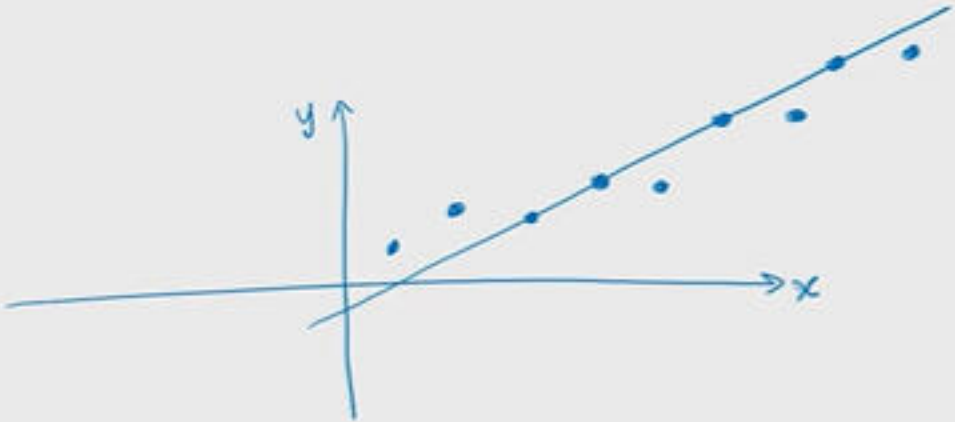


$y = ax + b$

$y = \beta_0 +$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Polynomial Regression of Different Degrees

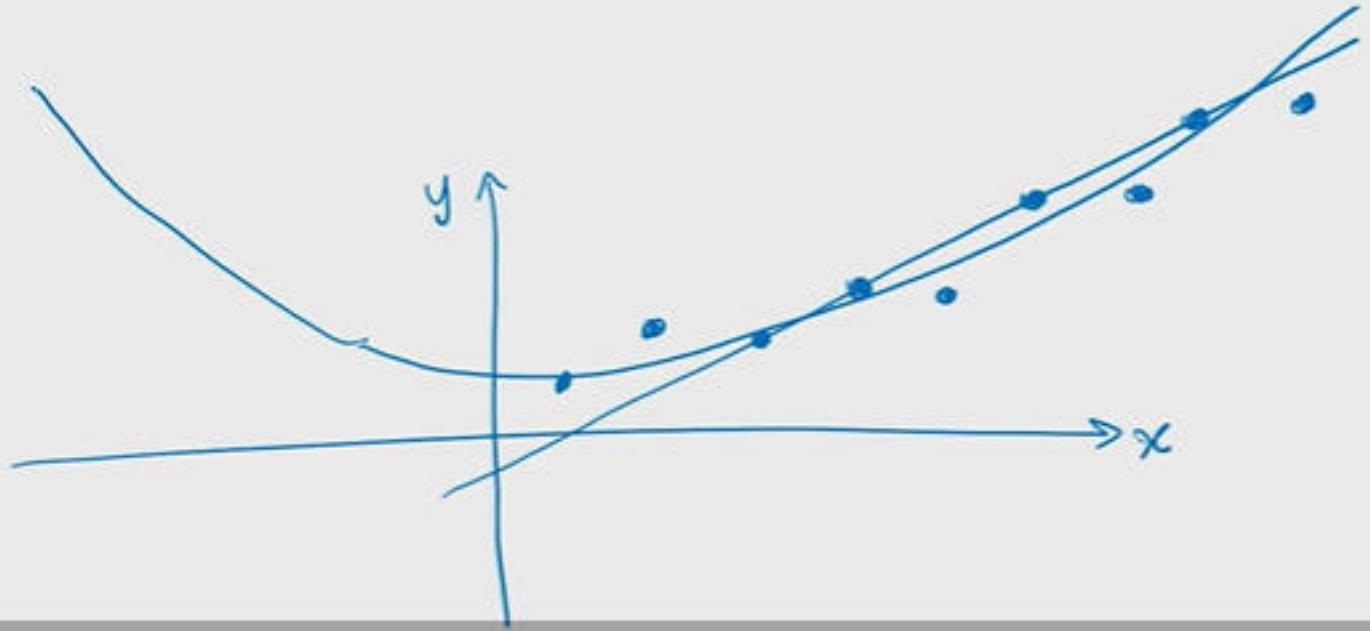


$$y = ax + b$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Polynomial Regression of Different Degrees



$$y = ax + b$$

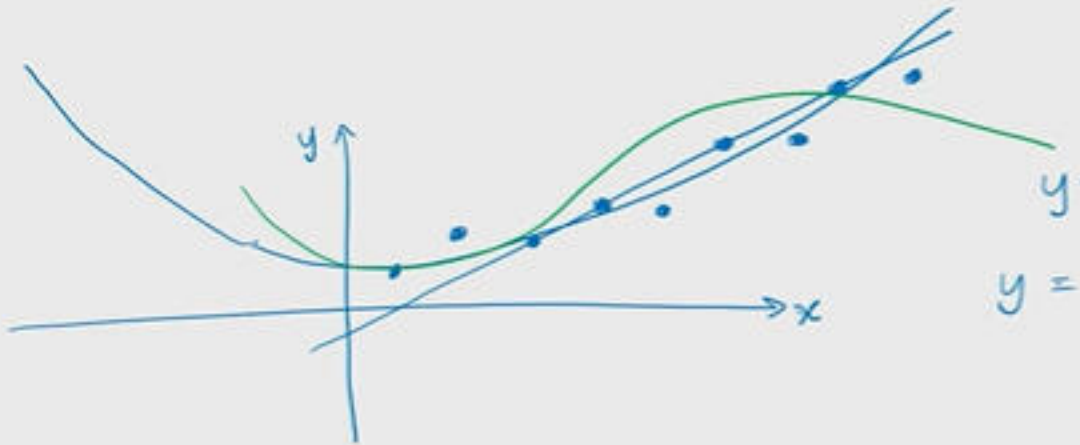
$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

-1:51

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Polynomial Regression of Different Degrees

aws training and certification



$y = ax + b$

$y = \beta_0 + \beta_1 x + \beta_2 x^2$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

-1:24

Polynomial Regression of Different Degrees

aws training and certification

$y = ax + b$

$y = \beta_0 + \beta_1 x + \beta_2 x^2$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

-0:56

Polynomial Regression of Different Degrees

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Solution: Regularization

aws training and certification

Regularization : penalize model complexity

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Solution: Regularization

Regularization : penalize model complexity

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Solution: Regularization

aws training and certification

Regularization : penalize model complexity

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Solution: Regularization

aws training and certification

Regularization : penalize model complexity

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Regularization

Earlier when you computed the coefficients of linear model (ß) using MSE or MLE, you did not opt any model for ß. This was because you were unaware of the coefficients and didn't know of a model to represent them. Now, we'll explore regularization from the Bayesian point of view, learn how to obtain the model for ß using Ridge Regression and LASSO Regression techniques, and interpret the models using geometrical figures.

aws training and certification

Regularization

-20:30

Topics

- The Bayesian Point of View
- Ridge Regression
- LASSO Regression

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Topics

- The Bayesian Point of View
- Ridge Regression
- LASSO Regression
- The Geometry of Regularization
- The Geometry of Ridge Regularization
- The Geometry of LASSO Regularization

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Topics

- The Bayesian Point of View
- Ridge Regression
- LASSO Regression
- The Geometry of Regularization
- The Geometry of Ridge Regularization
- The Geometry of LASSO Regularization
- Elastic Net Regression

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Topics

- The Bayesian Point of View
- Ridge Regression
- LASSO Regression
- The Geometry of Regularization
- The Geometry of Ridge Regularization
- The Geometry of LASSO Regularization
- Elastic Net Regression

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Bayesian Point of View

aws training and certification

MLE : $\bar{y} = X\bar{\beta} + N(0, \sigma^2)$

maximize $L(\beta)$: β^{MLE}

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Bayesian Point of View

aws training and certification

MLE : $\bar{y} = \underbrace{X\bar{\beta}} + \underbrace{N(0, \sigma^2)} = N(X\bar{\beta}, \sigma^2)$

$\rightarrow$ $\underbrace{\text{maximize } L(\beta)}$: β^{MLE}

The Bayesian Point of View

MLE : $\underbrace{\bar{y}} = \underbrace{X\bar{\beta}} + \underbrace{N(0, \sigma^2)} = N(\underbrace{X}\bar{\beta}, \sigma^2)$

$\underbrace{\text{maximize } L(\beta)} : \beta^{MLE}$

$$L(\beta) = \prod_{i=1}^{N} f(X, Y \mid \underline{\beta})$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Bayesian Point of View

MLE : $\underbrace{y} = \underbrace{X\bar{\beta}} + \underbrace{N(0, \sigma^2)} = N(\underbrace{X}\bar{\beta}, \sigma^2)$

$\underline{\text{maximize } L(\beta)} : \beta^{MLE}$

$$L(\beta) = \prod_{i=1}^{N} f(x, y \mid \beta)$$

$\underbrace{P(\beta)}$ ← probability distribution on β

$$\tilde{L}(\beta) = L(\beta)$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Ridge Regression

$$P(\beta) = e^{-\delta^2 \|\beta\|_2^2}$$

$\|\cdot\|_2^2 \rightarrow$ L2 norm of vector

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Ridge Regression

$$\underline{P(\beta)} = e^{-\delta^2 \|\beta\|_2^2}$$

$\|\cdot\|_2^2 \rightarrow$ L2 norm of vector

$$\tilde{L}(\beta) = L(\beta)\, P(\beta)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(-\frac{1}{2}\|\bar{y} - X\bar{\beta}\|_2^2\right) \cdot \exp\left(-\delta^2 \|\beta\|_2^2\right)$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Ridge Regression

$$\underline{P(\beta)} = e^{-\delta^2 \|\beta\|_2^2}$$

$\|\cdot\|_2^2 \rightarrow$ L2 norm of vector

$$\tilde{L}(\beta) = L(\beta)\, P(\beta)$$

$$= \underbrace{\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N} \exp\left(-\frac{1}{2}\|\bar{y} - X\bar{\beta}\|_2^2\right) \cdot \exp\left(-\delta^2 \|\beta\|_2^2\right)$$

$$\log\left(\tilde{L}(\beta)\right) = -\frac{N}{2}\log\left(2\pi\sigma^2\right)$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Ridge Regression

$$\underline{P(\beta)} = e^{-\delta^2 \|\beta\|_2^2}$$

$\|\cdot\|_2^2 \rightarrow$ L2 norm of vector

$$\tilde{L}(\beta) = L(\beta)\, P(\beta)$$

$$= \underbrace{\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N} \underbrace{\exp\left(-\frac{1}{2}\|\bar{y} - X\bar{\beta}\|_2^2\right) \cdot \exp\left(-\delta^2 \|\beta\|_2^2\right)}$$

$$\log(\tilde{L}(\beta)) = -\frac{N}{2}\log(2\pi\sigma^2) + \left(-\frac{1}{2}\|\bar{y} - X\bar{\beta}\|_2^2 - \delta^2\|\beta\|_2^2\right)$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Ridge Regression

$$\underbrace{P(\beta)} = e^{-\delta^2 \|\beta\|_2^2}$$

$\|\cdot\|_2^2 \rightarrow$ L2 norm of vector

$$\tilde{L}(\beta) = L(\beta)\, P(\beta)$$

$$= \underbrace{\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N} \underbrace{\exp\left(-\frac{1}{2}\|\bar{y} - X\bar{\beta}\|_2^2\right) \cdot \exp\left(-\delta^2 \|\beta\|_2^2\right)}$$

$$\underbrace{\log\left(\tilde{L}(\beta)\right)} = \underbrace{-\frac{N}{2}\log\left(2\pi\sigma^2\right)} + \underbrace{\left(-\frac{1}{2}\|\bar{y} - X\bar{\beta}\|_2^2 - \delta^2 \|\beta\|_2^2\right)}$$

minimize

$\rightarrow 0$

$\equiv$ minimize

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Ridge Regression

$$\underbrace{P(\beta)} = e^{-\delta^2 \|\beta\|_2^2}$$

$\|\cdot\|_2^2 \rightarrow$ L2 norm of vector

$$\tilde{L}(\beta) = L(\beta)\, P(\beta)$$

$$= \underbrace{\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N} \underbrace{\exp\left(-\frac{1}{2}\|\bar{y} - X\bar{\beta}\|_2^2\right) \cdot \exp\left(-\delta^2 \|\beta\|_2^2\right)}$$

$$\underbrace{\log\left(\tilde{L}(\beta)\right)} = \underbrace{-\frac{N}{2}\log(2\pi\sigma^2)}_{0} + \underbrace{\left(-\frac{1}{2}\|\bar{y} - X\bar{\beta}\|_2^2 - \delta^2 \|\beta\|_2^2\right)}$$

minimize

$$\equiv \text{minimize } \frac{1}{2}\|\bar{y} - X\bar{\beta}\|_2^2 + \delta^2 \|\beta\|_2^2$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Ridge Regression

$$P(\beta) = e^{-\delta^2 \|\beta\|_2^2}$$

$\|\cdot\|_2^2 \rightarrow$ L2 norm of vector

$$\tilde{L}(\beta) = L(\beta)\, P(\beta)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(-\frac{1}{2}\|\bar{y} - X\bar{\beta}\|_2^2\right) \cdot \exp\left(-\delta^2 \|\beta\|_2^2\right)$$

$$\log(\tilde{L}(\beta)) = -\frac{N}{2}\log(2\pi\sigma^2) + \left(-\frac{1}{2}\|\bar{y} - X\bar{\beta}\|_2^2 - \delta^2\|\beta\|_2^2\right)$$

minimize

0

$$\beta_{reg} = \underset{\beta}{\text{minimize}} \; \frac{1}{2}\|\bar{y} - X\bar{\beta}\|_2^2 + \delta^2 \|\beta\|_2^2$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Ridge Regression

$$\frac{d}{d\beta} \log \tilde{L}(\beta) = 0$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Ridge Regression

$$\frac{d}{d\beta} \log \tilde{L}(\beta) = 0$$

$$\Rightarrow \quad \beta^{ridge} = \left(X^T X + \delta^2 I \right)^{-1} X^T \underline{y}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Ridge Regression

$$\frac{d}{d\beta} \log \tilde{L}(\beta) = 0$$

$$\Rightarrow \quad \beta^{ridge} = \left(X^T X + \delta^2 \underbrace{I}_{\text{identity } r} \right)^{-1} X^T \underline{y}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Ridge Regression

$$\frac{d}{d\beta} \log \tilde{L}(\beta) = 0$$

$$\Rightarrow \quad \beta^{ridge} = \left(X^T X + \delta^2 \underbrace{I}_{\text{identity matrix}} \right)^{-1} X^T \underline{y}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

LASSO Regression

$$P(\beta) = e^{-\delta^2 \|\beta\|_1}$$

$\|\cdot\|_1$ = L1 -norm of vector

$$\log \tilde{L}(\beta) = \log L(\beta) + \log P(\beta)$$

$$= -$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

LASSO Regression

$$P(\beta) = e^{-\delta^2 \|\beta\|_1}$$

$\|\cdot\|_1$ = L1 -norm of vector

$$\log \tilde{L}(\beta) = \log L(\beta) + \log P(\beta)$$

$$= -\frac{N}{2}\log(2\pi\sigma^2) + \left(-\frac{1}{2}\|\bar{y} - X\bar{\beta}\|_2^2 - \delta^2\|\beta\|_1\right)$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

LASSO Regression

$$P(\beta) = e^{-\delta^2 \|\beta\|_1}$$

$\|\cdot\|_1$ = L1 -norm of vector

$$\log \tilde{L}(\beta) = \log L(\beta) + \log P(\beta)$$

$$= -\frac{N}{2}\log(2\pi\sigma^2) + \left(-\frac{1}{2}\|\bar{y} - X\bar{\beta}\|_2^2 - \delta^2\|\beta\|_1\right)$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

LASSO Regression

$$P(\beta) = e^{-\delta^2 \|\beta\|_1}$$

$\|\cdot\|_1$ = L1 -norm of vector

$$\log \tilde{L}(\beta) = \log L(\beta) + \log P(\beta)$$

$$= -\frac{N}{2}\log(2\pi\sigma^2) + \left(\underbrace{-\frac{1}{2}\|\bar{y} - X\bar{\beta}\|_2^2 - \delta^2\|\beta\|_1}\right)$$

$$\beta^{LASSO} = \min$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

LASSO Regression

$$P(\beta) = e^{-\delta^2 \|\beta\|_1}$$

$\|\cdot\|_1$ = L1 -norm of vector

$$\log \tilde{L}(\beta) = \log L(\beta) + \log P(\beta)$$

$$= -\frac{N}{2}\log(2\pi\sigma^2) + \underbrace{\left(-\frac{1}{2}\|\bar{y} - X\bar{\beta}\|_2^2 - \delta^2\|\beta\|_1\right)}$$

$$\beta^{LASSO} = \text{minimize} \underbrace{\left(\|\bar{y} - X\bar{\beta}\|_2^2 + \delta^2\|\beta\|_1\right)}$$

no closed form solution

gradient descent, SGD

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Geometry of Regularization

aws training and certification

The fact that our loss function is now the **sum of two loss functions** gives us a new geometric interpretation.

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Geometry of Regularization

The fact that our loss function is now the **sum of two loss functions** gives us a new geometric interpretation.

$$\text{loss function} = \|\bar{y} - X\bar{\beta}\|_2^2 + \delta^2 \|\beta\|.$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Geometry of Regularization

The fact that our loss function is now the **sum of two loss functions** gives us a new geometric interpretation.

$$\text{loss function} = \|\bar{y} - X\bar{\beta}\|_2^2 + \delta^2 \|\beta\|_{1/2}$$

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Geometry of Regularization

The fact that our loss function is now the **sum of two loss functions** gives us a new geometric interpretation.

$$\text{loss function} = \underbrace{\|\bar{y} - X\bar{\beta}\|_2^2}_{L1} + \underbrace{\delta^2 \|\beta\|_{1/2}}_{L2}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Geometry of Regularization

The fact that our loss function is now the **sum of two loss functions** gives us a new geometric interpretation.

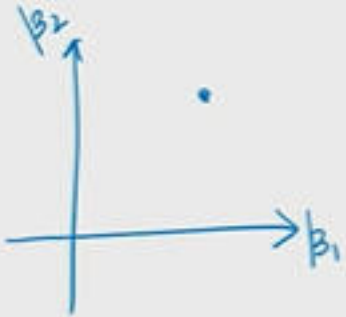
$$\text{loss function} = \underbrace{\|\bar{y} - X\bar{\beta}\|_2^2}_{\substack{L1 \\ \text{MLE}}} + \underbrace{\delta^2 \|\beta\|_{1/2}}_{\substack{L2 \\ \text{Reg}}}$$

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Geometry of Regularization

The fact that our loss function is now the **sum of two loss functions** gives us a new geometric interpretation.

$$\text{loss function} = \underbrace{\|\bar{y} - X\bar{\beta}\|_2^2}_{\substack{L1 \\ \text{MLE}}} + \underbrace{\delta^2 \|\beta\|_{1/2}}_{\substack{L2 \\ \text{Regularization term}}}$$



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Geometry of Regularization

The fact that our loss function is now the **sum of two loss functions** gives us a new geometric interpretation.

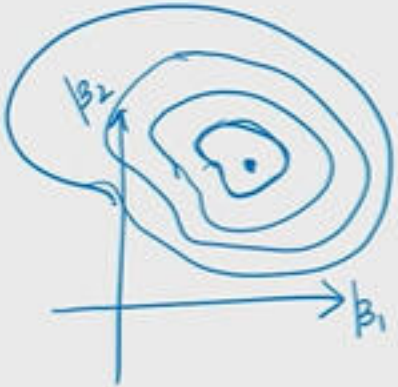
$$\text{loss function} = \underbrace{\|\bar{y} - X\bar{\beta}\|_2^2}_{\substack{L1 \\ \text{MLE}}} + \underbrace{\delta^2 \|\beta\|_{1/2}}_{\substack{L2 \\ \text{Regularization term}}}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Geometry of Regularization

The fact that our loss function is now the **sum of two loss functions** gives us a new geometric interpretation.

$$\text{loss function} = \underbrace{\|\bar{y} - X\bar{\beta}\|_2^2}_{\substack{L1 \\ \text{MLE}}} + \underbrace{\delta^2 \|\beta\|_{1/2}}_{\substack{L2 \\ \text{Regularization term}}}$$

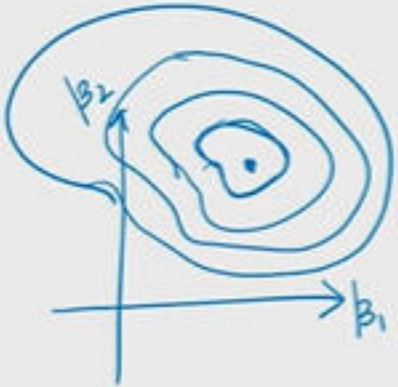


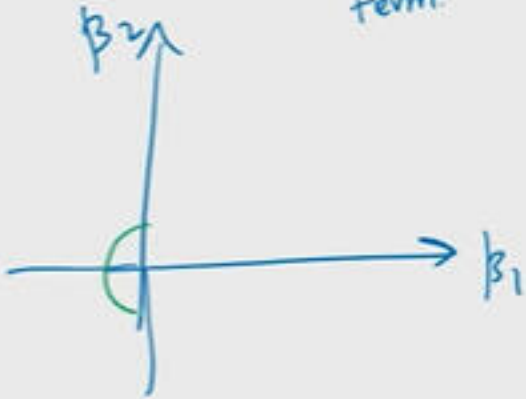
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Geometry of Regularization

The fact that our loss function is now the **sum of two loss functions** gives us a new geometric interpretation.

$$\text{loss function} = \underbrace{\|\bar{y} - X\bar{\beta}\|_2^2}_{\substack{L1 \\ \text{MLE}}} + \underbrace{\delta^2 \|\beta\|_{1/2}}_{\substack{L2 \\ \text{Regularization term}}}$$



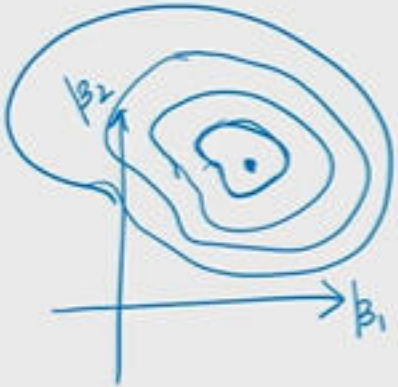


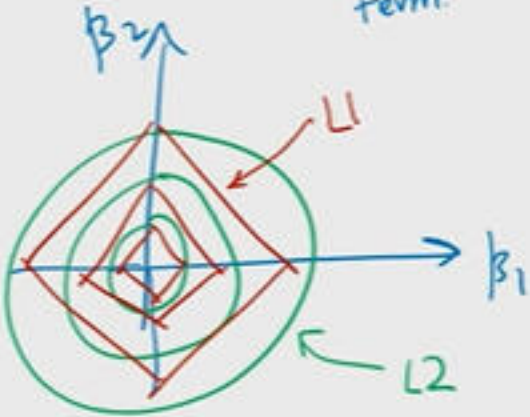
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Geometry of Regularization

The fact that our loss function is now the **sum of two loss functions** gives us a new geometric interpretation.

$$\text{loss function} = \underbrace{\|\bar{y} - X\bar{\beta}\|_2^2}_{\substack{L1 \\ \text{MLE}}} + \underbrace{\delta^2 \|\beta\|_{1/2}}_{\substack{L2 \\ \text{Regularization term}}}$$

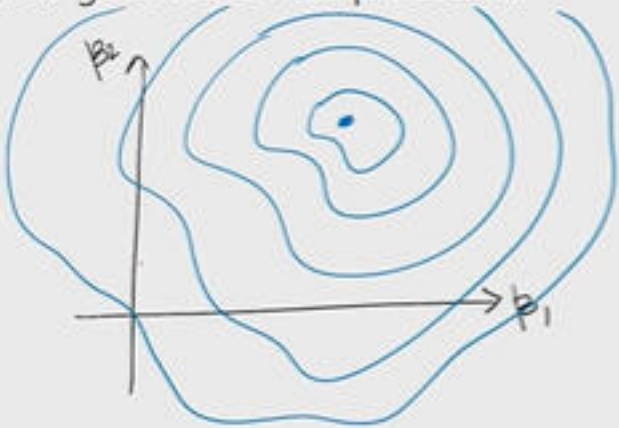




© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Geometry of Ridge Regularization

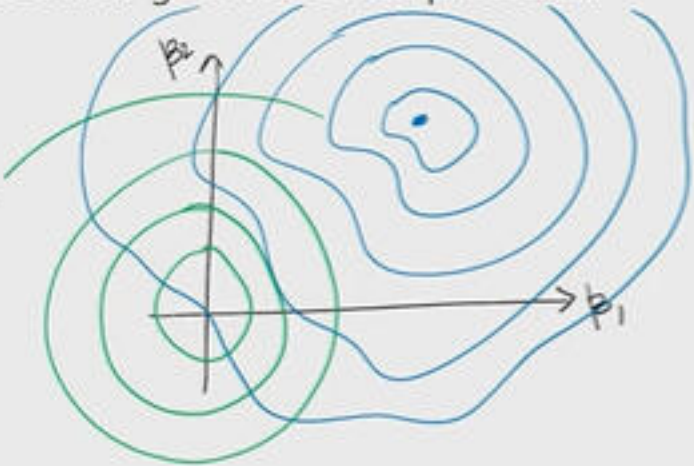
The fact that our loss function is now the **sum of two loss functions** gives us a new geometric interpretation.



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Geometry of Ridge Regularization

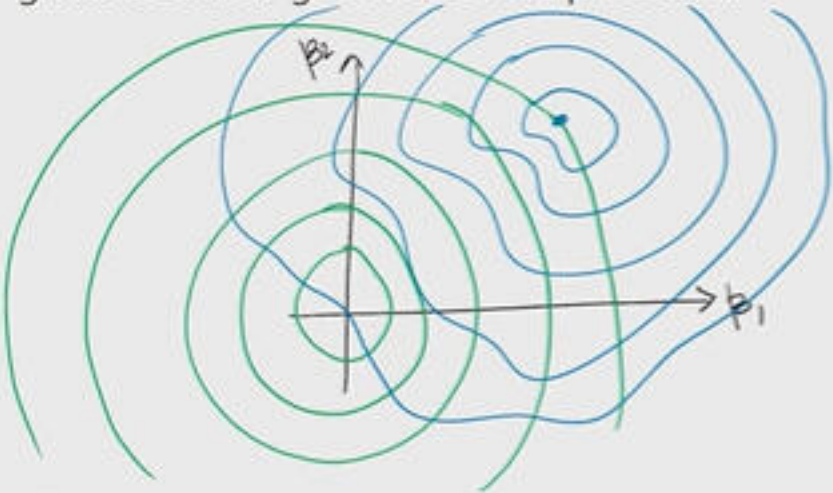
The fact that our loss function is now the **sum of two loss functions** gives us a new geometric interpretation.



© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Geometry of Ridge Regularization

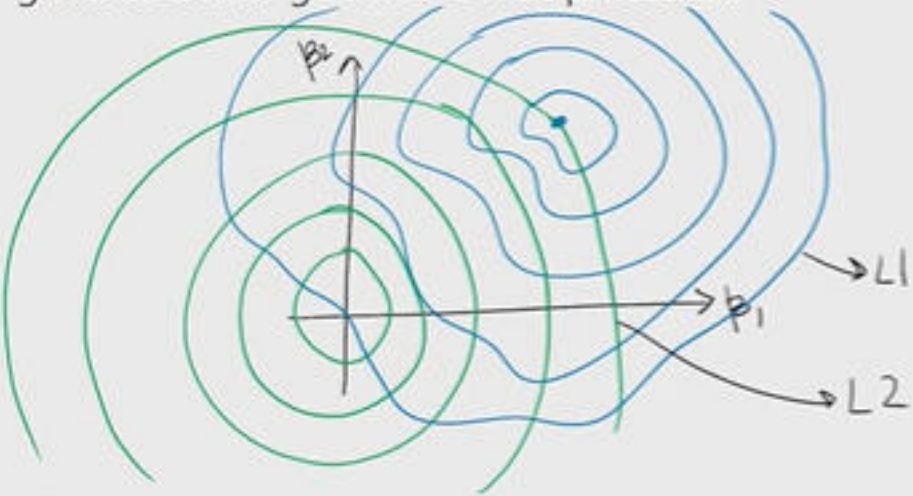
The fact that our loss function is now the **sum of two loss functions** gives us a new geometric interpretation.



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Geometry of Ridge Regularization

The fact that our loss function is now the **sum of two loss functions** gives us a new geometric interpretation.



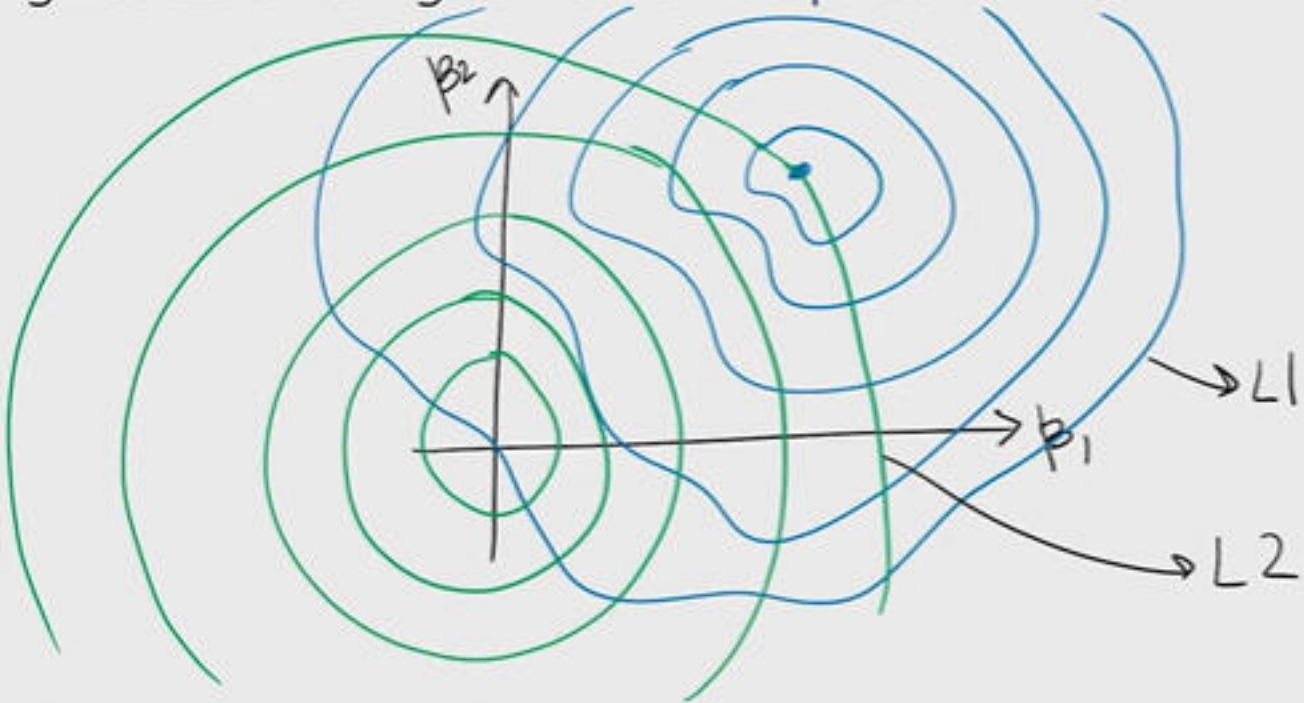
$$\frac{d\,\tilde{L}(\beta)}{d\beta} = 0$$

$$\frac{dL_1(\beta)}{d\beta} = -\frac{dL_2(\beta)}{d\beta}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Geometry of Ridge Regularization

The fact that our loss function is now the **sum of two loss functions** gives us a new geometric interpretation.



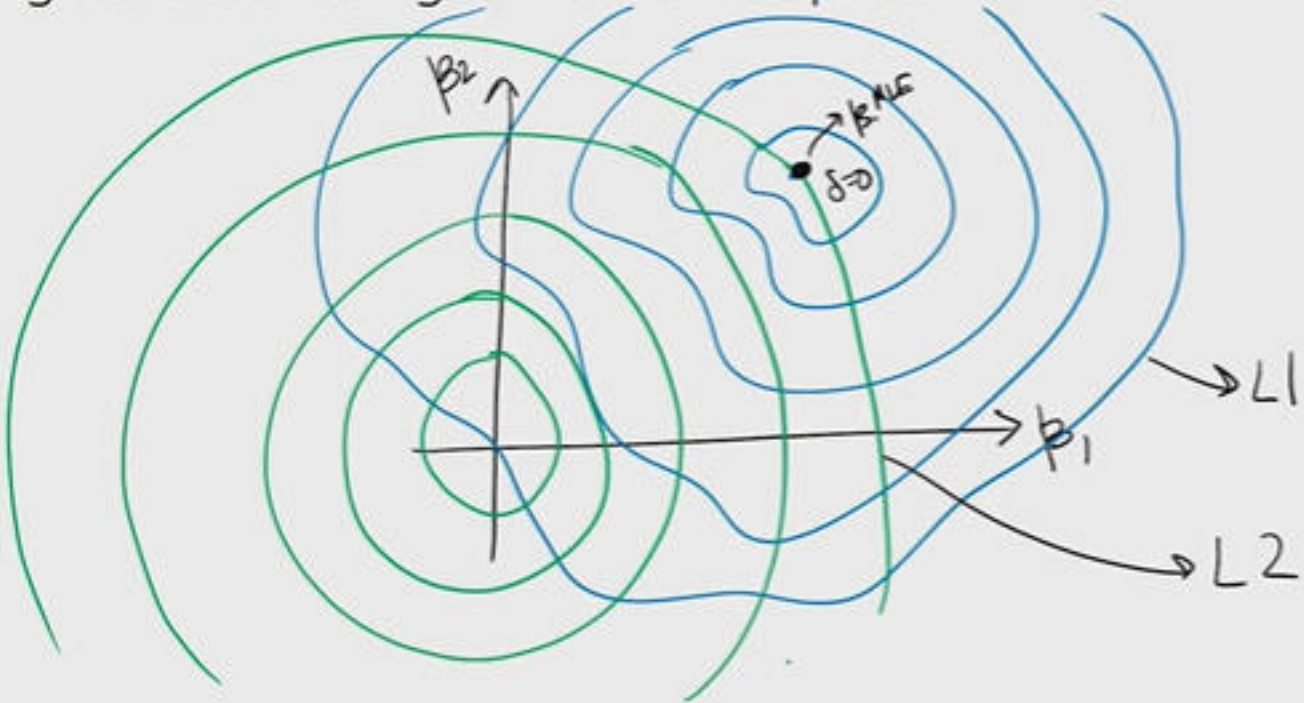
$$\frac{d\tilde{L}(\beta)}{d\beta} = 0$$

$$\underbrace{\frac{dL_1(\beta)}{d\beta}} = \underbrace{-\frac{\delta dL_2(\beta)}{d\beta}}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Geometry of Ridge Regularization

The fact that our loss function is now the **sum of two loss functions** gives us a new geometric interpretation.



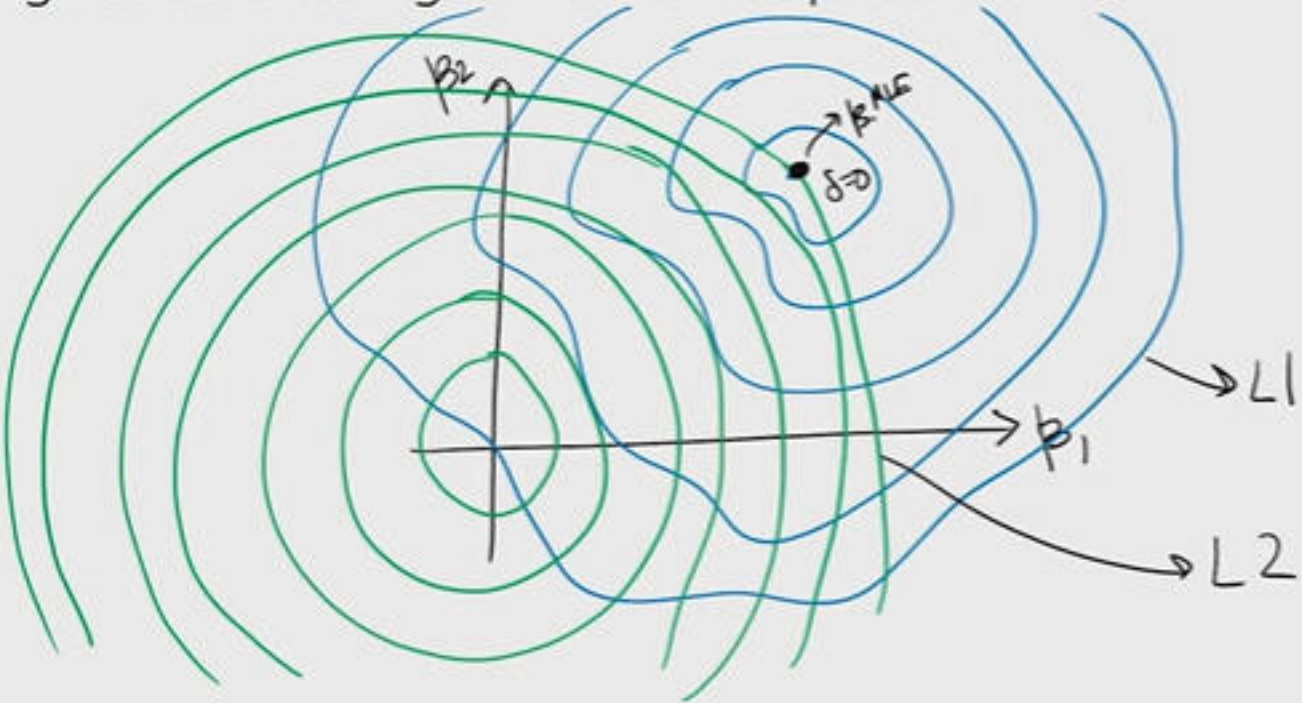
$$\frac{d\,\tilde{L}(\beta)}{d\beta} = 0$$

$$\underbrace{\frac{dL_1(\beta)}{d\beta}} = \underbrace{-\delta \frac{dL_2(\beta)}{d\beta}}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Geometry of Ridge Regularization

The fact that our loss function is now the **sum of two loss functions** gives us a new geometric interpretation.



$$\frac{d\,\tilde{L}(\beta)}{d\beta} = 0$$

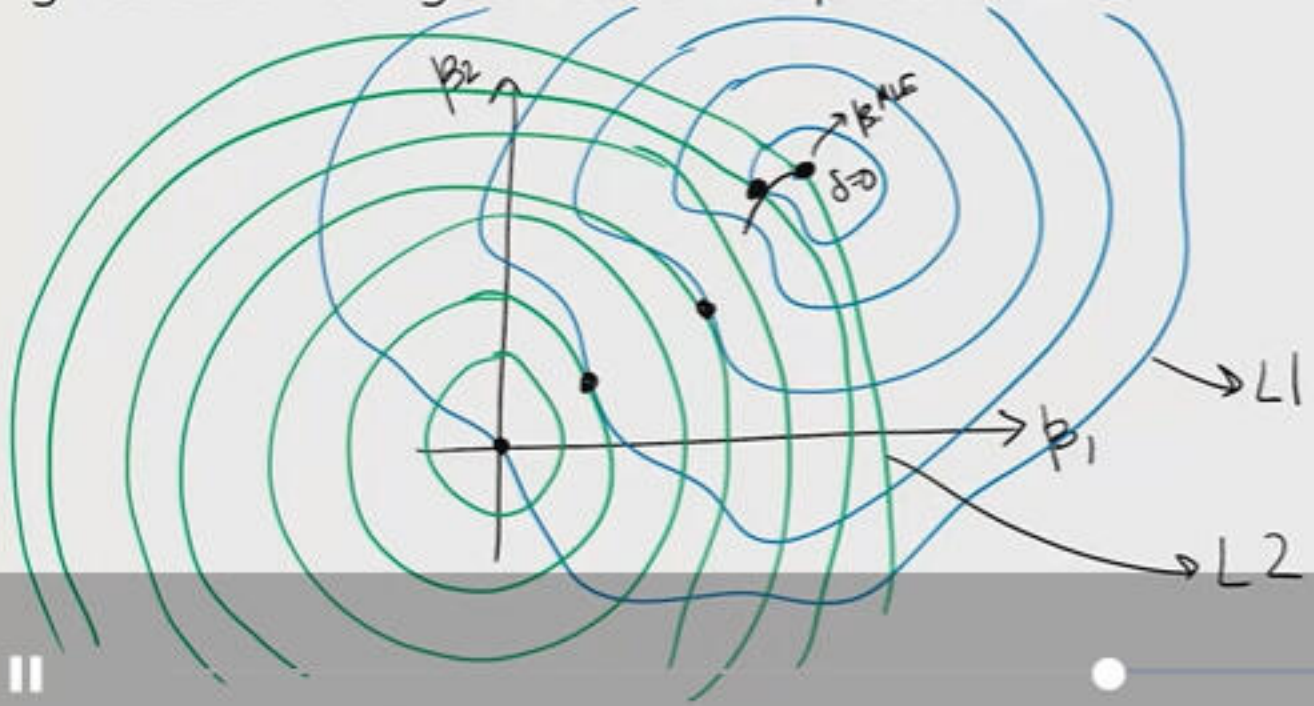
$$\underbrace{\frac{dL_1(\beta)}{d\beta}} = \underbrace{-\delta \frac{dL_2(\beta)}{d\beta}}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Geometry of Ridge Regularization

aws training and certification

The fact that our loss function is now the **sum of two loss functions** gives us a new geometric interpretation.



$$\frac{d\,\tilde{L}(\beta)}{d\beta} = 0$$

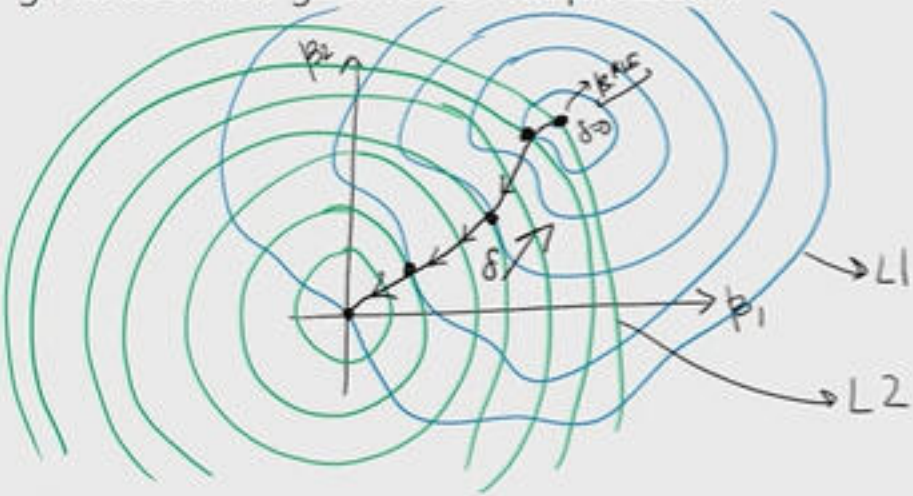
$$\underbrace{\frac{dL_1(\beta)}{d\beta}} = \underbrace{-\delta \frac{dL_2(\beta)}{d\beta}}$$

-5:04

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Geometry of Ridge Regularization

The fact that our loss function is now the **sum of two loss functions** gives us a new geometric interpretation.

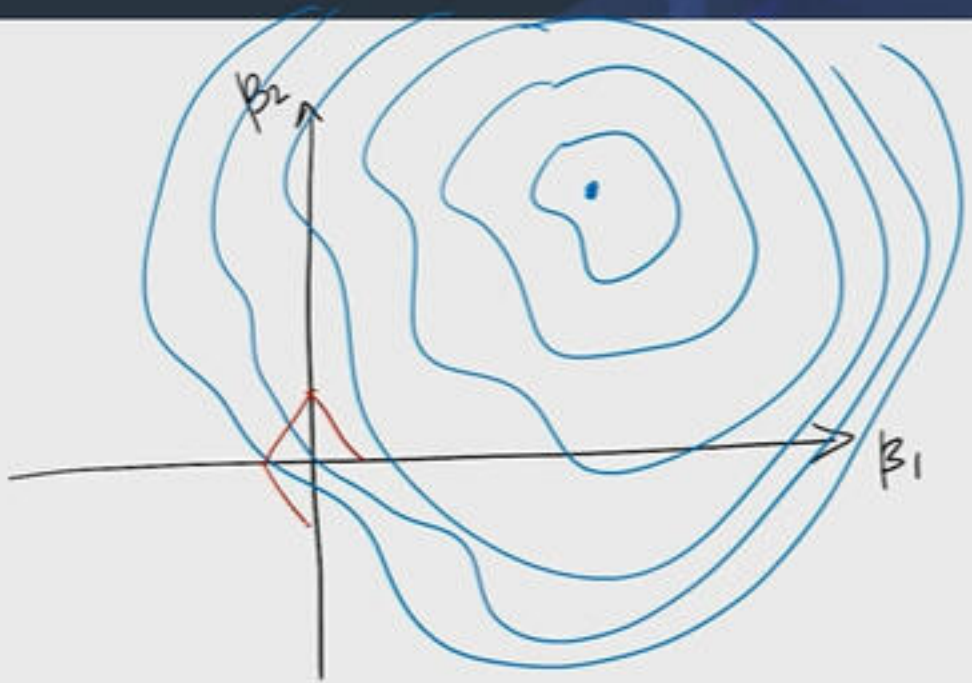


$$\frac{d\,\tilde{L}(\beta)}{d\beta} = 0$$

$$\underbrace{\frac{dL_1(\beta)}{d\beta}} = \underbrace{-\frac{\delta dL_2(\beta)}{d\beta}}$$

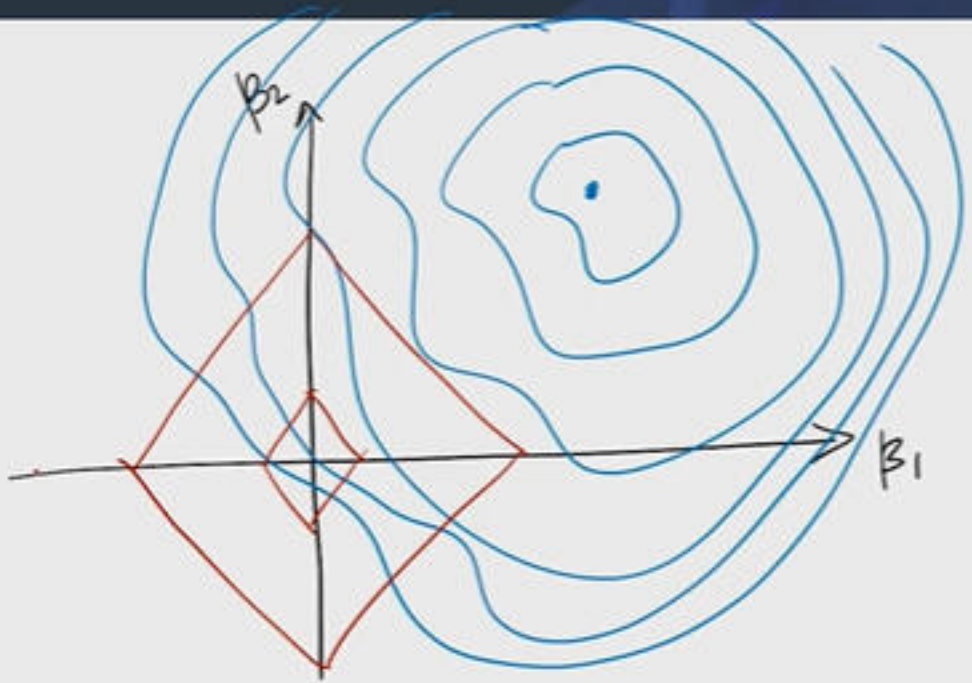
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Geometry of LASSO Regularization



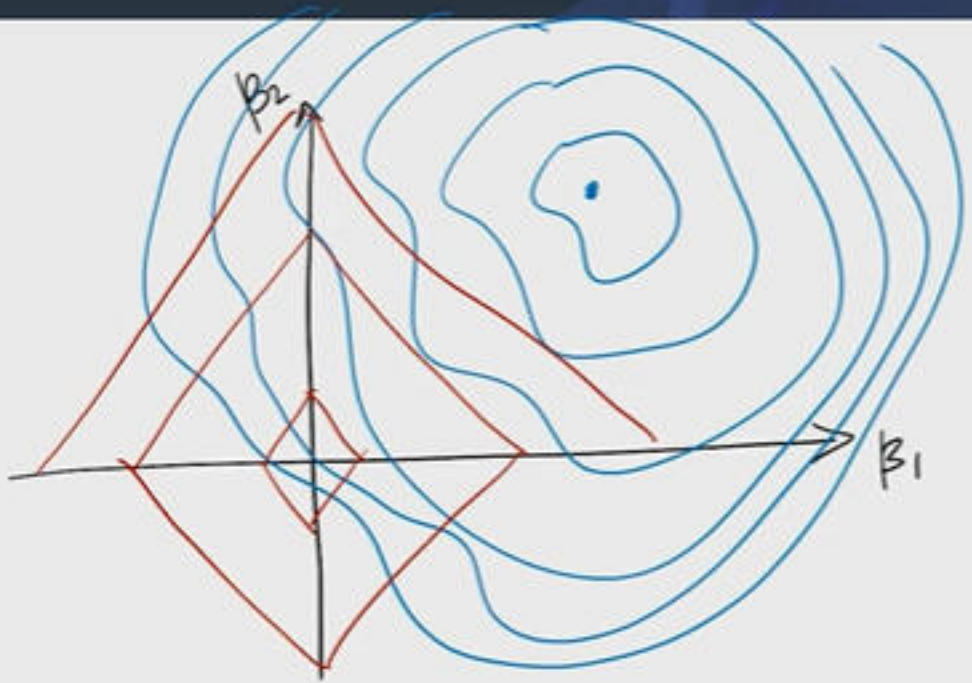
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Geometry of LASSO Regularization



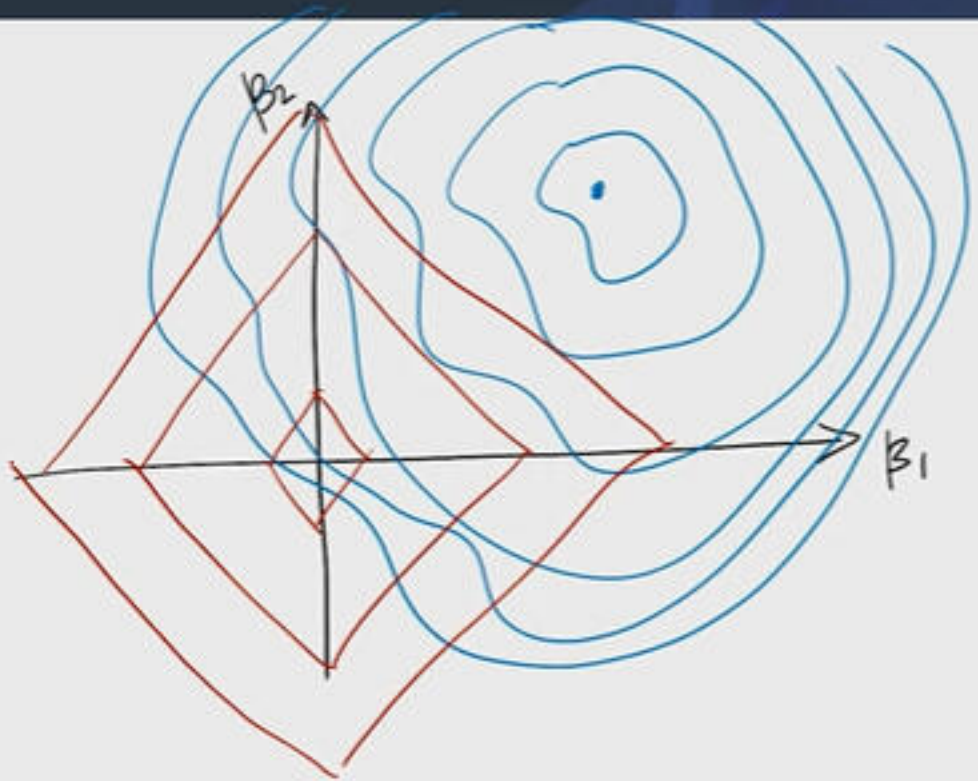
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Geometry of LASSO Regularization



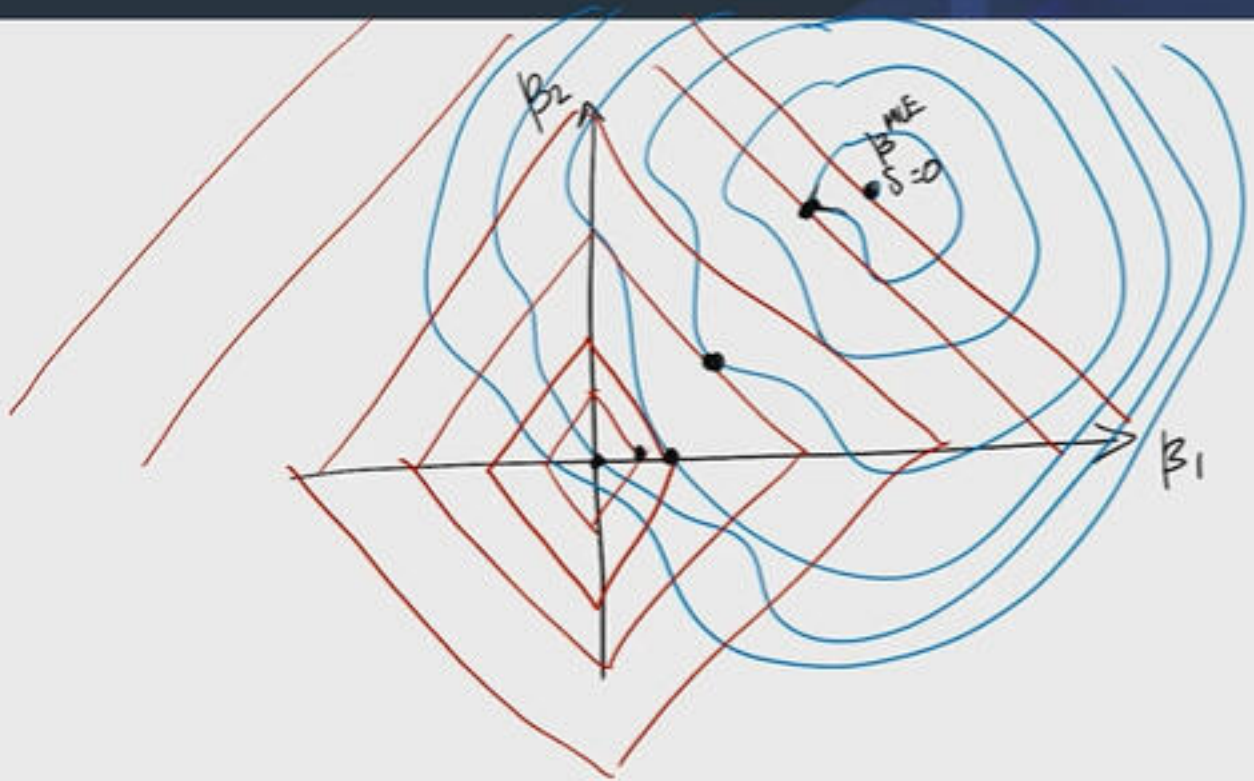
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Geometry of LASSO Regularization



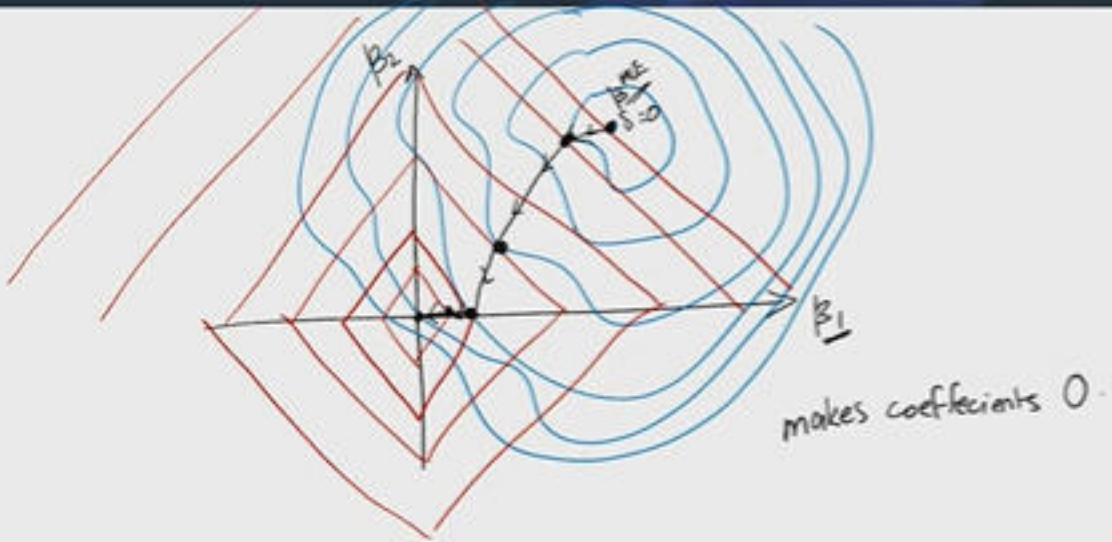
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Geometry of LASSO Regularization



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Geometry of LASSO Regularization



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Elastic Net Regression

This is nothing more than a combination of the two previous techniques.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

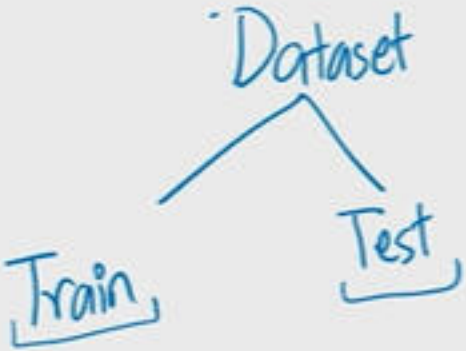
Elastic Net Regression

This is nothing more than a combination of the two previous techniques.

$$L = \|\bar{y} - X\bar{\beta}\|_2^2 + \delta_1^2 \|\beta\|_1 + \delta_2^2 \|\beta\|_2^2$$

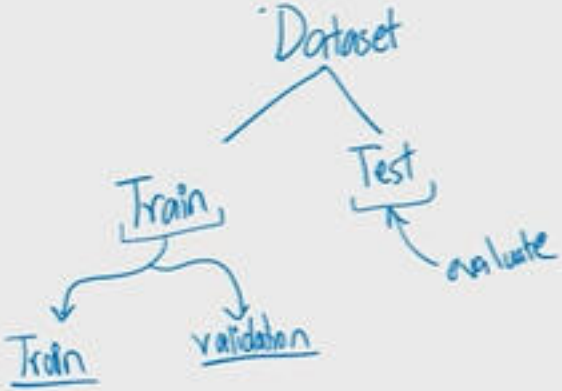
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

How do we pick δ ?



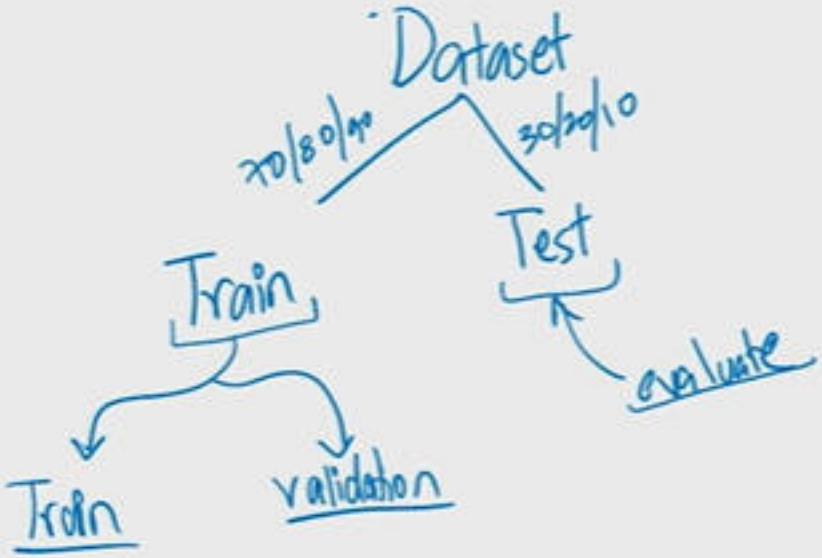
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

How do we pick δ ?



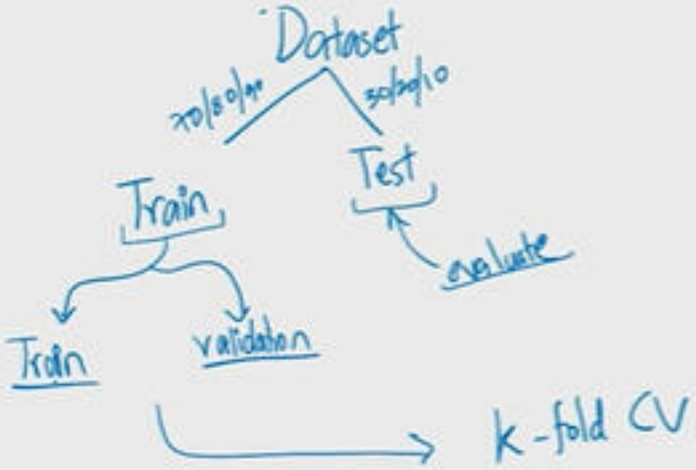
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

How do we pick δ ?



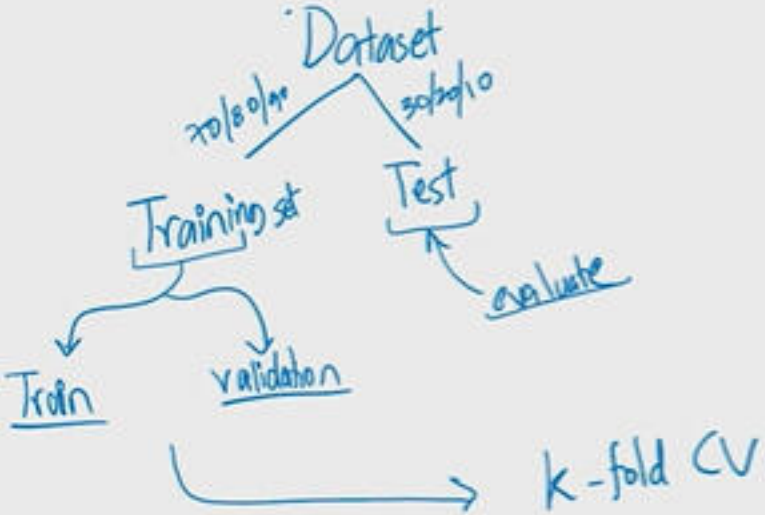
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

How do we pick δ ?



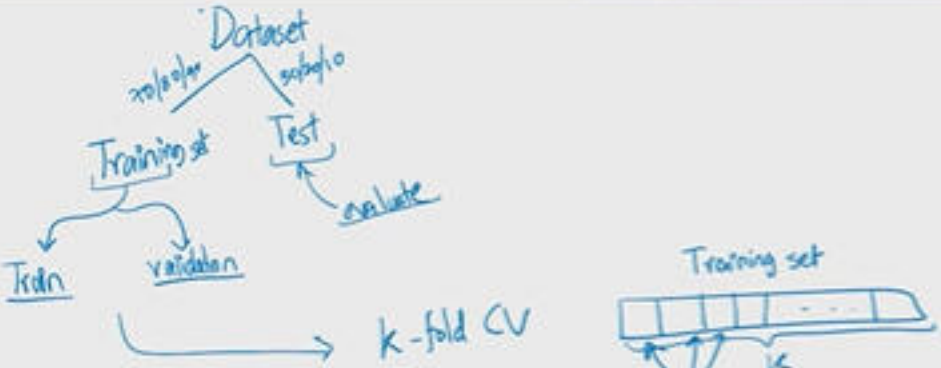
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

How do we pick δ ?



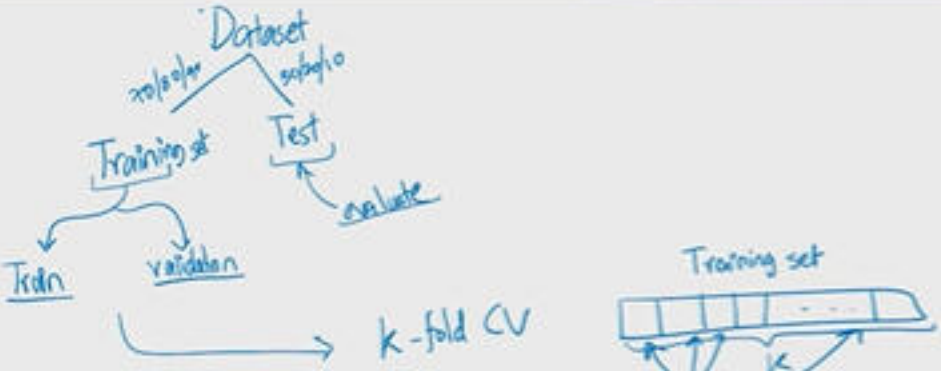
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

How do we pick δ ?



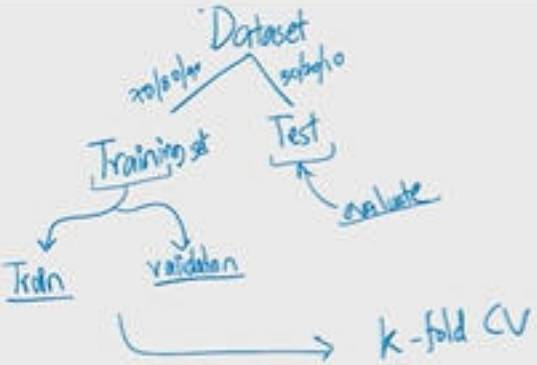
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

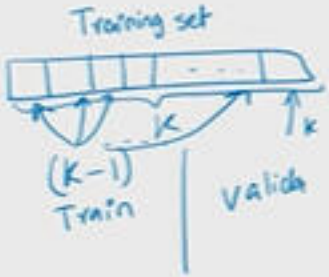
How do we pick δ ?



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

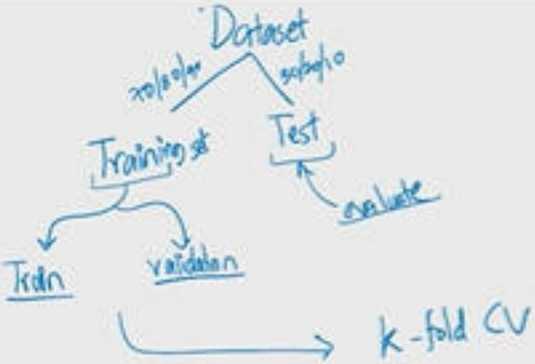
How do we pick δ ?

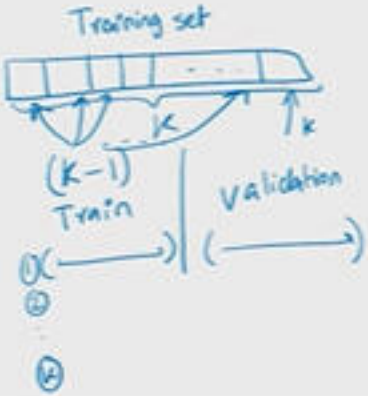




© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

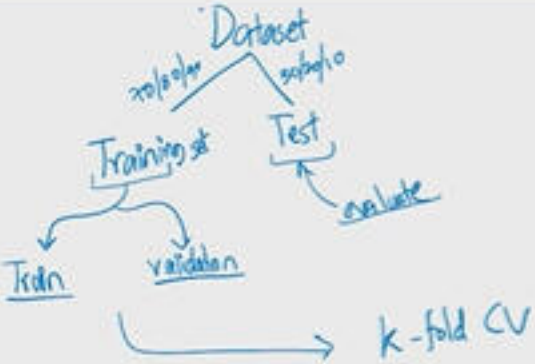
How do we pick δ ?

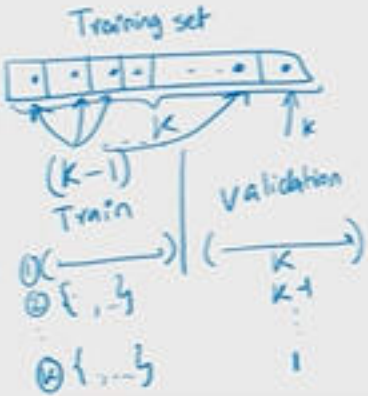




© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

How do we pick δ ?





© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Alternative Interpretation of Regularization

$$L = \|\bar{y} - X\bar{\beta}\|_2^2$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Alternative Interpretation of Regularization

aws training and certification

$$L = \|\bar{y} - X\bar{\beta}\|_2^2$$

unconstrained → overfitting

solution: penalize large values of "β".

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Alternative Interpretation of Regularization

Ridge regularization:

$$L = \|\bar{y} - X\bar{\beta}\|_2^2 + \delta.$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Alternative Interpretation of Regularization

Ridge regularization:

$$L = \|\bar{y} - X\bar{\beta}\|_2^2 + \delta_2 \underbrace{\|\beta\|_2^2}_{\text{uniformly penalizes}}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Alternative Interpretation of Regularization

aws training and certification

Ridge regularization:

$$\mathcal{L} = \|\bar{y} - X\bar{\beta}\|_2^2 + \delta_2 \|\beta\|_2^2$$

uniformly penalizes

LASSO regularization,

$$\mathcal{L} = \|\bar{y} - X\bar{\beta}\|_2$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Alternative Interpretation of Regularization

Ridge regularization:

$$\mathcal{L} = \|\bar{y} - X\bar{\beta}\|_2^2 + \delta_2 \|\beta\|_2^2$$

uniformly penalizes

LASSO regularization,

$$\mathcal{L} = \|\bar{y} - X\bar{\beta}\|_2^2 + \delta_1 \|\beta\|_1$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Alternative Interpretation of Regularization

Ridge regularization:

$$\mathcal{L} = \|\bar{y} - X\bar{\beta}\|_2^2 + \delta_2 \|\beta\|_2^2$$

uniformly penalizes

LASSO regularization:

$$\mathcal{L} = \|\bar{y} - X\bar{\beta}\|_2^2 + \delta_1 \|\beta\|_1$$

unimportant coeff

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Alternative Interpretation of Regularization

Ridge regularization:

$$\mathcal{L} = \|\bar{y} - X\bar{\beta}\|_2^2 + \delta_2 \|\beta\|_2^2$$

uniformly penalizes

LASSO regularization:

$$\mathcal{L} = \|\bar{y} - X\bar{\beta}\|_2^2 + \delta_1 \|\beta\|_1$$

unimportant coeff $\rightarrow 0$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Alternative Interpretation of Regularization

aws training and certification

Ridge regularization:

$$\mathcal{L} = \|\bar{y} - X\bar{\beta}\|_2^2 + \delta_2 \|\beta\|_2^2$$

uniformly penalizes

LASSO regularization,

$$\mathcal{L} = \|\bar{y} - X\bar{\beta}\|_2^2 + \delta_1 \|\beta\|_1$$

unimportant coeff $\to 0$

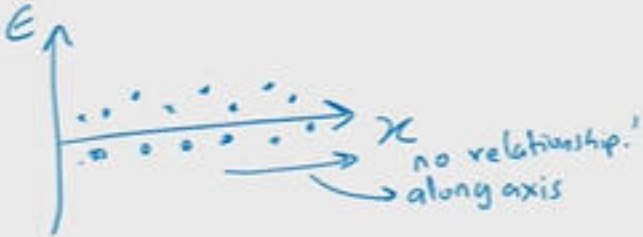
Elastic Net reg.:

$$\mathcal{L} = \|\bar{y} - X\bar{\beta}\|_2^2 + \delta_1 \|\beta\|_1 + \delta_2 \|\beta\|_2^2$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Non-Normality of Residuals?

Residual plot.

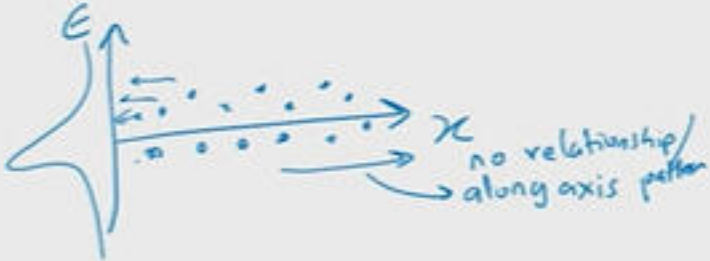


© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Non-Normality of Residuals?

aws training and certification

① Residual plot



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Non-Normality of Residuals?

aws training and certification

Identify ① Residual plot.

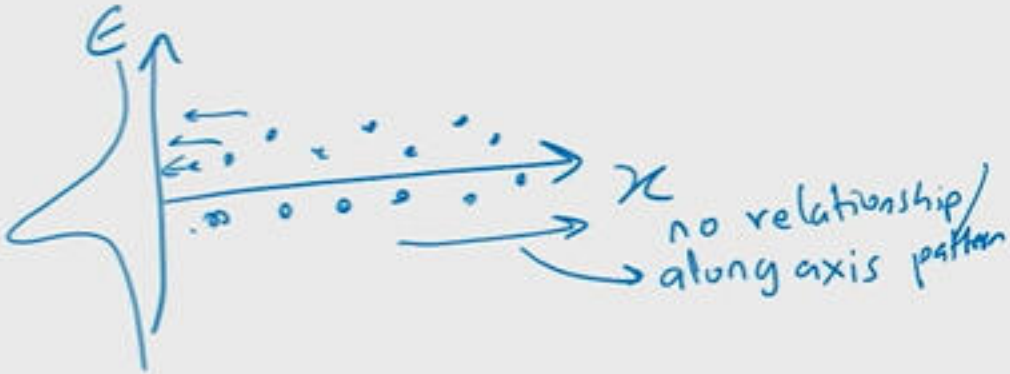
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Non-Normality of Residuals?

aws training and certification

Identify

1. Residual plot.
2. Q-Q plot
3. Kolmogorov-Smirnov test.



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Non-Normality of Residuals?

aws training and certification

Identify
1. Residual plot
2. Q-Q plot
3. Kolmogorov-Smirnov test.

Model: 1. Poisson

© 2016, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Non-Normality of Residuals?

aws training and certification

Identify
1. Residual plot
2. Q-Q plot
3. Kolmogorov-Smirnov test.

Model:
1. Poisson
2. Exponential

© 2016, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

$X^T X$ Not Invertible

$$\beta^* = (X^T X)^{-1} X^T \underline{y}$$

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

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

$X^T X$ Not Invertible

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

$$\beta^* = \underbrace{(X^T X)}^{-1} X^T \underline{y}$$

$$\beta^{\text{ridge}} = (\underbrace{X^T X}_{\text{p.s.d}} + \lambda^2 \underbrace{I}_{\text{identity matrix}})^{-1} X^T \underline{y}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

$X^T X$ Not Invertible

aws training and certification

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

$$\beta^* = \underbrace{(X^T X)}^{-1} X^T \underline{y}$$

$$\beta^{ridge} = \underbrace{\left(\underbrace{X^T X}_{p.s.d} + \lambda^2 \underbrace{I}_{pd}\right)}_{pd}^{-1} X^T \underline{y}$$

identity matrix

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

$X^T X$ Not Invertible

$$\hat{\underline{y}} = X\bar{\beta}$$

$$\beta^* = \underbrace{(X^T X)}^{-1} X^T \underline{y}$$

$$\beta^{ridge} = \underbrace{\left(\underbrace{X^T X}_{p.s.d} + \lambda^2 \underbrace{I}_{pd}\right)}_{pd}^{-1} X^T \underline{y}$$

identity matrix

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

$X^T X$ Not Invertible

aws training and certification

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

$$\beta^* = (X^T X)^{-1} X^T \underline{y}$$

$$\beta^{ridge} = \underbrace{\left(\underbrace{X^T X}_{p.s.d} + \lambda^2 \underbrace{I}_{pd}\right)}_{pd}{}^{-1} X^T \underline{y}$$

identity matrix

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Classification and Logistic Regularization

The OLS model is the best approximation model to predict data structures. However, you can't use it directly because the end model format doesn't fit the classification approach. In this course, we will explore a classification model that provides a finite number of discrete labels.

The following topics will be covered:

- The math behind logistic regression
- Visualizing logistic regression
- Logistic regression loss function and minimizing log likelihood function
- Batch/Full logistic regression
- Gradient descent for OLS

The following topics will be covered:

- The math behind logistic regression
- Visualizing logistic regression
- Logistic regression loss function and minimizing log likelihood function
- Batch/Full logistic regression
- Gradient descent for OLS
- Gradient descent for logistic regression
- Comparing OLS and logistic regression
- Multi-class logistic regression using Softmax function

Logistic Regression

aws training and certification

- Ordinary Least Squares (OLS)
- Logistic Regression
- The Setup
- The Issue
- Definitions
  - Link function
  - Logistic function
- Note on Choice of Link
- Visualizing Logistic Regression

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Logistic Regression

aws training and certification

- Ordinary Least Squares (OLS)
- Logistic Regression
- The Setup
- The Issue
- Definitions
  - Link function
  - Logistic function
- Note on Choice of Link
- Visualizing Logistic Regression

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Logistic Regression Loss

- Our Goal
- The Full Log-Likelihood
- A Note on Minimization

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

(Batch/Full) Gradient Descent

- The Gradient
- Algorithm

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Gradient Descent for Logistic Regression

- Numerical Minimization of Logistic Loss
- Hard Term
- Hard Term 2
- All Terms

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Summary of Calculations

- Written Out Nicely
- A Note on Normalization
- The Algorithm

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Comparison with OLS

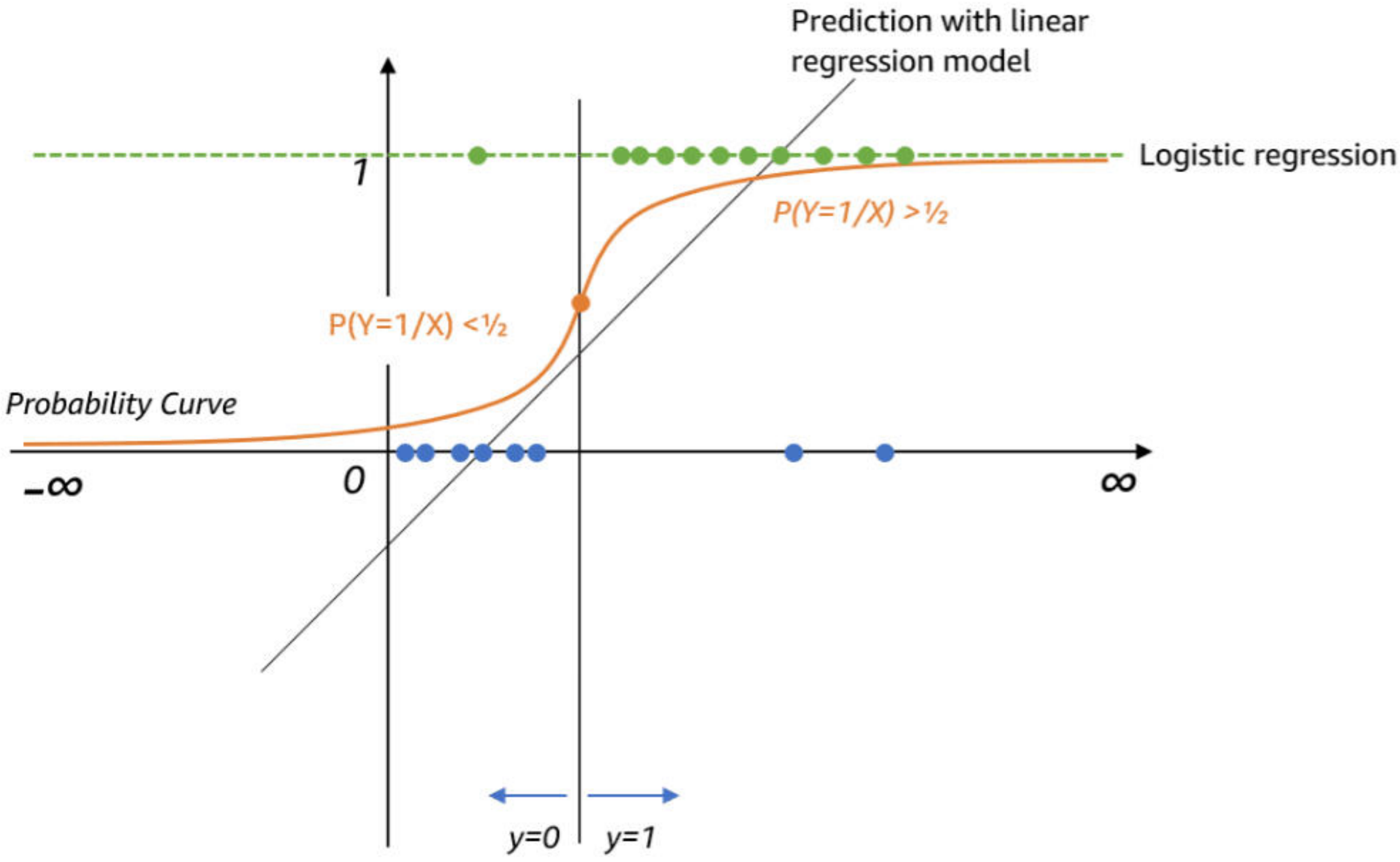
- The Two Update Steps
- Regularization

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Beyond Two Types

- The Softmax
- Comparison with Logistic Regression in the Case of Two Types

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.


Prediction with linear regression model
Logistic regression
1
P(Y=1/X) >½
P(Y=1/X) <½
Probability Curve
-∞
0
∞
y=0
y=1

Topics

- Ordinary Least Squares (OLS)
- The Issue
- Logistic Regression

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Topics

- Ordinary Least Squares (OLS)
- The Issue
- Logistic Regression
- The Setup

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Topics

- Ordinary Least Squares (OLS)
- The Issue
- Logistic Regression
- The Setup
- Definitions
  - Link function
  - Logistic function

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Topics

- Ordinary Least Squares (OLS)
- The Issue
- Logistic Regression
- The Setup
- Definitions
  - Link function
  - Logistic function
- Note on Choice of Link

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Topics

- Ordinary Least Squares (OLS)
- The Issue
- Logistic Regression
- The Setup
- Definitions
  - Link function
  - Logistic function
- Note on Choice of Link
- Visualizing Logistic Regression

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

OLS

aws training and certification

This is regression.

Predicts continuous and potentially unbounded labels based on given features.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

OLS

aws training and certification

This is regression.

Predicts continuous and potentially unbounded labels based on given features.

$$\hat{\bar{y}} = \underbrace{X}_{\text{data matrix}}\bar{\beta}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Issue

unbounded output $\not\Rightarrow$ classification.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Logistic Regression

This is classification.

Predicts discrete labels based on given features.

The Setup

$$y = \begin{matrix} 0 \\ 1 \end{matrix}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Setup

$x,$ $\quad y = \begin{matrix} 0 \\ 1 \end{matrix} \quad \left.\begin{matrix} 1-p \\ p \end{matrix}\right\}$ probability

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Setup

$x,$ $\quad y = \begin{matrix} 0 & 1-p \\ 1 & p \end{matrix} \Big\}$ probability

estim $\quad \boxed{P(y=1 \mid X)}$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definitions

Link function

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definitions

aws training and certification

Link function

relates mean of distribution to output of linear model

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definitions

aws training and certification

Link function

relates mean of distribution to output of linear model.

- unbounded ⇒ bounded
- continuous ⇒ discrete interpretation

© 2016, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definitions

aws training and certification

Link function

relates mean of distribution to output of linear model.

- → unbounded ⇒ bounded
- → continuous ⇒ discrete interpretation

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definitions

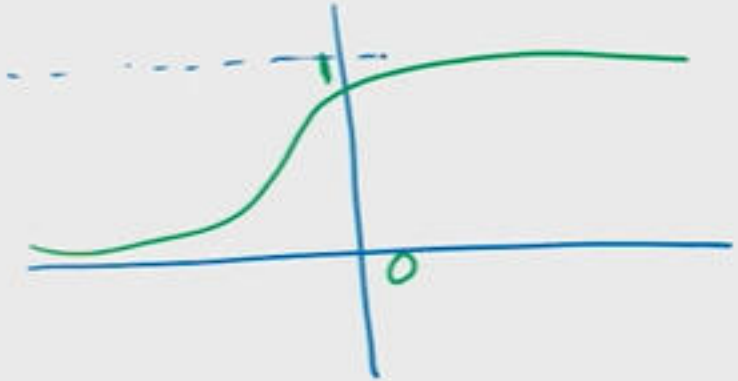
Logistic function

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definitions

aws training and certification

Logistic function



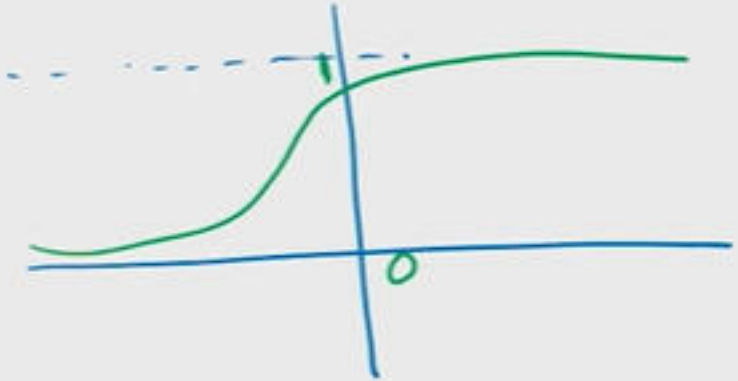
-7:34

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definitions

aws training and certification

Logistic function

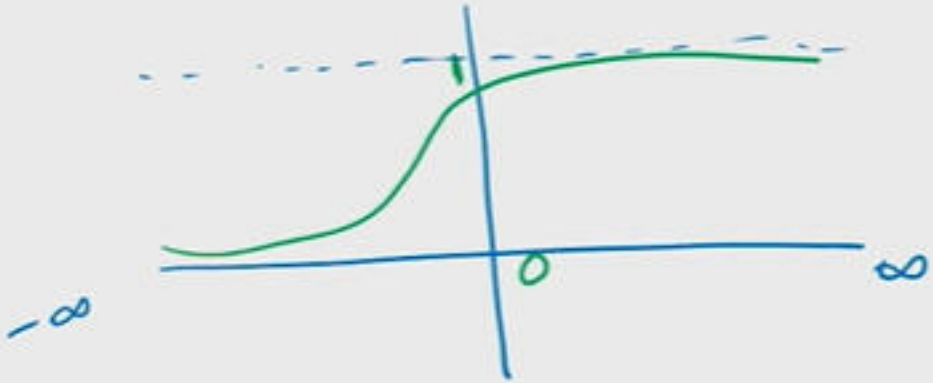


-7:34

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definitions

Logistic function



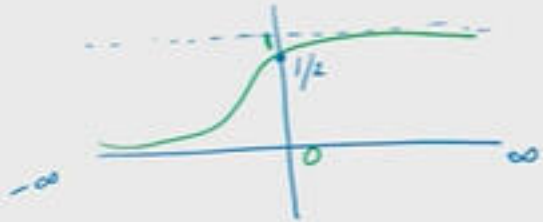
$$f : (-\infty, \infty) \to (0, 1)$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definitions

aws training and certification

Logistic function



$f(0) = 1/2$

$f(-\infty) = 0$

$f(\infty) = 1$

① $f : (-\infty, \infty) \to (0, 1)$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

-6:44

Definitions

Logistic function

$f(0) = 1/2$

$f(-\infty) = 0$

$f(\infty) = 1$

① $f : (-\infty, \infty) \to (0,1)$

$\mathbb{P}(Y=1 \mid X)$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Note on Choice of Link

aws training and certification

① bounded [0,1]

②

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Note on Choice of Link

aws training and certification

① bounded $[0,1]$

② domain $(-\infty, \infty)$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Note on Choice of Link

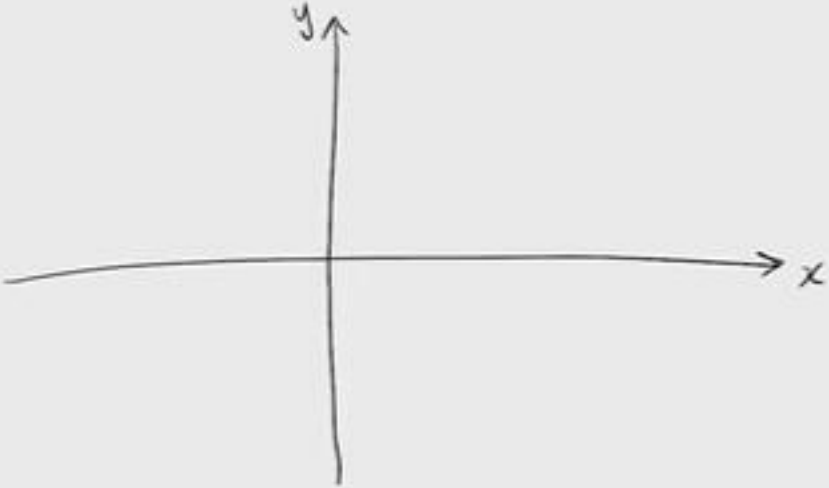
aws training and certification

1. bounded $[0,1]$
2. domain $(-\infty, \infty)$
3. differentiable everywhere.
4. increasing function.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Visualizing Logistic Regression

aws training and certification

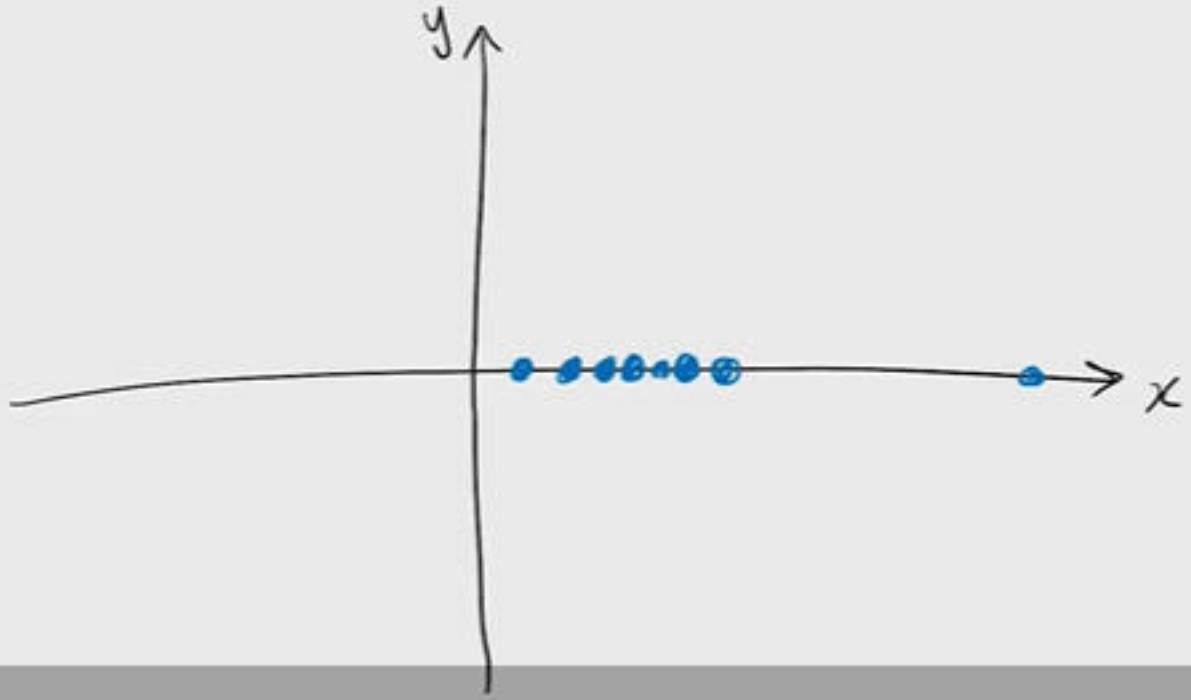


$y \in \{0, 1\}$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Visualizing Logistic Regression

aws training and certification

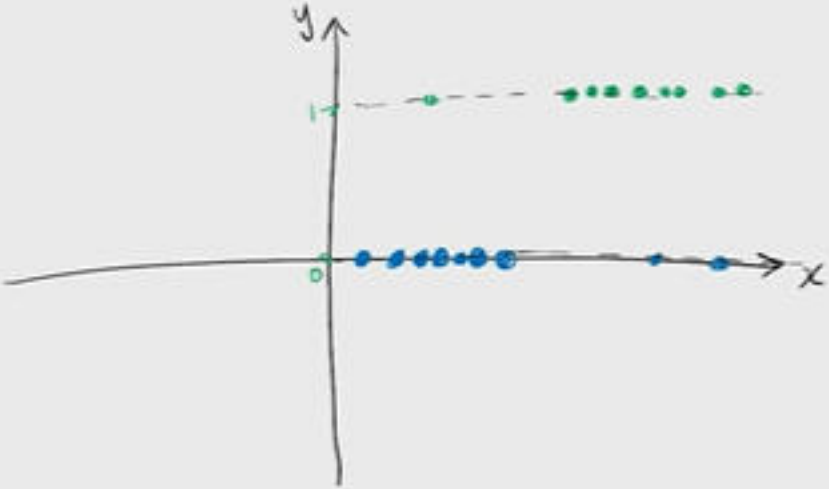


$y \in \{0, 1\}$

-3:10

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

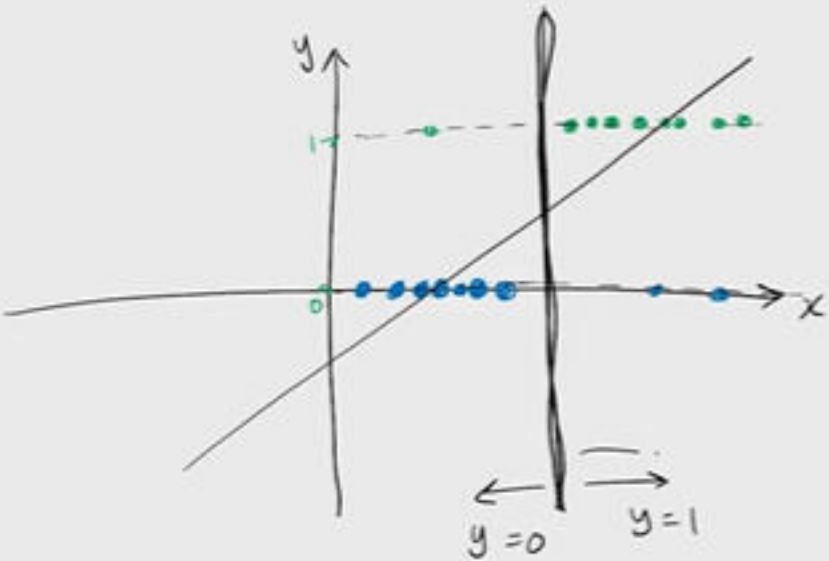
Visualizing Logistic Regression



$y \in \{0, 1$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

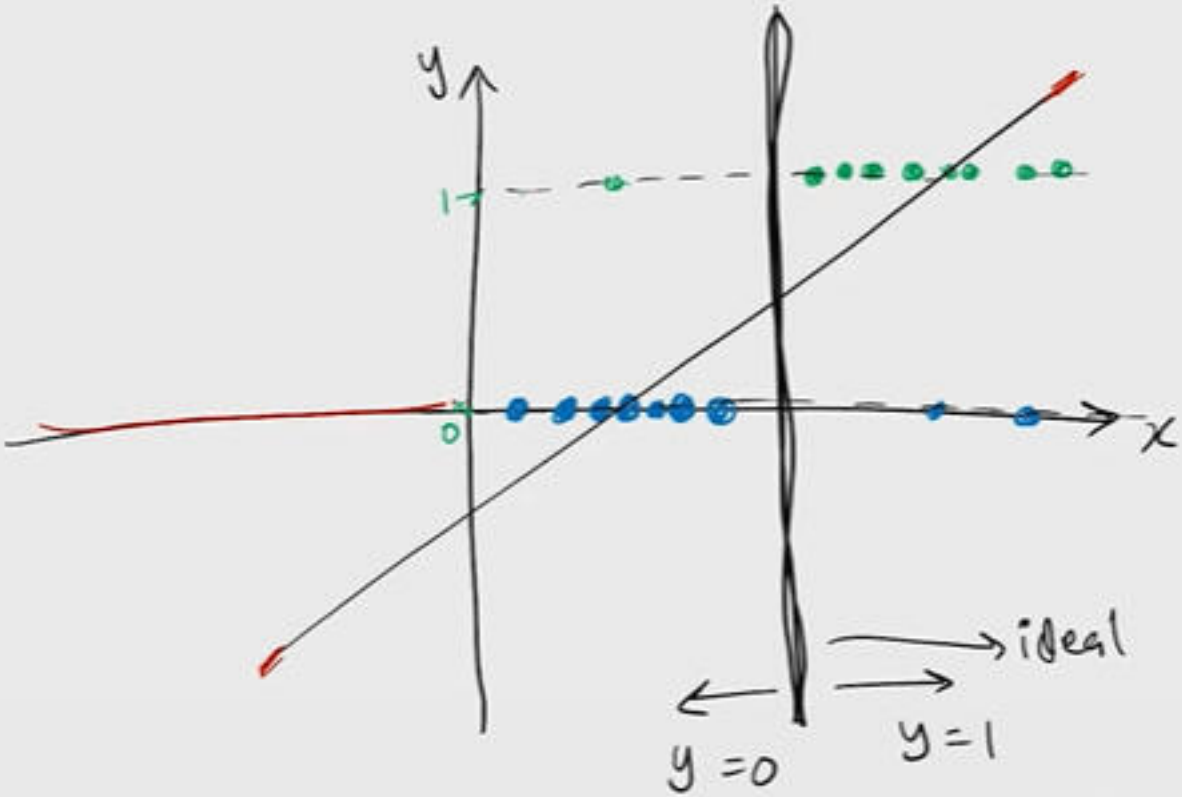
Visualizing Logistic Regression



$y \in \{0, 1$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Visualizing Logistic Regression

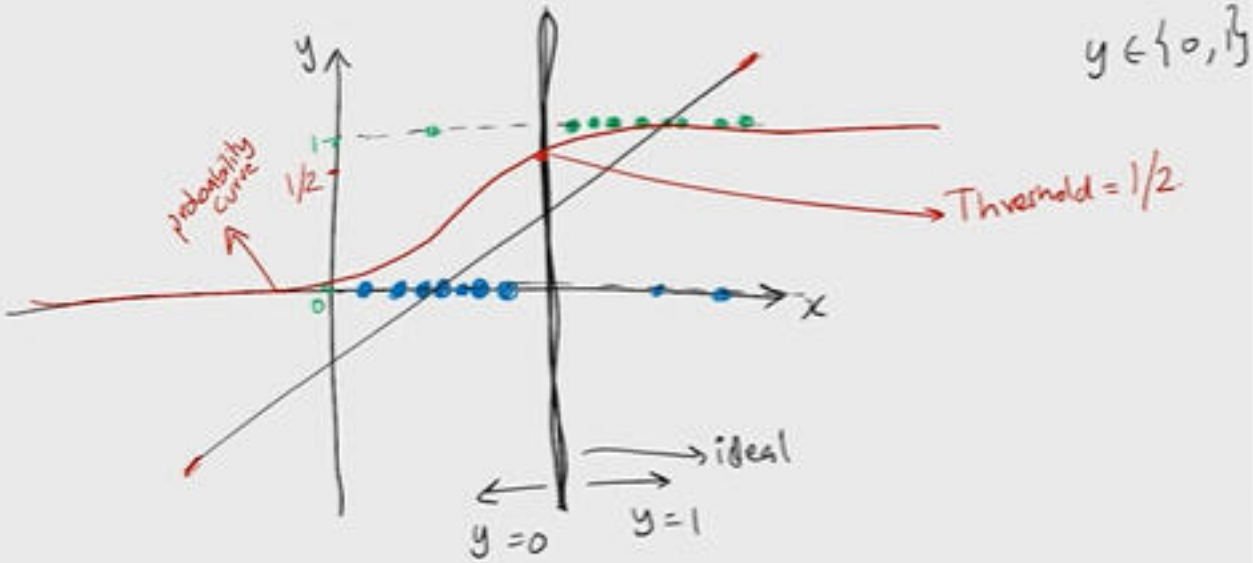


$y \in \{0, 1\}$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Visualizing Logistic Regression

aws training and certification



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

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Logistic Regression Loss

In this video we define the loss function for logistic regression and find its full log-likelihood estimation that we can minimize. However, unlike the OLS model, the logistic loss function doesn't have explicit solutions, so we'll need to minimize it numerically. Take a moment to explore before watching the video.

To find the logistic regression loss function:

1. Start with the logistic regression problem.
2. Find the full log likelihood.
3. Minimize -log(L).

Our Goal

$$\underbrace{P(y=1|X)}_{\text{probability to be estimated}}$$

$$\phi(x) = \frac{1}{1+e^{-x}}$$

$$\bar{y} = X\bar{\beta}$$

$$\boxed{\bar{y} = \Phi(X\bar{\beta})}$$

$$\Phi = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_N \end{bmatrix}$$

$$y_i = \phi(\bar{x}_i \cdot \bar{\beta})$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Our Goal

$$\underbrace{P(y=1|X)}_{\text{probability to be estimated}}$$

$$\phi(x) = \frac{1}{1+e^{-x}}$$

$$\bar{y} = X\bar{\beta}$$

$$\boxed{\bar{y} = \Phi(X\bar{\beta})}$$

$$\Phi = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_N \end{bmatrix}$$

$$y_i = \phi(\bar{x}_i \cdot \bar{\beta})$$

MLE: $\mathcal{L} = \prod_{i=1}^{N} P(Y=y_i | \bar{x}_i$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Our Goal

$$\underbrace{P(y=1|X)}_{\text{probability to be estimated}}$$

$$\phi(x) = \frac{1}{1+e^{-x}}$$

$$\bar{y} = X\bar{\beta}$$

$$\boxed{\bar{y} = \Phi(X\bar{\beta})}$$

$$\Phi = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_N \end{bmatrix}$$

$$y_i = \phi(\bar{x}_i \cdot \bar{\beta})$$

MLE: $\mathcal{L} = \prod_{i=1}^{N} P(Y=y_i | \bar{x}_i$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Our Goal

$Y = \begin{cases} 0 \\ 1 \end{cases}$ $\quad \underbrace{P(y=1|X)}_{\text{probability to be estimated}}$

Bernoulli $\rightarrow p$

$$\phi(x) = \frac{1}{1+e^{-x}}$$

$$\underbrace{\bar{y} = X\bar{\beta}}$$

$$\boxed{\bar{y} = \Phi(X\bar{\beta})}$$

$$\Phi = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_N \end{bmatrix}$$

$$y_i = \phi(\bar{x}_i \cdot \bar{\beta})$$

MLE;

$$\mathcal{L} = \prod_{i=1}^{N} P(Y=y_i|\bar{x}_i) = \prod_{i=1}^{N} (p)^{y_i}(1-p)^{1-y_i}$$

$\leftarrow$ independent (joint $\rightarrow$ product)

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Our Goal

aws training and certification

$Y = \begin{cases} 0 \\ 1 \end{cases}$ $\quad \underbrace{P(y=1|X)}_{\text{probability to be estimated}}$

Bernoulli $\rightarrow p$

$$\phi(x) = \frac{1}{1+e^{-x}}$$

$$\bar{y} = X\bar{\beta}$$

$$\boxed{\bar{y} = \Phi(X\bar{\beta})}$$

$$\Phi = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_N \end{bmatrix}$$

$$y_i = \phi(\bar{x}_i \cdot \bar{\beta})$$

MLE ;

$$L = \prod_{i=1}^{N} P(Y=y_i \mid \bar{x}_i) = \prod_{i=1}^{N} (p)^{y_i}(1-p)^{1-y_i}$$

$\leftarrow$ independent (joint $\rightarrow$ product)

log-likelihood, $\log(L) =$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

-3:29

Our Goal

$Y = \{0, 1\}$ $\quad \underbrace{P(y=1|X)}_{\text{probability to be estimated}}$

Bernoulli $\rightarrow p$

$$\phi(x) = \frac{1}{1+e^{-x}}$$

$$\bar{y} = X\bar{\beta}$$

$$\boxed{\bar{y} = \Phi(X\bar{\beta})}$$

$$\Phi = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_N \end{bmatrix}$$

$$y_i = \phi(\bar{x}_i \cdot \bar{\beta})$$

MLE;

$$\mathcal{L} = \prod_{i=1}^{N} P(Y=y_i | \bar{x}_i) \leftarrow \text{independent (joint} \rightarrow \text{product)}$$

$$= \prod_{i=1}^{N} (p)^{y_i} (1-p)^{1-y_i}$$

log-likelihood, $\log(L) = \sum_{i=1}^{N} \left[y_i \log p + (1-y_i) \log(1-p) \right]$

$=$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Our Goal

$Y = \{0, 1\}$

Bernoulli $\rightarrow p$

$$\underbrace{P(y=1|X)}_{\text{probability to be estimated}}$$

$$\phi(x) = \frac{1}{1+e^{-x}}$$

$$\bar{y} = X\bar{\beta}$$

$$\boxed{\bar{y} = \Phi(X\bar{\beta})}$$

$$\Phi = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_N \end{bmatrix}$$

$$\hat{y}_i = \phi(\bar{x}_i \cdot \bar{\beta})$$

$\hookrightarrow P(y_i | \bar{x}_i)$

MLE;

$$\mathcal{L} = \prod_{i=1}^{N} \underbrace{P(Y=y_i | \bar{x}_i)}_{(P)^{y_i}(1-p)^{1-y_i}} \leftarrow \text{independent (joint} \rightarrow \text{product)}$$

$$= \prod_{i=1}^{N} (P)^{y_i}(1-p)^{1-y_i}$$

log-likelihood, $\log(L) = \sum_{i=1}^{N} \left[y_i \log p + (1-y_i) \log(1-p) \right]$

$=$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Our Goal

$Y = \{0, 1\}$ $\quad \underbrace{P(y=1|X)}_{\text{probability to be estimated}}$

Bernoulli $\rightarrow p$

$$\phi(x) = \frac{1}{1+e^{-x}}$$

$$\bar{y} = X\bar{\beta}$$

$$\boxed{\bar{y} = \Phi(X\bar{\beta})}$$

$$\Phi = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_N \end{bmatrix}$$

$$\hat{y}_i = \phi(\bar{x}_i \cdot \bar{\beta})$$

$\hookrightarrow P(y_i | \bar{x}_i)$

MLE;

$$\mathcal{L} = \prod_{i=1}^{N} \underbrace{P(Y=y_i | \bar{x}_i)}_{(P)^{y_i}(1-p)^{1-y_i}} \leftarrow \text{independent (joint} \rightarrow \text{product)}$$

$$= \prod_{i=1}^{N} (P)^{y_i}(1-p)^{1-y_i}$$

log-likelihood, $\log(L) = \sum_{i=1}^{N} \left[y_i \log p + (1-y_i) \log(1-p) \right]$

$$= \sum_{i=1}^{N} \left[y_i \log \phi(\bar{x}_i \cdot \bar{\beta}) + (1-y_i) \times \log(1-\phi(\bar{x}_i \cdot \bar{\beta})) \right]$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Our Goal

$Y = \{0, 1\}$ Bernoulli $\rightarrow p$

$$\underbrace{P(y=1 \mid X)}_{\text{probability to be estimated}}$$

$$\phi(x) = \frac{1}{1+e^{-x}}$$

$$\bar{y} = X\bar{\beta}$$

$$\boxed{\bar{y} = \Phi(X\bar{\beta})}$$

$$\Phi = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_N \end{bmatrix}$$

$$\hat{y}_i = \phi(\bar{x}_i \cdot \bar{\beta})$$

$\hookrightarrow P(y_i \mid \bar{x}_i)$

MLE:

$$\mathcal{L} = \prod_{i=1}^{N} P(Y=y_i \mid \bar{x}_i) \leftarrow \text{independent (joint} \rightarrow \text{product)}$$

$$= \prod_{i=1}^{N} (p)^{y_i}(1-p)^{1-y_i}$$

log-likelihood, $\log(L) = \sum_{i=1}^{N} \left[y_i \log p + (1-y_i)\log(1-p) \right]$

$$= \sum_{i=1}^{N} \left[y_i \log \phi(\bar{x}_i \cdot \bar{\beta}) + (1-y_i) \times \log(1-\phi(\bar{x}_i \cdot \bar{\beta})) \right]$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Our Goal

aws training and certification

$Y = \{0, 1\}$ Bernoulli $\rightarrow p$

$\underbrace{P(y=1 \mid X)}_{\text{probability to be estimated}}$

$$\phi(x) = \frac{1}{1+e^{-x}}$$

$$\bar{y} = X\bar{\beta}$$

$$\boxed{\bar{y} = \Phi(X\bar{\beta})}$$

$$\Phi = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_N \end{bmatrix}$$

$$\hat{y}_i = \phi(\bar{x}_i \cdot \bar{\beta})$$

$\hookrightarrow P(y_i \mid \bar{x}_i)$

MLE:

$$\mathcal{L} = \prod_{i=1}^{N} P(Y=y_i \mid \bar{x}_i) \leftarrow \text{independent (joint} \rightarrow \text{product)}$$

$$= \prod_{i=1}^{N} (p)^{y_i}(1-p)^{1-y_i}$$

log-likelihood, $\log(\mathcal{L}) = \sum_{i=1}^{N} \left[y_i \log p + (1-y_i)\log(1-p) \right]$

$$= \sum_{i=1}^{N} \left[y_i \log \phi(\bar{x}_i \cdot \bar{\beta}) + (1-y_i) \times \log(1-\phi(\bar{x}_i \cdot \bar{\beta})) \right]$$

neg. log likelihood : $\boxed{-\log(\mathcal{L})}$

-1:14

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Our Goal

$Y = \{0, 1\}$ $\quad \underbrace{P(y=1|X)}_{\text{probability to be estimated}}$

Bernoulli $\rightarrow p$

$$\phi(x) = \frac{1}{1+e^{-x}}$$

$$\bar{y} = X\bar{\beta}$$

$$\boxed{\bar{y} = \Phi(X\bar{\beta})}$$

$$\Phi = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_N \end{bmatrix}$$

$$\hat{y}_i = \phi(\bar{x}_i \cdot \bar{\beta})$$

$\hookrightarrow P(y_i | \bar{x}_i)$

MLE:

$$\mathcal{L} = \prod_{i=1}^{N} P(Y=y_i | \bar{x}_i) \leftarrow \text{independent (joint} \rightarrow \text{product)}$$

$$= \prod_{i=1}^{N} (p)^{y_i}(1-p)^{1-y_i}$$

log-likelihood, $\log(\mathcal{L}) = \sum_{i=1}^{N} \left[y_i \log p + (1-y_i)\log(1-p) \right]$

$$= \sum_{i=1}^{N} \left[y_i \log \phi(\bar{x}_i \cdot \bar{\beta}) + (1-y_i) \times \log(1-\phi(\bar{x}_i \cdot \bar{\beta})) \right]$$

minimize

neg. log likelihood : $\boxed{-\log(\mathcal{L})}$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Full Log-Likelihood

$$\mathcal{L}(\vec{\beta}) = -\sum_i \left[y_i \log(\phi(\vec{x_i} \cdot \vec{\beta})) + (1 - y_i)\log(1 - \phi(\vec{x_i} \cdot \vec{\beta}))\right]$$

minimize
in
Logistic

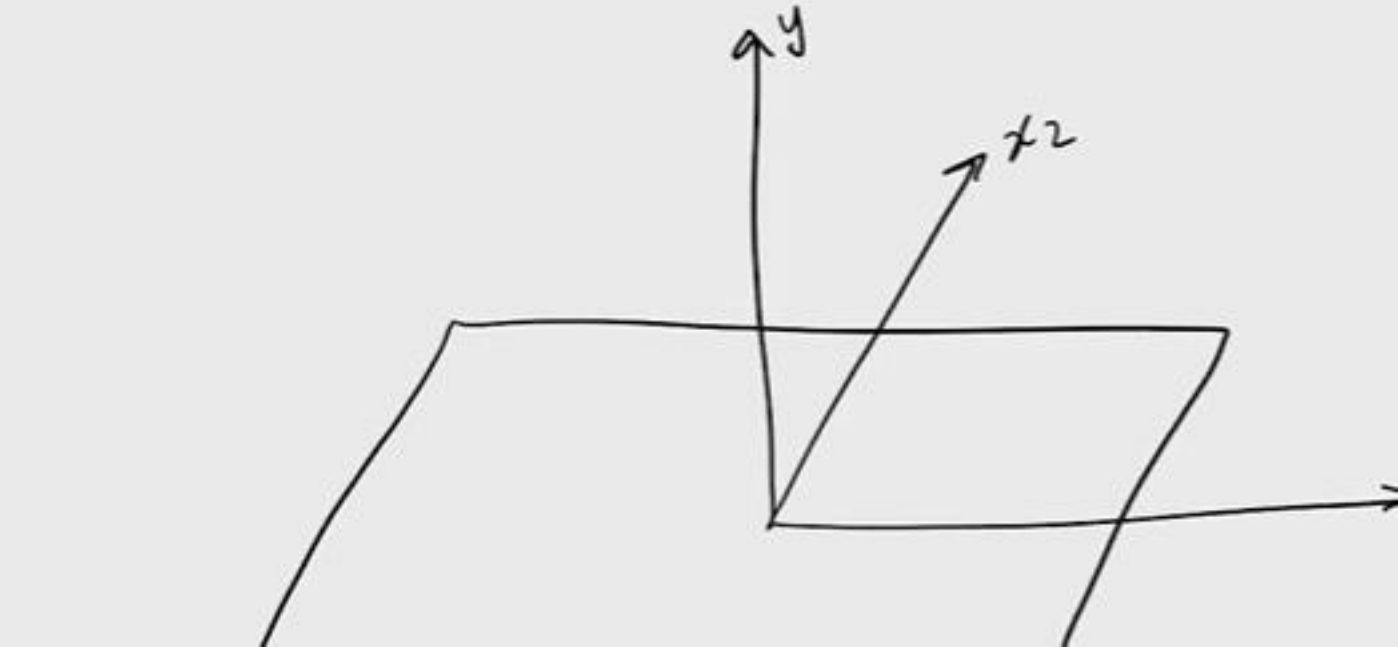
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

A Note on Minimization

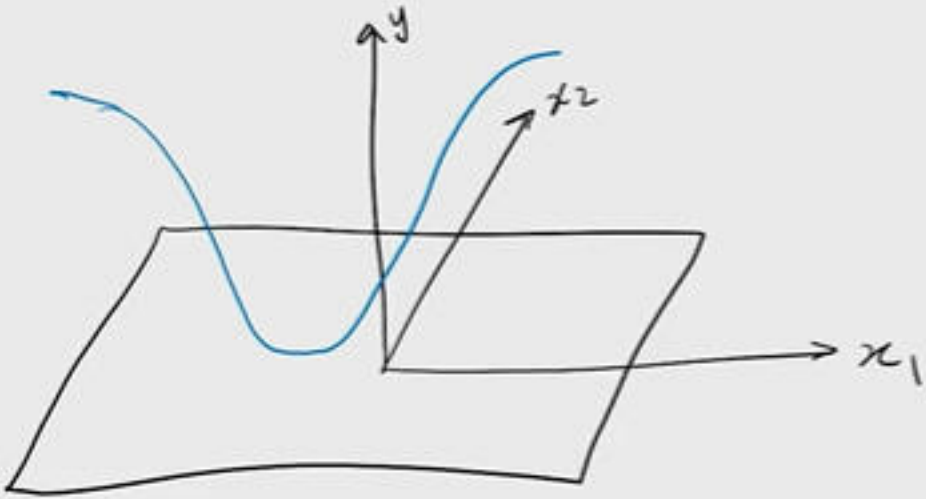
This does not have an explicit solution,
so we need to
minimize numerically

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Gradient



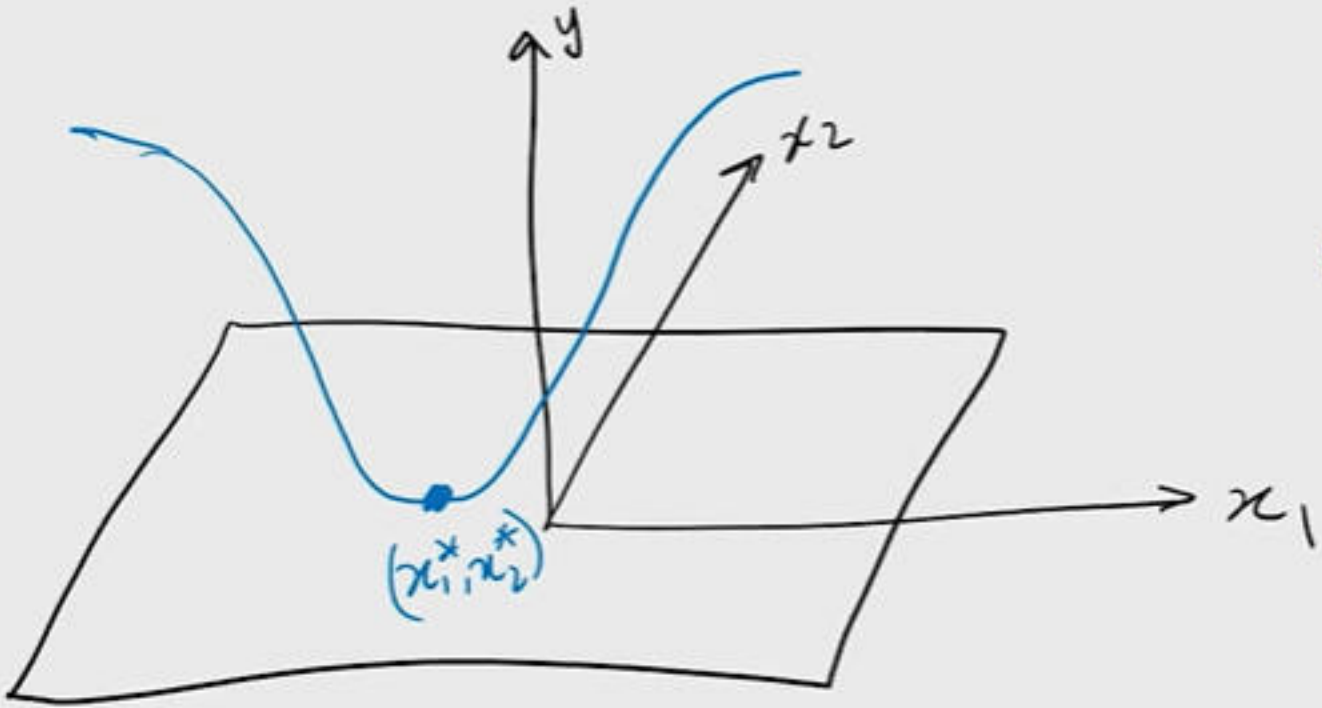
The Gradient



$y = f(x_1, x_2)$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Gradient



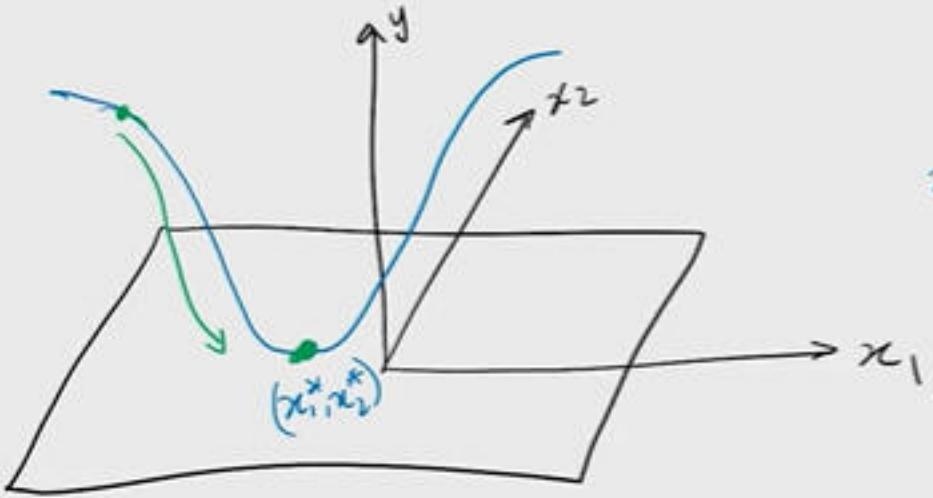
$y = f(x_1, x_2)$

$x_1^*, x_2^* = \arg\min$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Gradient

aws training and certification

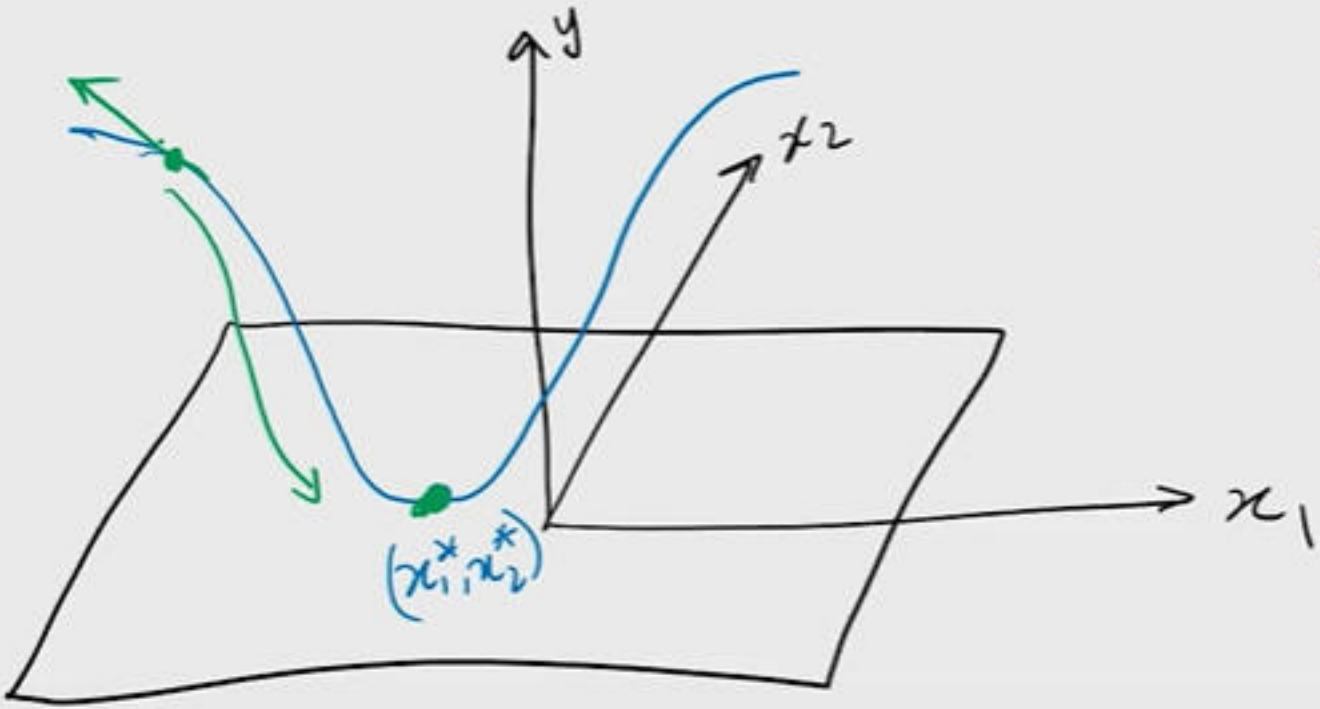


$y = f(x_1, x_2)$

$x_1^*, x_2^* = \arg\min_{x_1, x_2} f(x_1, x_2)$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Gradient

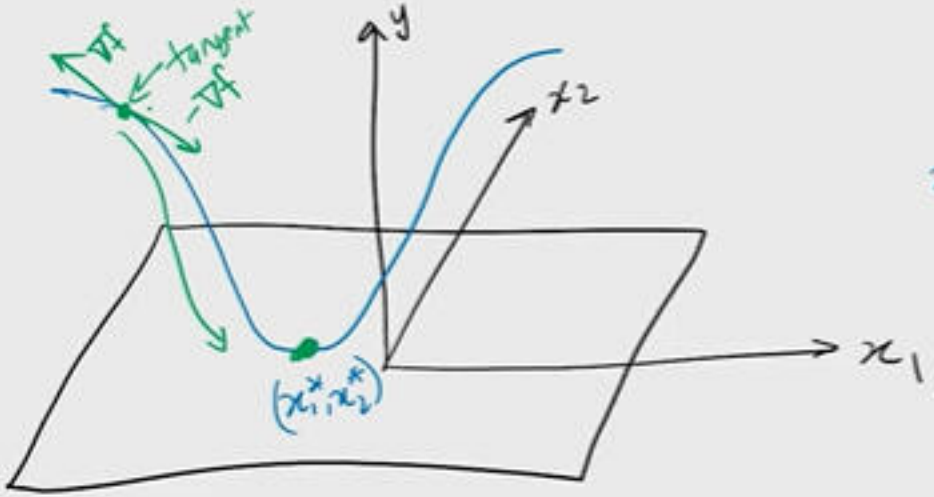


$$y = f(x_1, x_2)$$

$$x_1^*, x_2^* = \arg\min_{x_1, x_2} f(x_1, x_2$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Gradient

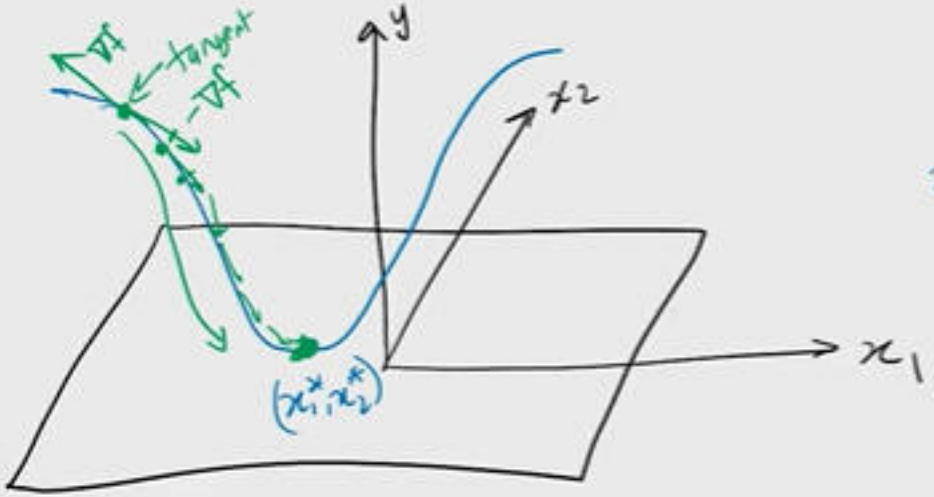


$y = f(x_1, x_2)$

$x_1^*, x_2^* = \arg\min_{x_1, x_2} f(x_1, x_2)$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Gradient



$y = f(x_1, x_2)$

$x_1^*, x_2^* = \arg\min_{x_1, x_2} f(x_1, x_2)$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Algorithm

① Random choice of $\bar{x}$

② Compute gradient of f at $\bar{x}$ $\nabla f(\bar{x})$

③ Step in the direction of the negative of the gradient.

x:

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Algorithm

1. Random choice of $\bar{x}$
2. Compute gradient of f at $\bar{x}$ $\nabla f(\bar{x})$
3. Step in the direction of the negative of the gradient.

$$\bar{x} \leftarrow \bar{x} - \underbrace{\eta}_{\text{step size}} \nabla f(\bar{x})$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Algorithm

aws training and certification

① Random choice of $\bar{x}$

② Compute gradient of f at $\bar{x}$ $\nabla f(\bar{x})$

③ Step in the direction of the negative of the gradient.

$$\bar{x} \leftarrow \bar{x} - \eta \nabla f(\bar{x})$$

η → step size → how fast do you want to minimize?

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Numerical Minimization of OLS

We work with the **mean log-likelihood**.

!

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Numerical Minimization of OLS

We work with the **mean log-likelihood**.

$$L = -\frac{1}{N} \|\bar{y} - X\bar{\beta}\|_2^2$$

N - observations.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Numerical Minimization of OLS

We work with the **mean log-likelihood**.

$$L = -\frac{1}{N}\|\bar{y} - X\bar{\beta}\|_2^2$$

N - observations.

$$= -\frac{1}{N}(\bar{y} - X\bar{\beta})^T(\bar{y} - X\bar{\beta})$$

$$\nabla L = -\frac{2}{N}\sum_{i=1}^{1} y_i - \bar{x}_i \cdot$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Numerical Minimization of OLS

We work with the **mean log-likelihood**.

$$L = -\frac{1}{N}\|\bar{y} - X\bar{\beta}\|_2^2$$

N – observations.

$$= -\frac{1}{N}(\bar{y} - X\bar{\beta})^T(\bar{y} - X\bar{\beta})$$

$$\nabla L = -\frac{2}{N}\sum_{i=1}^{N}(y_i - \bar{x}_i \cdot \bar{\beta})(\bar{x}_i)$$

$$= -\frac{2}{N}\underbrace{X}_{\text{data matrix}}^T(\bar{y} - X\bar{\beta})$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Numerical Minimization of OLS

We work with the **mean log-likelihood**.

$$L = -\frac{1}{N}\|\bar{y} - X\bar{\beta}\|_2^2$$

N – observations.

$$= -\frac{1}{N}(\bar{y} - X\bar{\beta})^T(\bar{y} - X\bar{\beta})$$

$$\nabla L = -\frac{2}{N}\sum_{i=1}^{N}(y_i - \bar{x}_i \cdot \bar{\beta})(\bar{x}_i)$$

$$= -\frac{2}{N}\underbrace{X}_{\text{data matrix}}^T(\bar{y} - X\bar{\beta})$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Algorithm

The algorithm runs as follows:

- Pick a learning rate η
- Initialize $\vec{\beta}$ with a random guess
- Iterate

$$\vec{\beta} \leftarrow \vec{\beta} - \eta \frac{\partial \mathcal{L}}{\partial \vec{\beta}}$$

$$= \vec{\beta} + \frac{2\eta}{N} X^T [\vec{y} - X\vec{\beta}]$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Gradient Descent for Logistic Regression

aws training and certification

Gradient Descent for Logistic Regression

-8:23

Numerical Minimization of Logistic Loss

aws training and certification

$$\mathcal{L}(\vec{\beta}) = -\sum_i \left[\underbrace{y_i \log(\phi(\vec{x}_i \cdot \vec{\beta}))} + \underbrace{(1 - y_i) \log(1 - \phi(\vec{x}_i \vec{\beta}))} \right]$$

$$\phi(x) = \frac{1}{1+e^{-}}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Numerical Minimization of Logistic Loss

$$\mathcal{L}(\vec{\beta}) = -\sum_i \left[\underbrace{y_i \log(\phi(\vec{x}_i \cdot \vec{\beta}))} + \underbrace{(1 - y_i) \log(1 - \phi(\vec{x}_i \vec{\beta}))} \right]$$

$$\phi(x) = \frac{1}{1+e^{-x}}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Numerical Minimization of Logistic Loss

$$\mathcal{L}(\vec{\beta}) = -\sum_i \left[y_i \log(\phi(\vec{x}_i \cdot \vec{\beta})) + (1 - y_i)\log(1 - \phi(\vec{x}_i \vec{\beta})) \right]$$

$$\phi(x) = \frac{1}{1+e^{-x}}$$

$$\phi(x) + \phi(-x) = 1 \quad \forall x.$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Hard Term

$$\frac{\partial}{\partial \beta_{j0}}[\log(\phi\,(\vec{x}_i \cdot \overrightarrow{\beta}))] =$$

$$\frac{\partial}{\partial \overrightarrow{\beta}}[\log(\phi\,(\vec{x}_i \overrightarrow{\beta}))] = \vec{x}_i^T(1 - \phi\,(\vec{x}_i \cdot \overrightarrow{\beta}))$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Hard Term

$$\frac{\partial}{\partial \beta_{j0}}[\log(\phi(\vec{x}_i \cdot \overrightarrow{\beta}))] = \frac{1}{\phi(\bar{x}_i \cdot \bar{\beta})} \frac{\partial}{\partial \beta_{j0}} \phi(\bar{x}_i \cdot \bar{\beta})$$

$$= \left(\frac{1+e^{-(\bar{x}_i \cdot \bar{\beta})}}{1}\right) \frac{\partial}{\partial \beta_{j0}} \left(\frac{1}{1+e^{-\bar{x}_i \cdot \bar{\beta}}}\right)$$

$$= \left(1+e^{-(\bar{x}_i \cdot \bar{\beta})}\right) \frac{1}{\left(1+e^{-\bar{x}_i \cdot \bar{\beta}}\right)^2} e^{-\bar{x}_i \cdot \bar{\beta}} \left(-x_{ij0}\right)$$

$$\frac{\partial}{\partial \vec{\beta}}[\log(\phi(\vec{x}_i \overrightarrow{\beta}))] = \vec{x}_i^T(1 - \phi(\vec{x}_i \cdot \overrightarrow{\beta}))$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Hard Term

$$\frac{\partial}{\partial \beta_{j0}}[\log(\phi(\vec{x}_i \cdot \overrightarrow{\beta}))] = \frac{1}{\phi(\bar{x}_i \cdot \bar{\beta})} \frac{\partial}{\partial \beta_{j0}} \phi(\bar{x}_i \cdot \bar{\beta})$$

$$= \left(\frac{1+e^{-(\bar{x}_i \cdot \bar{\beta})}}{1}\right) \frac{\partial}{\partial \beta_{j0}} \left(\frac{1}{1+e^{-\bar{x}_i \cdot \bar{\beta}}}\right)$$

$$= \left(1+e^{-(\bar{x}_i \cdot \bar{\beta})}\right) \frac{1}{\left(1+e^{-\bar{x}_i \cdot \bar{\beta}}\right)^2} e^{-\bar{x}_i \cdot \bar{\beta}} \left(-x_{ij0}\right)$$

$$\frac{\partial}{\partial \vec{\beta}}[\log(\phi(\vec{x}_i \overrightarrow{\beta}))] = \vec{x}_i^T(1 - \phi(\vec{x}_i \cdot \overrightarrow{\beta}))$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Hard Term

$$\frac{\partial}{\partial \beta_{j0}}[\log(\phi(\vec{x}_i \cdot \overrightarrow{\beta}))] = \frac{1}{\phi(\bar{x}_i \cdot \bar{\beta})} \frac{\partial}{\partial \beta_{j0}} \phi(\bar{x}_i \cdot \bar{\beta})$$

$$= \left(\frac{1+e^{-(\bar{x}_i \cdot \bar{\beta})}}{1}\right) \frac{\partial}{\partial \beta_{j0}} \left(\frac{1}{1+e^{-\bar{x}_i \cdot \bar{\beta}}}\right)$$

$$= \cancel{\left(1+e^{-(\bar{x}_i \cdot \bar{\beta})}\right)} \frac{1}{\left(1+e^{-\bar{x}_i \cdot \bar{\beta}}\right)^{\cancel{2}}} e^{-\bar{x}_i \cdot \bar{\beta}} (-x_{ij0})$$

$$= \frac{1}{1+e^{+\cdot}}$$

$$\frac{\partial}{\partial \vec{\beta}}[\log(\phi(\vec{x}_i \overrightarrow{\beta}))] = \vec{x}_i^T(1-\phi(\vec{x}_i \cdot \overrightarrow{\beta}))$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Hard Term

$$\frac{\partial}{\partial \beta_{j0}}[\log(\phi(\vec{x}_i \cdot \overrightarrow{\beta}))] = \frac{1}{\phi(\bar{x}_i \cdot \bar{\beta})} \frac{\partial}{\partial \beta_{j0}} \phi(\bar{x}_i \cdot \bar{\beta})$$

$$= \left(\frac{1+e^{-(\bar{x}_i \cdot \bar{\beta})}}{1}\right) \frac{\partial}{\partial \beta_{j0}} \left(\frac{1}{1+e^{-\bar{x}_i \cdot \bar{\beta}}}\right)$$

$$= \cancel{\left(1+e^{-(\bar{x}_i \cdot \bar{\beta})}\right)} \frac{1}{\left(1+e^{-\bar{x}_i \cdot \bar{\beta}}\right)^{\cancel{2}}} e^{-\bar{x}_i \cdot \bar{\beta}} (-x_{ij0})$$

$$= \frac{1}{1+e^{+\cdot}}$$

$$\frac{\partial}{\partial \vec{\beta}}[\log(\phi(\vec{x}_i \overrightarrow{\beta}))] = \vec{x}_i^T(1-\phi(\vec{x}_i \cdot \overrightarrow{\beta}))$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Hard Term

$$\frac{\partial}{\partial \beta_{jo}}[\log(\phi(\vec{x}_i \cdot \overrightarrow{\beta}))] = \frac{1}{\phi(\bar{x}_i \cdot \bar{\beta})} \frac{\partial}{\partial \beta_{jo}} \phi(\bar{x}_i \cdot \bar{\beta})$$

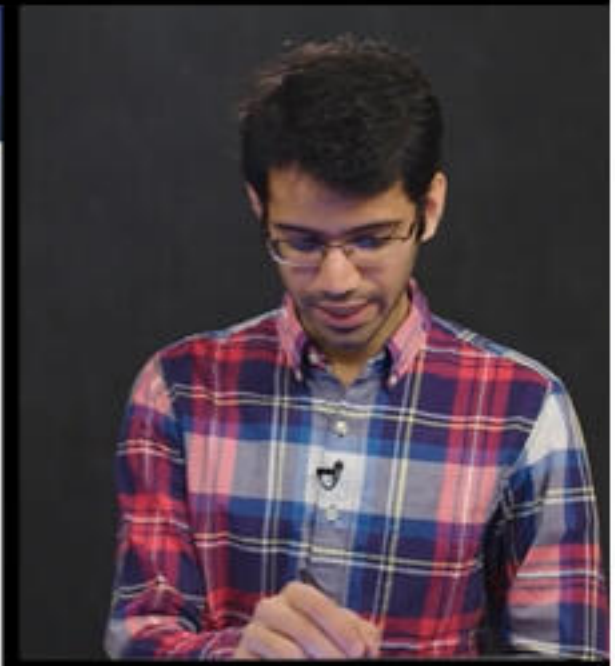
$$= \left(\frac{1+e^{-\bar{x}_i \cdot \bar{\beta}}}{1}\right) \frac{\partial}{\partial \beta_{jo}} \left(\frac{1}{1+e^{-\bar{x}_i \cdot \bar{\beta}}}\right)$$

$$= \cancel{\left(1+e^{-\bar{x}_i \cdot \bar{\beta}}\right)} \frac{1}{\left(1+e^{-\bar{x}_i \cdot \bar{\beta}}\right)^{\cancel{2}}} e^{-\bar{x}_i \cdot \bar{\beta}} \left(-x_{ij_0}\right)$$

$$= \frac{1}{1+e^{+\bar{x}_i \cdot \bar{\beta}}} \left(x_{ij_0}\right)$$

$$\frac{\partial}{\partial \overrightarrow{\beta}}[\log(\phi(\vec{x}_i \overrightarrow{\beta}))] = \vec{x}_i^T(1 - \phi(\vec{x}_i \cdot \overrightarrow{\beta}))$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



Hard Term

$$\frac{\partial}{\partial \beta_{jo}}[\log(\phi(\vec{x}_i \cdot \overrightarrow{\beta}))] = \frac{1}{\phi(\bar{x}_i \cdot \bar{\beta})} \frac{\partial}{\partial \beta_{jo}} \phi(\bar{x}_i \cdot \bar{\beta})$$

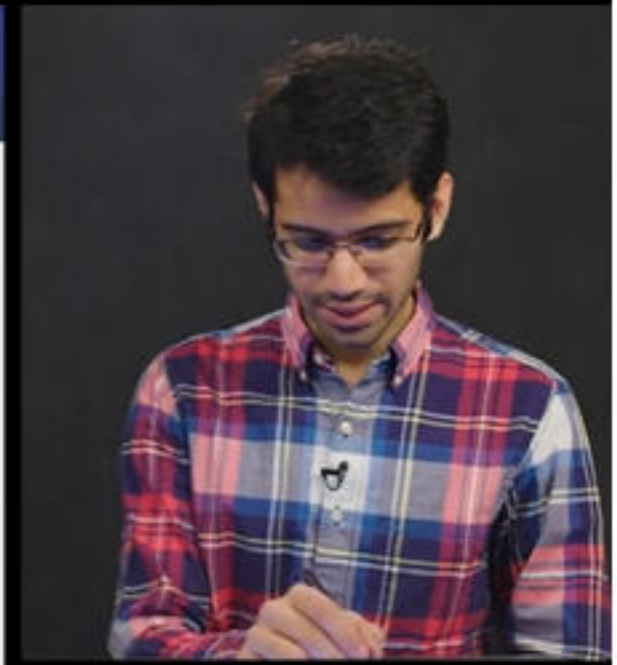
$$= \left(\frac{1+e^{-\bar{x}_i \cdot \bar{\beta}}}{1}\right) \frac{\partial}{\partial \beta_{jo}} \left(\frac{1}{1+e^{-\bar{x}_i \cdot \bar{\beta}}}\right)$$

$$= \cancel{\left(1+e^{-\bar{x}_i \cdot \bar{\beta}}\right)} \frac{1}{\left(1+e^{-\bar{x}_i \cdot \bar{\beta}}\right)^{\cancel{2}}} e^{-\bar{x}_i \cdot \bar{\beta}} \left(-x_{ij_0}\right)$$

$$= \frac{1}{1+e^{+\bar{x}_i \cdot \bar{\beta}}} \left(x_{ij_0}\right)$$

$$\frac{\partial}{\partial \overrightarrow{\beta}}[\log(\phi(\vec{x}_i \overrightarrow{\beta}))] = \vec{x}_i^T(1 - \phi(\vec{x}_i \cdot \overrightarrow{\beta}))$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



Hard Term

$$\frac{\partial}{\partial \beta_{j0}}[\log(\phi(\vec{x}_i \cdot \overrightarrow{\beta}))] = \frac{1}{\phi(\bar{x}_i \cdot \bar{\beta})} \frac{\partial}{\partial \beta_{j0}} \phi(\bar{x}_i \cdot \bar{\beta})$$

$$= \left(\frac{1+e^{-(\bar{x}_i \cdot \bar{\beta})}}{1}\right) \frac{\partial}{\partial \beta_{j0}} \left(\frac{1}{1+e^{-\bar{x}_i \cdot \bar{\beta}}}\right)$$

$$= \left(\cancel{1+e^{-(\bar{x}_i \cdot \bar{\beta})}}\right) \frac{1}{\left(1+e^{-\bar{x}_i \cdot \bar{\beta}}\right)^2} e^{-\bar{x}_i \cdot \bar{\beta}} \left(-x_{ij0}\right)$$

$$= \frac{1}{1+e^{+\bar{x}_i \cdot \bar{\beta}}} \left(x_{ij0}\right) = \phi(-\bar{x}_i \cdot \bar{\beta})\, x_{ij0} =$$

$$\frac{\partial}{\partial \vec{\beta}}[\log(\phi(\vec{x}_i \overrightarrow{\beta}))] = \vec{x}_i^T (1 - \phi(\vec{x}_i \cdot \overrightarrow{\beta}))$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Hard Term

$$\frac{\partial}{\partial \beta_{j0}}[\log(\phi(\vec{x}_i \cdot \overrightarrow{\beta}))] = \frac{1}{\phi(\bar{x}_i \cdot \bar{\beta})} \frac{\partial}{\partial \beta_{j0}} \phi(\bar{x}_i \cdot \bar{\beta})$$

$$= \left(\frac{1+e^{-(\bar{x}_i \cdot \bar{\beta})}}{1}\right) \frac{\partial}{\partial \beta_{j0}} \left(\frac{1}{1+e^{-\bar{x}_i \cdot \bar{\beta}}}\right)$$

$$= \left(\cancel{1+e^{-(\bar{x}_i \cdot \bar{\beta})}}\right) \frac{1}{\left(1+e^{-\bar{x}_i \cdot \bar{\beta}}\right)^{\cancel{2}}} e^{-\bar{x}_i \cdot \bar{\beta}} (-x_{ij0})$$

$$= \frac{1}{1+e^{+\bar{x}_i \cdot \bar{\beta}}} (x_{ij0}) = \phi(-\bar{x}_i \cdot \bar{\beta})\, x_{ij0} = (1-\phi(\bar{x}_i \cdot \bar{\beta}))\, x_{ij0}$$

$$\frac{\partial}{\partial \vec{\beta}}[\log(\phi(\vec{x}_i \overrightarrow{\beta}))] = \vec{x}_i^T (1 - \phi(\vec{x}_i \cdot \overrightarrow{\beta}))$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Hard Term 2

$$\frac{\partial}{\partial \beta}[\log(1 - \phi(\vec{x}_i \cdot \overrightarrow{\beta}))] = \frac{\partial}{\partial \beta} \log(\phi(-\vec{x}_i \cdot \vec{\beta}))$$

$$\frac{\partial}{\partial \vec{\beta}}[\log(1 - \phi(\vec{x}_i \overrightarrow{\beta}))] = -\vec{x}_i^T \phi(\vec{x}_i \cdot \overrightarrow{\beta})$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Hard Term 2

$$\frac{\partial}{\partial \beta}[\log(1-\phi(\vec{x}_i \cdot \overrightarrow{\beta}))] = \frac{\partial}{\partial \beta} \log\left(\phi(-\bar{x}_i \cdot \bar{\beta})\right)$$

$$= -x_i^T \left(1 - \phi(-\bar{x}_i \cdot \bar{\beta})\right)$$

$$=$$

$$\frac{\partial}{\partial \vec{\beta}}[\log(1-\phi(\vec{x}_i \overrightarrow{\beta}))] = -\vec{x}_i^T \phi(\vec{x}_i \cdot \overrightarrow{\beta})$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

All Terms

$$\mathcal{L}(\vec{\beta}) = -\sum_i \left[y_i \log(\phi(\vec{x}_i \cdot \overrightarrow{\beta})) + (1 - y_i)\log(1 - \phi(\vec{x}_i \cdot \overrightarrow{\beta}))\right]$$

$$\frac{\partial \mathcal{L}(\vec{\beta})}{\partial \vec{\beta}} = -\sum_i y_i x_i^T \left(1 - \phi(\vec{x}_i \cdot \vec{\beta})\right) + 1 -$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

All Terms

$$\mathcal{L}(\vec{\beta}) = -\sum_i \left[y_i \log(\phi(\vec{x}_i \cdot \overrightarrow{\beta})) + (1 - y_i)\log(1 - \phi(\vec{x}_i \cdot \overrightarrow{\beta}))\right]$$

$$\frac{\partial \mathcal{L}(\vec{\beta})}{\partial \vec{\beta}} = -\sum_i y_i x_i^T \left(1 - \phi(\vec{x}_i \cdot \vec{\beta})\right) + (1 - y_i)(-x_i^T)\, \phi(\vec{x}_i \cdot \vec{\beta})$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

All Terms

$$\mathcal{L}(\vec{\beta}) = -\sum_i \left[y_i \log(\phi(\vec{x}_i \cdot \vec{\beta})) + (1 - y_i)\log(1 - \phi(\vec{x}_i \cdot \vec{\beta}))\right]$$

$$\frac{\partial \mathcal{L}(\vec{\beta})}{\partial \vec{\beta}} = -\sum_i \left[y_i x_i^T \left(1 - \phi(\vec{x}_i \cdot \vec{\beta})\right) + (1 - y_i)(-x_i^T)\,\phi(\vec{x}_i \cdot \vec{\beta})\right]$$

$$= -\sum_i \left[y_i x_i^T - y_i x_i^T \phi(\vec{x}_i \cdot \vec{\beta}) + x_i^T y_i \phi(\vec{x}_i \cdot \vec{\beta}) - x_i^T \phi(\vec{x}_i \cdot \vec{\beta})\right]$$

-0:58

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

All Terms

$$\mathcal{L}(\vec{\beta}) = -\sum_i \left[y_i \log(\phi(\vec{x}_i \cdot \vec{\beta})) + (1 - y_i)\log(1 - \phi(\vec{x}_i \cdot \vec{\beta}))\right]$$

$$\frac{\partial \mathcal{L}(\vec{\beta})}{\partial \vec{\beta}} = -\sum_i \left[y_i x_i^T \left(1 - \phi(\vec{x}_i \cdot \vec{\beta})\right) + (1 - y_i)(-x_i^T)\,\phi(\vec{x}_i \cdot \vec{\beta})\right]$$

$$= -\sum_i \left[y_i x_i^T - \cancel{y_i x_i^T \phi(\vec{x}_i \cdot \vec{\beta})} + \cancel{x_i^T y_i \phi(\vec{x}_i \cdot \vec{\beta})} - x_i^T \phi(\vec{x}_i \cdot \vec{\beta})\right]$$

$$= -\sum_i x_i^T \left(y_i - \phi(\vec{x}_i \cdot \vec{\beta})\right)$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Topics

- Written Out Nicely
- A Note on Normalization
- The Algorithm

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Written Out Nicely

$$\frac{\partial \mathcal{L}}{\partial \vec{\beta}} = -X^T(\vec{y} - \Phi(X\vec{\beta}))$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Written Out Nicely

$$\frac{\partial \mathcal{L}}{\partial \vec{\beta}} = -X^T(\vec{y} - \phi(X\vec{\beta}))$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

A Note on Normalization

aws training and certification

To work with different sized datasets, it is often better to divide everything by the number of observations N to get a **mean loss**.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

A Note on Normalization

aws training and certification

To work with different sized datasets, it is often better to divide everything by the number of observations N to get a **mean loss**.

This constant is carried through all the derivatives, as well as giving the final algorithm.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Algorithm

The algorithm runs as follows. Pick a learning rate n.

- Initialize $\vec{\beta}$ with a random guess.
- Iterate

$$\vec{\beta} \leftarrow \vec{\beta} - \eta \frac{\partial \mathcal{L}}{\partial \vec{\beta}} = \vec{\beta} + \frac{\eta}{N} X^T [\vec{y} - \phi(X\vec{\beta})].$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Two Update Steps

$$\vec{\beta} \leftarrow \vec{\beta} + \underbrace{\frac{2\eta}{N} X^T}_{\text{OLS}} \underbrace{[\vec{y} - X\vec{\beta}]}_{\text{residual error}}$$

vs.

$$\vec{\beta} \leftarrow \vec{\beta} + \underbrace{\frac{\eta}{N} X^T}_{\text{Logistic}} [\underbrace{\vec{y} - \phi(X\vec{\beta})}_{\text{error}}]$$

$\in \{0,1$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Two Update Steps

$$\vec{\beta} \leftarrow \vec{\beta} + \underbrace{\frac{2\eta}{N}}_{} X^T \underbrace{[\vec{y} - X\vec{\beta}]}_{\text{residual error}}$$

OLS

vs.

$$\vec{\beta} \leftarrow \vec{\beta} + \underbrace{\frac{\eta}{N}}_{} X^T [\vec{y} - \phi(X\vec{\beta})]$$

Logistic

error

$\in \{0, 1$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Two Update Steps

$$\vec{\beta} \leftarrow \vec{\beta} + \underbrace{\frac{2\eta}{N} X^T}_{\text{OLS}} \underbrace{[\vec{y} - X\vec{\beta}]}_{\text{residual error}}$$

vs.

$$\vec{\beta} \leftarrow \vec{\beta} + \underbrace{\frac{\eta}{N} X^T}_{\text{Logistic}} [\underbrace{\vec{y} - \phi(X\vec{\beta})}_{\text{error}}]$$

$\in \{0, 1\}$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Regularization

aws training and certification

OLS :

Regularization, $\mathcal{L} = \|\bar{y} - X\bar{\beta}\|_2^2 + \delta^2 \|\beta\|_{2/1}^{2/1}$

Logistic :

Reg, $\mathcal{L} = \underbrace{\mathcal{L}(\beta)}_{\text{logistic loss function}} + \delta^2 \|\beta\|_{2/1}^{2/1}$

Ridge ← uniform penalization

LAS

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Softmax

$$P(Y = c_j \mid x_i) = \frac{e^{\bar{\beta}_j \cdot \bar{x}_i}}{\sum_{j=1}^{M} e^{\bar{\beta}_j \cdot \bar{x}_i}}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Softmax

$$P(Y = c_j \mid x_i) = \frac{e^{\bar{\beta}_j \cdot \bar{x}_i}}{\sum_{j=1}^{M} e^{\bar{\beta}_j \cdot \bar{x}_i}}$$

$y \in \{1, \dots, M\}$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Softmax

$$P(Y = c_j \mid x_i) = \frac{e^{\bar{\beta}_j \cdot \bar{x}_i}}{\sum_{j=1}^{M} e^{\bar{\beta}_j \cdot \bar{x}_i}}$$

$y \in \{1, \dots, M\}$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Softmax

$$P(Y = c_j \mid x_i) = \frac{e^{\bar{\beta}_j \cdot \bar{x}_i}}{\sum_{j=1}^{M} e^{\bar{\beta}_j \cdot \bar{x}_i}}$$

$y \in \{1, \dots, M\}$

S

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Comparison with Logistic in the Case of Two Types

$$P(Y=1 \mid \bar{x}_i) = \frac{1}{1+e^{-\bar{x}_i \cdot \bar{\beta}}}$$

$$P(Y=1 \mid \bar{x}_i) = \frac{e^{\bar{x}_i \cdot \bar{\beta}_1}}{e^{\bar{x}_i \cdot \bar{\beta}_1} + e^{\bar{x}_i \cdot \bar{\beta}}}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Comparison with Logistic in the Case of Two Types

$$P(Y=1 \mid \bar{x}_i) = \frac{1}{1+e^{-\bar{x}_i \cdot \bar{\beta}}}$$

$$P(Y=1 \mid \bar{x}_i) = \frac{e^{\bar{x}_i \cdot \bar{\beta}_1} \times e^{-\bar{x} \cdot \bar{\beta}_1}}{e^{\bar{x}_i \cdot \bar{\beta}_1} + e^{\bar{x}_i \cdot \bar{\beta}_2} \times} \qquad Y \in \{1, 2\}$$

$$= \frac{1}{1+e^{-\bar{x}_i \cdot (\bar{\beta}_1 - \bar{\beta}_2)}}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Comparison with Logistic in the Case of Two Types

$$\boxed{P(Y=1 \mid \bar{x}_i) = \frac{1}{1+e^{-\bar{x}_i \cdot \bar{\beta}}}}$$

$Y \in \{1,2\}$

$$P(Y=1 \mid \bar{x}_i) = \frac{e^{\bar{x}_i \cdot \bar{\beta}_1} \times e^{-\bar{x} \cdot \bar{\beta}_1}}{e^{\bar{x}_i \cdot \bar{\beta}_1} + e^{\bar{x}_i \cdot \bar{\beta}_2} \times}$$

$$= \frac{1}{1+e^{-\bar{x}_i \cdot (\bar{\beta}_1 - \bar{\beta}_2)}}$$

$\beta \equiv \beta_1 - \beta_2$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Gradients

This can all be computed as before. Let's all avoid the pain. First you need the classification probabilities:

$$\underbrace{\pi_{i,j}}_{\text{probabilities (observation } i\text{, class } j)} = \frac{e^{\overrightarrow{x_i}\vec{\beta}_j}}{\sum_k e^{\overrightarrow{x_i}\vec{\beta}_k}}, \vec{\pi}_j = \begin{pmatrix} \pi_{1,j} \\ \vdots \\ \pi_{N,j} \end{pmatrix}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Gradients

Once you compute the loss and take the derivative, you see that:

$$\frac{\partial \mathcal{L}}{\partial \vec{\beta}_j} = -X^T(\vec{y_j} - \vec{\pi}_j).$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Gradients

Once you compute the loss and take the derivative, you see that:

$$\frac{\partial \mathcal{L}}{\partial \vec{\beta}_j} = -X^T(\vec{y_j} - \vec{\pi}_j).$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Gradients

Once you compute the loss and take the derivative, you see that:

$$\frac{\partial \mathcal{L}}{\partial \vec{\beta}_j} = -X^T(\vec{y}_j - \vec{\pi}_j).$$

compare: $-X^T(\bar{y} - \underbrace{\Phi(X\bar{\beta})}_{\text{er}})$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Gradients

Once you compute the loss and take the derivative, you see that:

$$\frac{\partial \mathcal{L}}{\partial \vec{\beta}_j} = -X^T(\overbrace{\vec{y_j} - \vec{\pi}_j}^{\text{error}}).$$

compare: $-X^T(\underbrace{\bar{y} - \sigma(X\bar{\beta})}_{\text{error}})$

$$\bar{y}_i = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \end{bmatrix}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Gradients

Once you compute the loss and take the derivative, you see that:

$$\frac{\partial \mathcal{L}}{\partial \vec{\beta}_j} = -X^T(\overbrace{\vec{y}_j - \vec{\pi}_j}^{\text{error}}).$$

compare: $-X^T(\underbrace{\vec{y} - \vec{\pi}(X\vec{\beta})}_{\text{error}})$

$$\vec{y}_i = \begin{array}{c} \\ \\ j^{th} \\ \\ \\ \end{array}\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \end{bmatrix} \leftarrow y_i = j$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Gradient Descent Algorithm

aws training and certification

And thus the algorithm is largely unchanged.

0 Initialize η

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Gradient Descent Algorithm

aws training and certification

And thus the algorithm is largely unchanged.

① Initialize η, random choice for $(\bar{\beta}_1, \ldots, \bar{\beta}_M)$

② $\bar{\beta}_j \leftarrow \bar{\beta}_j + \eta$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Gradient Descent Algorithm

aws training and certification

And thus the algorithm is largely unchanged.

① Initialize η, random choice for $(\bar{\beta}_1, \ldots, \bar{\beta}_M)$

② $$\boxed{\bar{\beta}_j \leftarrow \bar{\beta}_j + \frac{\eta}{N} X^T (\bar{y}_j - \bar{\pi}_j)}$$

GD to solve sa

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

A Warning

Learning rate, η $\longrightarrow$ small $\Rightarrow$ long time to converge

large $\Rightarrow$ convergence not happen

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

A Warning

Learning rate, η → small ⇒ long time to converge

large ⇒ convergence not happen

monitor loss function

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

A Warning

Learning rate, η → small ⇒ long time to converge

large ⇒ convergence not happen

monitor loss function

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

A Warning

Learning rate, η → small ⇒ long time to converge

large ⇒ convergence not happen

→ monitor loss function

→ normalizing correctly $\left(\frac{1}{N}\right)$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

All Gradient Descents Are The Same

OLS
Logistic
Softmax

$\rightarrow$

$\bar{\beta}$ random

$\bar{\beta}$ —

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Optimization

There is no limit on how stochastic gradient descent (SGD) can be modified to try to avoid pitfalls.

Optimization

Which method will work best depends on your data, and the only way to really know for sure is to **try** them out and select the method that produces the **best test loss**.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Optimization

aws training and certification

Honestly, many machine learning practitioners just pick a favorite.

For example, some practitioners like the comfort of a large body of literature with SGD.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Images

aws training and certification

- https://thenounproject.com/search/?q=science&i=1445239
- https://thenounproject.com/search/?q=pitfall&i=1189524
- https://thenounproject.com/search/?q=try%20again&i=451520
- https://thenounproject.com/search/?q=art&i=225590

-0:06

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Numerical Methods for Optimization

In this final section, we explore optimization models that are more efficient than the logistic gradient descent method. We also explain how to address the learning rate issue and explore a few more optimization hacks that are based on the current and previous classifications of gradients.

The following topics will be covered:

- Numerical methods for optimizing learning models
- A review of the gradient descent method
- Other optimization techniques
  - Newton's method
  - Stochastic gradient descent
  - Mini-batch algorithm
- Optimization techniques based on current and historical gradient values
  - Momentum
  - AdaGrad

Review

- Gradient Descent

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Other Methods

- Other Optimization Methods

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Newton's Method

- Definition of Newton's Method
- Recall: The Hessian

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Newton's Method for Ordinary Least Squares (OLS)

aws training and certification

- Gradients
- Hessians
- The Update Step

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Logistic Regression

- Deriving the Update Step

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Issues With Newton's Method

- An Issue with Newton's Method

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Stochastic Gradient Descent

- The Idea
- What Order?
- Approximate Computational Complexity
- The Big Gains

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Stochastic Gradient Descent

aws training and certification

- The Idea
- What Order?
- Approximate Computational Complexity
- The Big Gains

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Mini Batch

aws training and certification

- Idea
- Pictorially
- A Note on the Use of Noise
- Update Step (and Regularization)
- Definition of Epoch

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Momentum

- A Useful Hack
- Note $\alpha = 0$
- Benefits
- Downsides

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

AdaGrad

- The Expression
- Benefits
- Downsides

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

AdaGrad

- The Expression
- Benefits
- Downsides

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Optimization in Practice

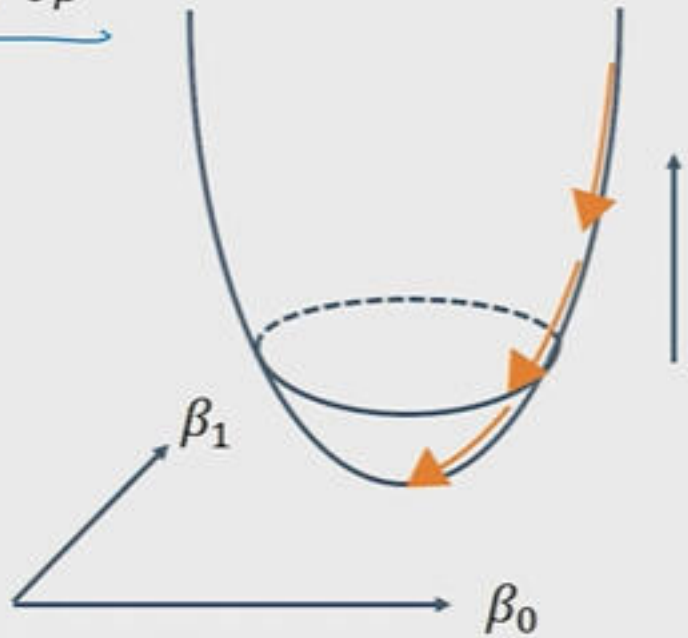
aws training and certification

- Additional Optimization Methods

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Gradient Descent

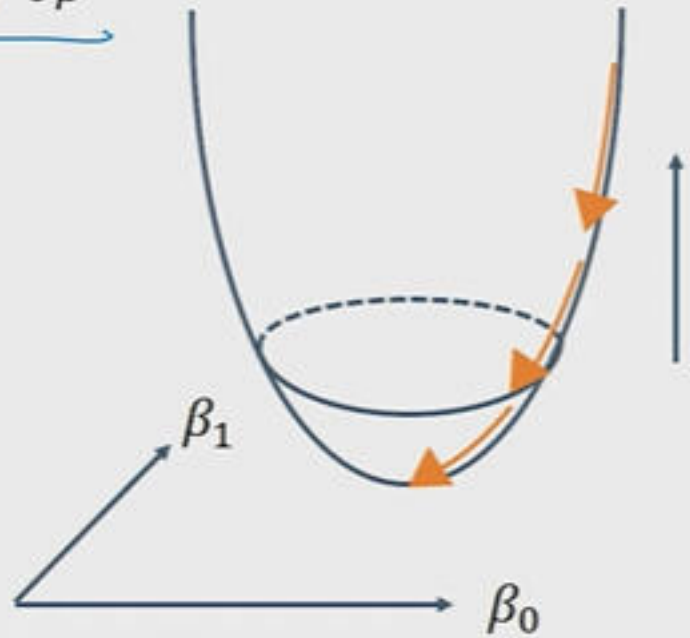
- Generate random parameters $\vec{\beta}$
- Iterate: $\vec{\beta} \leftarrow \vec{\beta} - \eta \frac{\partial \mathcal{L}}{\partial \vec{\beta}}$



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Gradient Descent

- Generate random parameters $\vec{\beta}$
- Iterate: $\vec{\beta} \leftarrow \vec{\beta} - \eta \frac{\partial \mathcal{L}}{\partial \vec{\beta}}$



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

For Regression

$$\frac{\partial \mathcal{L}}{\partial \vec{\beta}} = -X^T(\bar{y} - X\bar{\beta})$$

$$= -X^T(\bar{y} - f(X\bar{\beta}))$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

For Regression

$$\frac{\partial \mathcal{L}}{\partial \vec{\beta}} = -X^T(\bar{y} - X\bar{\beta})$$

$$\frac{\partial \mathcal{L}}{\partial \vec{\beta}} = -X^T(\bar{y} - f(X\bar{\beta}))$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

For Regression

$$\frac{\partial \mathcal{L}}{\partial \vec{\beta}} = -X^T(\bar{y} - X\bar{\beta})$$

$$\frac{\partial \mathcal{L}}{\partial \bar{\beta}} = -X^T(\bar{y} - \underbrace{f}_{\text{link function}}(X\bar{\beta}))$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

For Regression

$$\frac{\partial \mathcal{L}}{\partial \vec{\beta}} = -X^T(\bar{y} - X\bar{\beta})$$

$$\frac{\partial \mathcal{L}}{\partial \bar{\beta}} = -X^T(\bar{y} - \underbrace{f}_{\text{link function}}(X\bar{\beta}))$$

$$f(x) = x$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

For Regression

$$\frac{\partial \mathcal{L}}{\partial \vec{\beta}} = -X^T(\bar{y} - X\bar{\beta})$$

$$\frac{\partial \mathcal{L}}{\partial \vec{\beta}} = -X^T(\bar{y} - \underbrace{f}_{\text{link function}}(X\bar{\beta}))$$

$$f(x) = x \leftarrow \text{OLS}$$

$$f(x) = \frac{1}{1+e^{-x}} \leftarrow$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

For Regression

$$\frac{\partial \mathcal{L}}{\partial \vec{\beta}} = -X^T(\bar{y} - X\bar{\beta})$$

$$\frac{\partial \mathcal{L}}{\partial \bar{\beta}} = -X^T(\bar{y} - \underbrace{f}_{\text{link function}}(X\bar{\beta}))$$

$$f(x) = x \leftarrow \text{OLS}$$

$$f(x) = \frac{1}{1+e^{-x}} \leftarrow \text{Logistic}$$

$$f(x) = e^x \leftarrow \text{Poisson}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Other Optimization Methods

In this video, we briefly discuss other optimization models that are reliant on the gradient descent technique but have a modification.

Optimization techniques based on gradient descent:

- Newton's method ← uses a higher order derivative.
- Stochastic gradient descent ← randomly selects one data point in each iteration.
- Mini-batch gradient descent ← randomly selects data points from set of m-data points.
- Momentum gradient descent ← incorporates additional parameter to capture historical gradients.
- AdaGrad ← modifies learning rate to capture historical gradient.

Other Optimization Methods

aws training and certification

1. Newton's method ← higher order derivative

2. Stochastic Gradient Descent ← random selection of single data point for update.

3. Mini-batch Gradient Descent ← 'm' data points → batch

4. Momentum ← incorporate additional parameter to capture historical gradients

5. AdaGrad ← modifies learning rate to capture historical gradients.

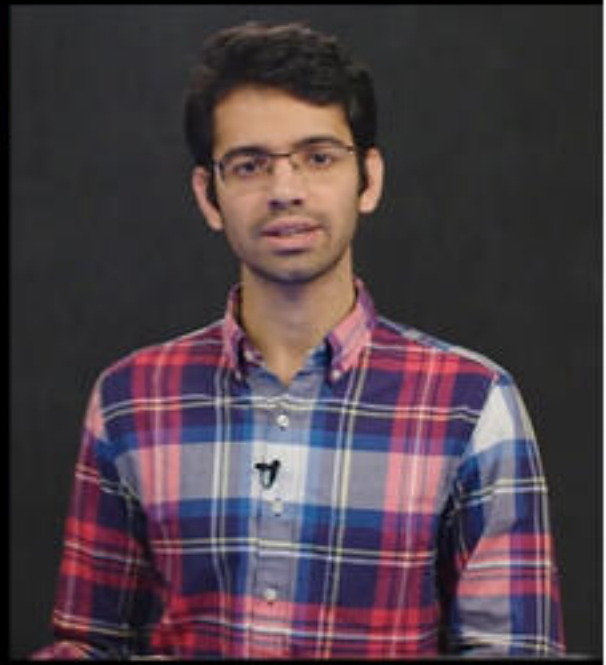
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Other Optimization Methods

aws training and certification

1. Newton's method ← higher order derivative
2. Stochastic Gradient Descent ← random selection of single data point for update
3. Mini-batch Gradient Descent ← 'm' data points → batch
4. Momentum ← incorporate additional parameter to capture historical gradients
5. Adagrad ← modifies learning rate to capture historical gradients

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



-0:13

Other Optimization Methods

aws training and certification

1. Newton's method ← higher order derivative

2. Stochastic Gradient Descent ← random selection of single data point for update.

3. Mini-batch Gradient Descent ← 'm' data points → batch

4. Momentum ← incorporate additional parameter to capture historical gradients

5. AdaGrad ← modifies learning rate to capture historical gradien

-0:14

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definition

$$\vec{\beta} = \vec{\beta} - \underbrace{[HL(\vec{\beta})]^{-1}}\frac{\partial \mathcal{L}}{\partial \vec{\beta}}$$

speeding up convergence
due to higher order derivative

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Recall: The Hessian

aws training and certification

matrix: $$H\mathcal{L}(\bar{\beta}) = \frac{\partial}{\partial \bar{\beta}}(\nabla \mathcal{L}) = \frac{\partial}{\partial \bar{\beta}}\underbrace{\left(\frac{\partial \mathcal{L}}{\partial \bar{\beta}}\right)}_{\text{gradient}}$$

M ind vars $\rightarrow (M+1)$

N observations

$$H\mathcal{L}(\bar{\beta}) = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial \beta_0 \partial \beta_0} & \cdots & \frac{\partial \mathcal{L}}{\partial \beta_0 \partial \beta_M} \\ & & \\ & & \end{bmatrix}_{\times M1}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws training and certification

Newton's Method for Ordinary Least Squares (OLS)

-4:09

Gradients

$$\nabla_{\bar{\beta}} \mathcal{L} = -2 \underbrace{X^T}_{\text{data matrix}} \underbrace{(\bar{y} - X\bar{\beta})}_{\text{residual error}}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Update Step

$$\bar{\beta} \leftarrow \bar{\beta} - \left(H L(\beta)\right)^{-1} \left(\nabla_{\beta} L\right)$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Update Step

$$\bar{\beta} \leftarrow \bar{\beta} - (H L(\beta))^{-1} (\nabla_{\beta} L)$$

$$\bar{\beta} - (2X^T X)^{-1} ($$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Update Step

$$\bar{\beta} \leftarrow \bar{\beta} - \left(H L(\bar{\beta})\right)^{-1} \left(\nabla_{\bar{\beta}} L\right)$$

$$= \bar{\beta} - \left(\cancel{2} X^T X\right)^{-1} \left(-\cancel{2} X^T (\bar{y} - X\bar{\beta})\right)$$

$$= \bar{\beta} + (X^T X)^{-1} \left(X^T y - X^T X \beta\right)$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Update Step

$$\bar{\beta} \leftarrow \bar{\beta} - (H L(\bar{\beta}))^{-1} (\nabla_{\bar{\beta}} L)$$

$$= \bar{\beta} - (2X^T X)^{-1} \left(-2X^T(\bar{y} - X\bar{\beta})\right)$$

$$= \bar{\beta} + (X^T X)^{-1} \left(X^T y - X^T X \beta\right)$$

$$= \bar{\beta} + (X^T X)^{-1} X^T y - (X^T X)^{-1} (X^T X)$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Update Step

$$\bar{\beta} \leftarrow \bar{\beta} - (H L(\bar{\beta}))^{-1} (\nabla_{\bar{\beta}} L)$$

$$= \bar{\beta} - (\cancel{2} X^T X)^{-1} \left(-\cancel{2} X^T (\bar{y} - X\bar{\beta}) \right)$$

$$= \bar{\beta} + (X^T X)^{-1} \left(X^T y - X^T X \bar{\beta} \right)$$

$$= \cancel{\bar{\beta}} + (X^T X)^{-1} X^T y - \cancel{(X^T X)^{-1} (X^T X) \bar{\beta}}$$

$$\boxed{\bar{\beta} = (X^T X)^{-1} X^T y}$$

© 2016, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Update Step

- This is the same as our original closed-form solution!

- For OLS, the Newton's method is exact in **one step**, and is an **alternate way** to derive the exact solutions that have been previously discussed.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Deriving the Update Step

$$\nabla_\beta \mathcal{L} = -X^T\left(\bar{y} - \Phi(X\bar{\beta})\right)$$

$$H\mathcal{L}(\beta) = \frac{\partial}{\partial \beta}\left(-X^T\left(\underbrace{\bar{y}}_{0} - \Phi(X\bar{\beta})\right)\right)$$

$$= X^T \Phi(X\bar{\beta}) \circ (\bar{1}$$

$$\Phi(x) = \begin{bmatrix} \phi(x_1) \\ \phi(x_2) \\ \vdots \\ \phi(x_N) \end{bmatrix}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Deriving the Update Step

$$\nabla_\beta \mathcal{L} = -X^T\left(\bar{y} - \Phi(X\bar{\beta})\right)$$

$$H\mathcal{L}(\beta) = \frac{\partial}{\partial \beta}\left(-X^T\left(\bar{y} - \Phi(X\bar{\beta})\right)\right)$$

$$= X^T \Phi(X\bar{\beta}) \circ \left(\bar{1} - \Phi(X\bar{\beta})\right) X$$

element-wise product

$$\Phi(x) = \begin{bmatrix} \phi(x_1) \\ \phi(x_2) \\ \vdots \\ \phi(x_N) \end{bmatrix}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Deriving the Update Step

$$\nabla_\beta \mathcal{L} = -X^T(\bar{y} - \Phi(X\bar{\beta}))$$

$$H\mathcal{L}(\beta) = \frac{\partial}{\partial \beta}\left(-X^T(\bar{y} - \Phi(X\bar{\beta}))\right)$$

$$= X^T \Phi(X\bar{\beta}) \circ (\bar{1} - \Phi(X\bar{\beta}))X$$

element-wise product

$$\Phi(x) = \begin{bmatrix} \phi(x_1) \\ \phi(x_2) \\ \vdots \\ \phi(x_N) \end{bmatrix}$$

$$\left[\phi(\bar{x}_i \cdot \bar{\beta})(1 - \phi(\bar{x}_i \cdot \bar{\beta})) \right]$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Deriving the Update Step

aws training and certification

$$\nabla_\beta \mathcal{L} = -X^T(\bar{y} - \Phi(X\bar{\beta}))$$

$$H\mathcal{L}(\beta) = \frac{\partial}{\partial \beta}\left(-X^T(\bar{y} - \Phi(X\bar{\beta}))\right)$$

$$= X^T \Phi(X\bar{\beta}) \circ (\bar{1} - \Phi(X\bar{\beta}))\, X$$

element-wise product

vector of 1's

$$\Phi(x) = \begin{bmatrix} \phi(x_1) \\ \phi(x_2) \\ \vdots \\ \phi(x_N) \end{bmatrix}$$

$$\phi(\bar{x}_i \cdot \bar{\beta})\,(1 - \phi(\bar{x}_i \cdot \bar{\beta}))$$

$$\beta \leftarrow \beta - (H\mathcal{L}(\beta))^{-1} \nabla_\beta \mathcal{L}$$

$$= \beta - X^T \Phi(X\bar{\beta}) \circ (\bar{1} - \Phi(X$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Deriving the Update Step

aws training and certification

$$\nabla_\beta \mathcal{L} = -X^T\left(\bar{y} - \Phi(X\bar{\beta})\right)$$

$$H\mathcal{L}(\beta) = \frac{\partial}{\partial \beta}\left(-X^T\left(\bar{y} - \Phi(X\bar{\beta})\right)\right)$$

$$= X^T \Phi(X\bar{\beta}) \circ \left(\bar{1} - \Phi(X\bar{\beta})\right) X$$

element-wise product

vector of $\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}$

$$\Phi(x) = \begin{bmatrix} \phi(x_1) \\ \phi(x_2) \\ \vdots \\ \phi(x_N) \end{bmatrix}$$

$$\left[\phi(\bar{x}_i \cdot \bar{\beta})\left(1 - \phi(\bar{x}_i \cdot \bar{\beta})\right)\right]$$

$$\beta \leftarrow \beta - \left(H\mathcal{L}(\beta)\right)^{-1} \nabla_\beta \mathcal{L}$$

$$= \beta - \left[X^T \Phi(X\bar{\beta}) \circ \left(\bar{1} - \Phi(X\bar{\beta})\right) X\right]^{-1}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Deriving the Update Step

$$\nabla_\beta \mathcal{L} = -X^T(\bar{y} - \Phi(X\bar{\beta}))$$

$$H\mathcal{L}(\beta) = \frac{\partial}{\partial \beta}\left(-X^T(\bar{y} - \Phi(X\bar{\beta}))\right)$$

$$= X^T \Phi(X\bar{\beta}) \circ \left(\bar{1} - \Phi(X\bar{\beta})\right) X$$

element-wise product

vector of 1's

$$\Phi(x) = \begin{bmatrix} \phi(x_1) \\ \phi(x_2) \\ \vdots \\ \phi(x_N) \end{bmatrix}$$

$$\left[\phi(\bar{x}_i \cdot \bar{\beta})\left(1 - \phi(\bar{x}_i \cdot \bar{\beta})\right)\right]$$

$$\beta \leftarrow \beta - \left(H\mathcal{L}(\beta)\right)^{-1} \nabla_\beta \mathcal{L}$$

$$= \beta + \left[X^T \Phi(X\bar{\beta}) \circ \left(\bar{1} - \Phi(X\bar{\beta})\right) X\right]^{-1} X^T(\bar{y} - \Phi$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Deriving the Update Step

aws training and certification

$$\nabla_\beta \mathcal{L} = -X^T(\bar{y} - \Phi(X\bar{\beta}))$$

$$H\mathcal{L}(\beta) = \frac{\partial}{\partial \beta}\left(-X^T(\bar{y} - \Phi(X\bar{\beta}))\right)$$

$$= X^T \Phi(X\bar{\beta}) \circ \left(\bar{1} - \Phi(X\bar{\beta})\right) X$$

element-wise product

vector of 1's

$$\Phi(x) = \begin{bmatrix} \phi(x_1) \\ \phi(x_2) \\ \vdots \\ \phi(x_N) \end{bmatrix}$$

$$\left[\phi(\bar{x}_i \cdot \bar{\beta})\left(1 - \phi(\bar{x}_i \cdot \bar{\beta})\right)\right]$$

$$\beta \leftarrow \beta - \left(H\mathcal{L}(\beta)\right)^{-1} \nabla_\beta \mathcal{L}$$

$$= \beta + \left[X^T \Phi(X\bar{\beta}) \circ \left(\bar{1} - \Phi(X\bar{\beta})\right) X\right]^{-1} X^T(\bar{y} - \Phi(X\bar{\beta}))$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

An Issue with Newton's Method

aws training and certification

Slow

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

An Issue with Newton's Method

aws training and certification

Slow

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

An Issue with Newton's Method

aws training and certification

Slow

quadratic dependence # fea

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

An Issue with Newton's Method

aws training and certification

Slow

quadratic dependence # features.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws training and certification

Stochastic Gradient Descent

-11:53

aws training and certification

Stochastic Gradient Descent

-11:54

Topics

- The Idea
- What Order?
  - Cycle through
  - Random elements
  - Shuffle then cycle through
- Approximate Computational Complexity
- The Big Gains

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Idea

$$\bar{\beta} \leftarrow \bar{\beta} - \underbrace{\eta}_{\text{learning rate}} \underbrace{(\nabla_{\bar{\beta}} \mathcal{L})}_{\text{gradient of } \mathcal{L}}$$

$\longrightarrow$ defined on $X, \bar{y}$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Idea

$$\bar{\beta} \leftarrow \bar{\beta} - \underbrace{\eta}_{\text{learning rate}} \underbrace{\left(\nabla_{\bar{\beta}} \mathcal{L}\right)}_{\text{gradient of } \mathcal{L}}$$

$\longrightarrow$ defined on $X, \bar{y}$

GD:

$$\Big\downarrow \text{OLS}$$

$$\bar{\beta} \leftarrow \bar{\beta} + X^T(\bar{y} - X\bar{\beta})$$

SGD : single observation in each update

β

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The idea

$$\bar{\beta} \leftarrow \bar{\beta} - \underbrace{\eta}_{\text{learning rate}} \underbrace{\left(\nabla_{\bar{\beta}} \mathcal{L}\right)}_{\text{gradient of } \mathcal{L}}$$

$\longrightarrow$ defined on $X, \bar{y}$

GD:

OLS

$$\bar{\beta} \leftarrow \bar{\beta} + \eta X^T (\bar{y} - X\bar{\beta})$$

SGD : single observation in each update.

$$\boxed{\bar{\beta} \leftarrow \bar{\beta} + \eta\, \bar{x}_i^T (y_i - \bar{x}_i \cdot \bar{\beta})}$$

observation : one at a time.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The idea

$$\bar{\beta} \leftarrow \bar{\beta} - \underbrace{\eta}_{\text{learning rate}} \underbrace{\left(\nabla_{\bar{\beta}} \mathcal{L}\right)}_{\text{gradient of } \mathcal{L}} \longrightarrow \text{defined on } X, \bar{y}$$

GD:

OLS

$$\bar{\beta} \leftarrow \bar{\beta} + \eta X^T (\bar{y} - X\bar{\beta})$$

SGD : single observation in each update.

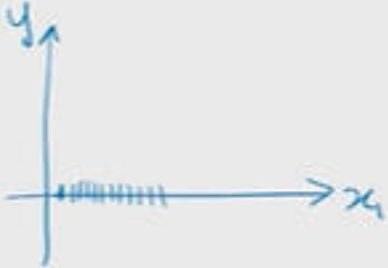
$$\boxed{\bar{\beta} \leftarrow \bar{\beta} + \eta\, \bar{x}_i^T (y_i - \bar{x}_i \cdot \bar{\beta})}$$

observation : one at a time.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Order: Cycle Through?

- Not robust to patterns in the dataset



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Order: Random Elements?

aws training and certification

- You'll see some data points **many times** before you see others even **once**.

1 2 3 . . . . . N

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Order: Shuffle Then Cycle Through?

aws training and certification

- Shuffle, then cycle through, and repeat
- Take advantage of the benefits of both

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Order: Shuffle Then Cycle Through?

aws training and certification

- Shuffle, then cycle through, and repeat
- Take advantage of the benefits of both

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Order: Shuffle Then Cycle Through?

aws training and certification

- Shuffle, then cycle through, and repeat
- Take advantage of the benefits of both

randomization cycling.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Approximate Computational Complexity

aws training and certification

① Computational complexity of each iteration

② # iterations needed to reach an accuracy of ϵ.

GD	SGD
N	1

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Approximate Computational Complexity

aws training and certification

① Computational complexity of each iteration

② # iterations needed to reach an accuracy of ϵ

③ Comp complexity to reach

GD	SGD
N	1
$\log(1/\epsilon)$	$1/\epsilon$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Approximate Computational Complexity

aws training and certification

① Computational complexity of each iteration

② # iterations needed to reach an accuracy of ϵ

③ Comp complexity to reach accuracy

GD	SGD
N	1
$\log(1/\epsilon)$	$1/\epsilon$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Approximate Computational Complexity

aws training and certification

① Computational complexity of each iteration

② # iterations needed to reach an accuracy of ϵ

③ Comp complexity to reach accuracy

GD	SGD
N	1
$\log(1/\epsilon)$	$1/\epsilon$
$N \log(1/\epsilon)$	$1/\epsilon$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Gain: Work and Accuracy

aws training and certification

Work and accuracy are directly related

No longer conflate the notions of the amount of data and desired accuracy

© 2016, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Gain: Work and Accuracy

aws training and certification

Work and accuracy are directly related

No longer conflate the notions of the amount of data and desired accuracy

SGD ϵ : $(1/\epsilon)$

GD ϵ : $N \log(1/\epsilon)$

© 2016, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Gain: Work and Accuracy

aws training and certification

Work and accuracy are directly related

No longer conflate the notions of the amount of data and desired accuracy

SGD ϵ : $(1/\epsilon)$

GD ϵ : $\underline{N} \log(1/\epsilon)$

© 2016, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Gain: Work and Accuracy

aws training and certification

Work and accuracy are directly related

No longer conflate the notions of the amount of data and desired accuracy

SGD ϵ : $(1/\epsilon)$

GD ϵ : $\underline{N} \log(1/\epsilon))$

© 2016, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Gain: Online Learning

aws training and certification

Can be done *online*

→ iteration only uses a single observation

→ assumption: distribution of data doesn't change w/time

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Gain: Online Learning

aws training and certification

Can be done *online*

→ iteration only uses a single observation

→ assumption: distribution of data doesn't change w/time

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Update Step (and Regularization)

aws training and certification

$$\frac{\mathcal{L}}{N} = \underbrace{\frac{1}{N}\|\bar{y} - X\bar{\beta}\|_2^2}_{\mathcal{L}_1} + \underbrace{\frac{\delta_2^2}{N}\|\beta\|_2^2}_{\mathcal{L}_2} \longleftarrow \text{Ridge reg.}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Update Step (and Regularization)

$$\frac{\mathcal{L}}{N} = \underbrace{\frac{1}{N}\|\bar{y} - X\bar{\beta}\|_2^2}_{\mathcal{L}_1} + \underbrace{\frac{\delta_2^2}{N}\|\beta\|_2^2}_{\mathcal{L}_2} \longleftarrow \text{Ridge reg.}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Update Step (and Regularization)

aws training and certification

$$\frac{\mathcal{L}}{N} = \underbrace{\frac{1}{N}\|\bar{y} - X\bar{\beta}\|_2^2}_{\substack{\mathcal{L}_1 \\ \text{error}}} + \underbrace{\frac{\delta_2^2}{N}\|\beta\|_2^2}_{\mathcal{L}_2} \longleftarrow \text{Ridge reg.}$$

$$\underbrace{\frac{1}{m}}\underbrace{\|\bar{y} - X\bar{\beta}\|_2^2}_{\text{error}}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Update Step (and Regularization)

aws training and certification

$$\frac{\mathcal{L}}{N} = \underbrace{\frac{1}{N}\|\bar{y} - X\bar{\beta}\|_2^2}_{\substack{\mathcal{L}_1 \\ \text{error}}} + \underbrace{\left(\frac{\delta_2^2}{N}\|\beta\|_2^2\right)}_{\mathcal{L}_2} \longleftarrow \text{Ridge reg.}$$

$$\underbrace{\frac{1}{m}}\underbrace{\|\bar{y} - X\bar{\beta}\|_2^2}_{\text{error}} + \frac{m}{N}\delta_2^2\|\beta\|_2^2$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Update Step (and Regularization)

aws training and certification

$$\frac{\mathcal{L}}{N} = \underbrace{\frac{1}{N}\|\bar{y} - X\bar{\beta}\|_2^2}_{\substack{\mathcal{L}_1 \\ \text{error}}} + \underbrace{\left(\frac{\delta_2^2}{N}\|\beta\|_2^2\right)}_{\mathcal{L}_2} \longleftarrow \text{Ridge reg.}$$

$$\mathcal{L}_{m.b} = \underbrace{\frac{1}{m}}\underbrace{\|\bar{y} - X\bar{\beta}\|_2^2}_{\text{error}} + \frac{m}{N}\delta_2^2\|\beta\|_2^2$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definition of *Epoch*

SGD : 1 obs
M.B : m obs.
GD : N obs

update step

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definition of *Epoch*

SGD : 1 obs
M.B : m obs.
GD : N obs

update step

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definition of *Epoch*

aws training and certification

SGD : 1 obs
M.B : m obs.
GD : N obs

} update step

Epoch : Number of iterations to pass over the entire dataset.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definition of *Epoch*

aws training and certification

SGD : 1 obs
M.B : m obs.
GD : N obs.

} update step

Epoch : Number of iterations to pass over the entire dataset.

SGD : N iterations
MB.

} 1 epoch

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definition of *Epoch*

aws training and certification

SGD : 1 obs
M.B : m obs.
GD : N obs

} update step

Epoch : Number of iterations to pass over the entire dataset.

SGD : N iterations
MB

} 1 epoch

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definition of *Epoch*

aws training and certification

SGD : 1 obs
M.B : m obs.
GD : N obs.

} update step

Epoch : Number of iterations to pass over the entire dataset.

SGD : N iterations
MB

} 1 epoch

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definition of *Epoch*

aws training and certification

SGD : 1 obs
M.B : m obs.
GD : N obs.

} update step

Epoch : Number of iterations to pass over the entire dataset.

SGD : N iterations
MB :

} 1 epoch

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Definition of *Epoch*

aws training and certification

SGD : 1 obs
M.B : m obs.
GD : N obs.

} update step

Epoch : Number of iterations to pass over the entire dataset.

SGD : N iterations
MB : $\frac{N}{m}$ iterations
GD : 1 iterati

} 1 epoch

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Per Epoch is Optimal to Compare

aws training and certification

For most relatives to gradient descent, it takes almost the exact same clock-time to compute an epoch (for example, up to about 10 percent).

Thus, per-epoch improvement in loss can be used as a proxy for how long it will take you to learn something.

A Useful Hack

aws training and certification

GD: $\bar{\beta} \leftarrow \underbrace{\bar{\beta} - \eta(\nabla L)}$

Momentum: $\bar{\beta} \leftarrow \beta + \bar{v}$

$$\bar{v} \leftarrow (1-\alpha)\underbrace{(-\eta \nabla L)}_{GD} + \alpha \bar{v}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Note

$\alpha = 0$

$$\bar{v} \leftarrow (1-\alpha)\underbrace{(-\eta \nabla L)} + \underbrace{\alpha \bar{v}}_{\to 0 \text{ when } \alpha = 0}$$

$$\bar{\beta} \leftarrow \bar{\beta} + \bar{v}$$

when $\alpha = 0$, Momentum $\longrightarrow$ GD

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Note

$\alpha = 0$

$$\bar{v} \leftarrow (1-\alpha)\underbrace{(-\eta \nabla L)} + \underbrace{\alpha \bar{v}}_{\to 0 \text{ when } \alpha = 0}$$

$$\bar{\beta} \leftarrow \bar{\beta} + \bar{v}$$

when $\alpha = 0$, Momentum $\longrightarrow$ GD

when $\alpha = 1$ $\longrightarrow$ X

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Note

$\alpha = 0$

$$\bar{v} \leftarrow (1-\alpha)\underbrace{(-\eta \nabla L)} + \underbrace{\alpha \bar{v}}_{\to 0 \text{ when } \alpha = 0}$$

$$\bar{\beta} \leftarrow \bar{\beta} + \bar{v}$$

when $\alpha = 0$, Momentum $\longrightarrow$ GD

when $\alpha = 1$ $\longrightarrow$ X

$\alpha \approx 0.9$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Note

$\alpha = 0$

$$\bar{v} \leftarrow (1-\alpha)\underbrace{(-\eta \nabla L)} + \underbrace{\alpha \bar{v}}_{0 \text{ when } \alpha=0}$$

$$\bar{\beta} \leftarrow \bar{\beta} + \bar{v}$$

when $\alpha=0$, Momentum $\longrightarrow$ GD

when $\alpha=1$ $\longrightarrow$ X

$\alpha \approx 0.9 \leftarrow$ pro

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Benefits

aws training and certification

Smoothing

$$v \leftarrow (1-\alpha)(-\eta \nabla) + \alpha v$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Benefits

Smoothing

$$v \leftarrow (1-\alpha)(-\eta \nabla) + \alpha v$$

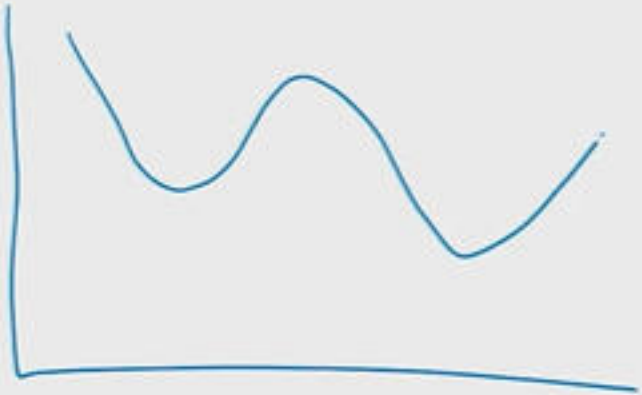
smoothening

EWA

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Benefits

Escapes local minima



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Benefits

aws training and certification

Escapes local minima

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Benefits

Escapes local minima

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Benefits

Escapes local minima

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Downsides

aws training and certification

Extra meta-parameter

α

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Downsides

aws training and certification

Extra meta-parameter

α

hyperparameter

→ Cross-validati

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Downsides

aws training and certification

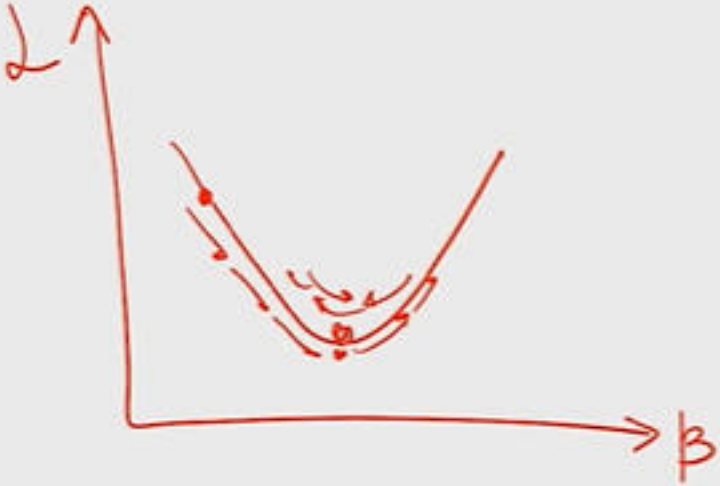
Oscillations

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Downsides

aws training and certification

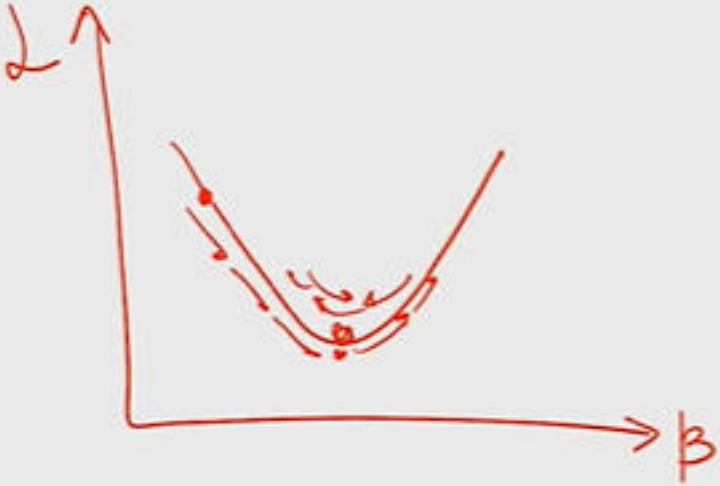
Oscillations



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Downsides

Oscillations



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws training and certification

AdaGrad

-5:19

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Topics

- The Expression
- Benefits
- Downsides

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

AdaGrad

aws training and certification

Attempts to adapt the **step size**
to make them all **of the same scale**

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

AdaGrad

aws training and certification

Attempts to adapt the **step size**
to make them all **of the same scale**

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Expression

$$\beta \leftarrow \beta + \eta \frac{\partial \mathcal{L}}{\partial \beta}$$

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Expression

$$\beta \leftarrow \beta + \eta \frac{\partial \mathcal{L}}{\partial \beta}$$

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Expression

$$\beta_i \leftarrow \beta_i + \eta \frac{\partial L}{\partial \beta_i} \longleftarrow \text{ADAGRA}$$

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Expression

$$\beta_i \leftarrow \beta_i + \eta \frac{\partial L}{\partial \beta_i} \longleftarrow \text{ADAGRAD}$$

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Expression

$$\beta_i \leftarrow \beta_i + \eta \frac{\partial L}{\partial \beta_i} \longleftarrow \text{ADAGRAD}$$

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Expression

$$\beta_i \leftarrow \beta_i + \eta \frac{\partial L}{\partial \beta_i} \longleftarrow \text{ADAGRAD.}$$

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Expression

$$\beta_i \leftarrow \beta_i + \frac{\eta}{\sqrt{c + \frac{\partial \mathcal{L}}{\partial \beta_i}}} \frac{\partial \mathcal{L}}{\partial \beta_i} \longleftarrow \text{ADAGRAD.}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Expression

$$\beta_i \leftarrow \beta_i + \frac{\eta}{\sqrt{c + \left(\frac{\partial L}{\partial \beta_i}\right)^2}} \frac{\partial L}{\partial \beta_i} \longleftarrow \text{ADAGRAD.}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Expression

$$\beta_i \leftarrow \beta_i + \frac{\eta}{\sqrt{c + \sum_t \left(\frac{\partial \mathcal{L}}{\partial \beta_i}\right)^2}} \frac{\partial \mathcal{L}}{\partial \beta_i} \longleftarrow \text{ADAGRAD.}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Expression

$$\beta_i \leftarrow \beta_i + \frac{\eta}{\sqrt{c + \sum_t \left(\frac{\partial L}{\partial \beta_i}\right)^2}} \frac{\partial L}{\partial \beta_i} \longleftarrow \text{ADAGRAD.}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Expression

$$\beta_i \leftarrow \beta_i + \frac{\eta}{\sqrt{c + \sum_{t'} \left(\frac{\partial \mathcal{L}}{\partial \beta_i}\right)^2}} \frac{\partial \mathcal{L}}{\partial \beta_i} \longleftarrow \text{ADAGRAD.}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Expression

$$\beta_i \leftarrow \beta_i + \frac{\eta}{\sqrt{c + \sum_{t=} \left(\frac{\partial L}{\partial \beta_i}\right)^2}} \underline{\frac{\partial L}{\partial \beta_i}} \longleftarrow \text{ADAGRAD.}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Expression

$$\beta_i \leftarrow \beta_i + \frac{\eta}{\sqrt{c + \sum_{t=1}^{s} \left(\frac{\partial L}{\partial \beta_i}\right)^2}} \frac{\partial L}{\partial \beta_i} \longleftarrow \text{ADAGRAD.}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Expression

$$\beta_i \leftarrow \beta_i + \frac{\eta}{\sqrt{c + \sum_{t=1}^{s} \left(\frac{\partial L}{\partial \beta_i}\right)^2}} \frac{\partial L}{\partial \beta_i} \longleftarrow \text{ADAGRAD.}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Expression

$$\beta_i \leftarrow \beta_i^{(s)} + \frac{\eta}{\sqrt{c + \sum_{t=1}^{s} \left(\frac{\partial \mathcal{L}}{\partial \beta_i}\right)^2}} \frac{\partial \mathcal{L}}{\partial \beta_i} \longleftarrow \text{ADAGRAD.}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Expression

$$\beta_i^{(s+1)} \longleftarrow \beta_i^{(s)} + \frac{\eta}{\sqrt{c + \sum_{t=1}^{s} \left(\frac{\partial \mathcal{L}}{\partial \beta_i}\right)^2}} \underline{\frac{\partial \mathcal{L}}{\partial \beta_i}} \longleftarrow \text{ADAGRAD.}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The Expression

$$\beta_i^{(s+1)} \longleftarrow \beta_i^{(s)} + \frac{\eta}{\sqrt{c + \sum_{t=1}^{s} \left(\frac{\partial \mathcal{L}}{\partial \beta_i^{(s)}}\right)^2}} \underline{\frac{\partial \mathcal{L}}{\partial \beta_i^{(s)}}} \longleftarrow \text{ADAGRAD.}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Benefits

aws training and certification

Not sensitive to meta-parameters

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Benefits

aws training and certification

Not sensitive to meta-parameters

η, c

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Benefits

aws training and certification

Very simple to implement

$$\frac{\partial}{\partial \beta_i^{(t)}}$$

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Benefits

aws training and certification

Very simple to implement

$$\left(\frac{\partial \mathcal{L}}{\partial \beta_i^{(t)}}\right) \quad \forall i, t.$$

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Downsides

aws training and certification

Learning rate death

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Downsides

Learning rate death

$$\frac{\eta}{\sqrt{c + \sum_{t=0}^{S} \left(\frac{\partial L}{\partial \beta_i^{(t)}} \right)^2}}$$

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Downsides

Learning rate death

$$\frac{1}{\sqrt{c + \sum_{t=0}^{S} \left(\frac{\partial L}{\partial \beta_i^{(t)}} \right)^2}}$$

large value. ⇒ learning rate becomes very s.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Downsides

Learning rate death

$$\frac{1}{\sqrt{c + \sum_{t=0}^{S} \left(\frac{\partial L}{\partial \beta_i^{(t)}} \right)^2}}$$

large value. $\Rightarrow$ learning rate becomes very small

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Optimization

aws training and certification

Optimization is equal parts

art

and

science

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Optimization

Which method will work best depends on your data, and the only way to really know for sure is to **try** them out and select the method that produces the **best test loss**.

Optimization

Which method will work best depends on your data, and the only way to really know for sure is to **try** them out and select the method that produces the **best test loss**.

Optimization

Which method will work best depends on your data, and the only way to really know for sure is to **try** them out and select the method that produces the **best test loss**.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Optimization

Honestly, many machine learning practitioners just pick a favorite.

For example, some practitioners like the comfort of a large body of literature with SGD.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Optimization

Honestly, many machine learning practitioners just pick a favorite.

For example, some practitioners like the comfort of a large body of literature with SGD.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Optimization

Honestly, many machine learning practitioners just pick a favorite.

For example, some practitioners like the comfort of a large body of literature with SGD.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Optimization

Honestly, many machine learning practitioners just pick a favorite.

For example, some practitioners like the comfort of a large body of literature with SGD.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Discussion of Methods

For a robust discussion of possible methods, see:
http://ruder.io/optimizing-gradient-descent/

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Images

aws training and certification

- https://thenounproject.com/search/?q=science&i=1445239
- https://thenounproject.com/search/?q=pitfall&i=1189524
- https://thenounproject.com/search/?q=try%20again&i=451520
- https://thenounproject.com/search/?q=art&i=225590

-0:08

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws training and certification

Thank You

© 2018 Amazon Web Services, Inc. or its affiliates. All rights reserved. This work may not be reproduced or redistributed, in whole or in part, without prior written permission from Amazon Web Services, Inc. Commercial copying, lending, or selling is prohibited. Corrections or feedback on the course, please email us at: aws-course-feedback@amazon.com. For all other questions, contact us at: https://aws.amazon.com/contact-us/aws-training/. All trademarks are the property of their owners.

aws training and certification

Regression for Machine Learning: In Summary

aws training and certification

Regression for Machine Learning: In Summary

Learning Outcomes

aws training and certification

In this track, you learnt how to:

- Describe how loss functions of regression models are motivated.
- Derive maximum likelihood estimators of simple statistical models.
- Adapt loss functions of regression models for problem-specific requirements including regularization.
- Understand batch, stochastic gradient training algorithms of regression models.
- Identify practical tricks for using training algorithms, such as momentum, Adagrad.

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Optional Reading

aws training and certification

- K. Murphy, Machine Learning: A Probabilistic Perspective
- C. Bishop, Pattern Recognition and Machine Learning
- T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction (available free online)

-0:14

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Knowledge Check: Regression - Course 5

aws training and certification

What is the Newton's method update step for OLS?

a. $\vec{\beta} - \frac{\eta}{N} X^T(\vec{y} - X\vec{\beta})$

b. $\vec{\beta} + (X^T X)^{-1} X^T \vec{y}$

c. $(X^T X)^{-1} X^T \vec{y}$

d. $\vec{\beta} + \frac{\eta}{N} X^T(\vec{y} - X\vec{\beta})$

Knowledge Check: Regression - Course 5

aws training and certification

What is the preferred ordering of data points for Stochastic Gradient Descent (SGD)?

a. Cycle through starting with the first element
b. Random selection
c. Shuffle and then cycle through
d. Cycle through starting with a random element

Knowledge Check: Regression - Course 5

aws training and certification

Which of the following is NOT an advantage of SGD over Gradient Descent?

a. SGD can be used with streaming data

b. SGD has lower computational complexity for large datasets

c. SGD requires fewer updates steps to converge (up to a small ϵ error)

d. SGD update step has lower memory requirements

Knowledge Check: Regression - Course 5

aws training and certification

Which of the following is a feature of Gradient Descent with Momentum? (choose all that apply)

a. It is a exponentially weighted average function of gradients computed at different coefficient values

b. It uses a single data point in each update step

c. It helps in avoiding convergence to local minima

d. Step size in a momentum update step is variable

Knowledge Check: Regression - Course 4

aws training and certification

Which of the following is the cost function $J(\beta)$ for logistic regression?

a. $\sum_{i=1}^{N} y_i \log\left(1 + e^{\beta^T x_i}\right) + (1 - y_i) \log\left(1 + e^{\beta^T x_i}\right)$

b. $\sum_{i=1}^{N} y_i \log\left(1 + e^{-\beta^T x_i}\right) + (1 - y_i) \log\left(1 + e^{-\beta^T x_i}\right)$

c. $\sum_{i=1}^{N} y_i \log\left(1 + e^{\beta^T x_i}\right) + (1 - y_i) \log\left(1 + e^{-\beta^T x_i}\right)$

d. $\sum_{i=1}^{N} y_i \log\left(1 + e^{-\beta^T x_i}\right) + (1 - y_i) \log\left(1 + e^{\beta^T x_i}\right)$

$$J(\beta) = -\sum_{i=1}^{N} \left[y_i \log\left(\phi(\bar{x}_i \cdot \bar{\beta})\right) + (1 - y_i) \log\left(1 - \phi(\bar{x}_i \cdot \bar{\beta})\right) \right]$$

$$\phi(x) = \frac{1}{1 + e^{-x}}$$

Knowledge Check: Regression - Course 4

aws training and certification

Which of the following is the gradient of $J(\beta)$ in matrix form?

a. $-X^T(\vec{y} - \Phi\left(X\vec{\beta}\right))$

b. $-(\vec{y} - \Phi\left(X\vec{\beta}\right))X^T$

c. $-(\vec{y} - \Phi\left(X\vec{\beta}\right))$

d. $-X^T(\vec{y} - \Phi\left(X\vec{\beta}\right))X^T$

Knowledge Check: Regression - Course 4

aws training and certification

Write down the update step for $J(\beta)$ with Gradient Descent?

a. $\vec{\beta} - \frac{\eta}{N} X^T (\vec{y} + \Phi\left(X\vec{\beta}\right))$

b. $\vec{\beta} - \frac{\eta}{N} X^T (\vec{y} - \Phi\left(X\vec{\beta}\right))$

c. $\vec{\beta} + \frac{\eta}{N} X^T (\vec{y} - \Phi\left(X\vec{\beta}\right))$

d. $\vec{\beta} + \eta X^T (\vec{y} + \Phi\left(X\vec{\beta}\right))$

Knowledge Check: Regression - Course 4

aws training and certification

Write down the update step for $J(\beta)$ with Gradient Descent?

a. $\vec{\beta} - \frac{\eta}{N} X^T(\vec{y} + \Phi\left(X\vec{\beta}\right))$

b. $\vec{\beta} - \frac{\eta}{N} X^T(\vec{y} - \Phi\left(X\vec{\beta}\right))$

c. $\vec{\beta} + \frac{\eta}{N} X^T(\vec{y} - \Phi\left(X\vec{\beta}\right))$

d. $\vec{\beta} + \eta X^T(\vec{y} + \Phi\left(X\vec{\beta}\right))$

$$\vec{\beta} \leftarrow \vec{\beta} + \frac{\eta}{N}\left(-\frac{\partial J}{\partial \vec{\beta}}\right)$$

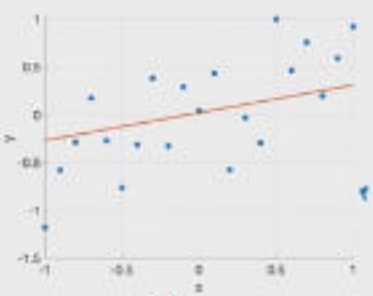
$$= \vec{\beta} + \frac{\eta}{N} X^T\left(\vec{y} - \Phi\left(X\vec{\beta}\right)\right)$$

Knowledge Check: Regression - Course 3

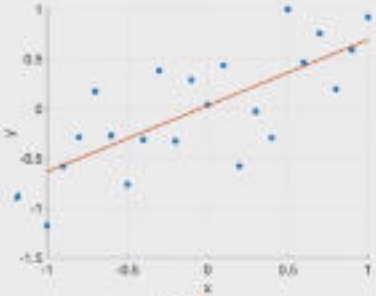
aws training and certification

Assign each figure below to the appropriate model. Assume that δ^2 is the hyperparameter for L2 regularization and λ as the hyperparameter for radial basis function (RBFs): $K(x, \mu_i, \lambda) = e^{-\left(\frac{1}{\lambda}\right)\|x - \mu_i\|^2}$

a. Linear regression δ2=0.01
b. Linear regression δ2=10
c. Linear regression with RBFs as features δ2=0.01 and λ=1
d. Linear regression with RBFs as features δ2=10 and λ=1



(b)



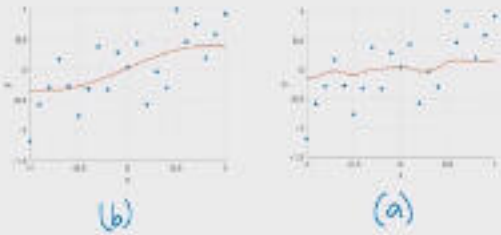
(a)

Knowledge Check: Regression - Course 3

aws training and certification

Assign each figure below to the appropriate model. Assume that δ^2 is the hyperparameter for L2 regularization and λ as the hyperparameter for radial basis function (RBFs): $K(x, \mu_i, \lambda) = e^{-\left(\frac{1}{\lambda}\right)\|x - \mu_i\|^2}$

a. Linear regression with RBFs as features δ2=10 and λ=0.01
b. Linear regression with RBFs as features δ2=10 and λ=1
c. Linear regression δ2=0.01
d. Linear regression δ2=10

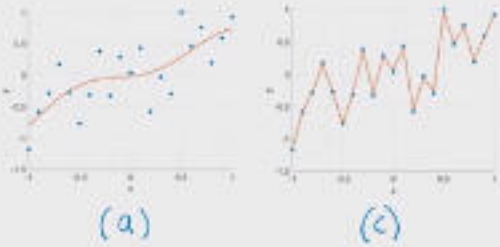


Knowledge Check: Regression - Course 3

aws training and certification

Assign each figure below to the appropriate model. Assume that δ^2 is the hyperparameter for L2 regularization and λ as the hyperparameter for radial basis function (RBFs): $K(x, \mu_i, \lambda) = e^{-\left(\frac{1}{\lambda}\right)\|x - \mu_i\|^2}$

a. Linear regression with RBFs as features δ2=0.01 and λ=1
b. Linear regression δ2=0.01
c. Linear regression with RBFs as features δ2=0.01 and λ=0.01
d. Linear regression δ2=10



Knowledge Check: Regression - Course 2

aws training and certification

Suppose I tossed a coin 5 times and observed the outcome: (T H T T H). The coin outcome follows a Bernoulli distribution. The Bernoulli distribution has a single parameter, 'p', which represents the probability of seeing a Head. What is the maximum likelihood estimate of p?

a. 1/6

b. 3/5

c. 2/5

d. 5/6

Knowledge Check: Regression - Course 2

aws training and certification

Suppose I tossed a coin 5 times and observed the outcome: (T H T T H). The coin outcome follows a Bernoulli distribution. The Bernoulli distribution has a single parameter, 'p', which represents the probability of seeing a Head. What is the maximum likelihood estimate of p?

a. 1/6

b. 3/5

c. 2/5

d. 5/6

$$L = P(T)P(H)P(T)P(T)P(H) \quad \left(= P(\{T, H, T, T, H\})\right)$$

$$= (1-p)(p)(1-p)(1-p)(p)$$

$$= p^2(1-p)^3$$

$$\log L = 2\log(p) + 3\log(1-p)$$

Knowledge Check: Regression - Course 2

aws training and certification

Suppose I tossed a coin 5 times and observed the outcome: (T H T T H). The coin outcome follows a Bernoulli distribution. The Bernoulli distribution has a single parameter, 'p', which represents the probability of seeing a Head. What is the maximum likelihood estimate of p?

a. 1/6

b. 3/5

c. 2/5

d. 5/6

$$L = P(T)P(H)P(T)P(T)P(H) \quad \left(= P(\{T, H, T, T, H\})\right)$$

$$= (1-p)(p)(1-p)(1-p)(p)$$

$$= p^2(1-p)^3$$

$$\log L = 2\log(p) + 3\log(1-p)$$

Knowledge Check: Regression - Course 2

aws training and certification

Suppose I tossed a coin 5 times and observed the outcome: (T H T T H). The coin outcome follows a Bernoulli distribution. The Bernoulli distribution has a single parameter, 'p', which represents the probability of seeing a Head. What is the maximum likelihood estimate of p?

a. 1/6

b. 3/5

c. 2/5

d. 5/6

$$L = P(T)P(H)P(T)P(T)P(H) \quad \left(= P(\{T, H, T, T, H\}) \right)$$

$$= (1-p)(p)(1-p)(1-p)(p)$$

$$= p^2(1-p)^3$$

$$\log L = 2\log(p) + 3\log(1-p)$$

$$\frac{d(\log L)}{dp} = 0 \Rightarrow$$

Knowledge Check: Regression - Course 2

aws training and certification

Suppose I tossed a coin 5 times and observed the outcome: (T H T T H). The coin outcome follows a Bernoulli distribution. The Bernoulli distribution has a single parameter, 'p', which represents the probability of seeing a Head. What is the maximum likelihood estimate of p?

a. 1/6

b. 3/5

c. 2/5

d. 5/6

$$L = P(T)P(H)P(T)P(T)P(H) \quad \left(= P(\{T, H, T, T, H\})\right)$$

$$= (1-p)(p)(1-p)(1-p)(p)$$

$$= p^2(1-p)^3$$

$$\log L = 2\log(p) + 3\log(1-p)$$

$$\frac{d(\log L)}{dp} = 0 \Rightarrow \frac{2}{p} + \frac{3}{1-p}(-1) = 0$$

$$\Rightarrow p =$$

Knowledge Check: Regression - Course 2

aws training and certification

Now, $J(\beta) = \sum_{i=1}^{N} \omega_i \left(y_i - \vec{x_i} \cdot \vec{\beta}\right)^2 = \left(\vec{y} - X\vec{\beta}\right)^T W(\vec{y} - X\vec{\beta})$

What does the expression $\frac{\partial J}{\partial \beta}$ evaluate to?

$$\frac{\partial J}{\partial \beta} = \frac{\partial}{\partial \vec{\beta}}\left(\vec{y}^T - \vec{\beta}^T X^T\right)\left(W\vec{y} - WX\vec{\beta}\right)$$

$$= \frac{\partial}{\partial \vec{\beta}}\left[\vec{y}^T W\vec{y} - \vec{\beta}^T X^T W\vec{y} + \vec{\beta}^T X^T W X\vec{\beta} - \vec{y}^T W X\vec{\beta}\right]$$

a. $-2X^T\vec{y} + 2X^TX\vec{\beta}$

b. $-X^TW\vec{y} + X^TWX\vec{\beta}$

c. $-2X^T\vec{y} + 2X^TWX\vec{\beta}$

d. $-2X^TW\vec{y} + 2X^TWX\vec{\beta}$

Knowledge Check: Regression - Course 2

aws training and certification

Now, $J(\beta) = \sum_{i=1}^{N} \omega_i \left(y_i - \vec{x_i} \cdot \vec{\beta}\right)^2 = \left(\vec{y} - X\vec{\beta}\right)^T W(\vec{y} - X\vec{\beta})$

What does the expression $\frac{\partial J}{\partial \beta}$ evaluate to?

a. $-2X^T\vec{y} + 2X^TX\vec{\beta}$

b. $-X^TW\vec{y} + X^TWX\vec{\beta}$

c. $-2X^T\vec{y} + 2X^TWX\vec{\beta}$

d. $-2X^TW\vec{y} + 2X^TWX\vec{\beta}$

$$\frac{\partial J}{\partial \beta} = \frac{\partial}{\partial \beta}\left(\vec{y}^T - \vec{\beta}^T X^T\right)\left(W\vec{y} - WX\vec{\beta}\right)$$

$$= \frac{\partial}{\partial \beta}\left[\underbrace{\vec{y}^T W \vec{y}}_{0} - \vec{\beta}^T X^T W \vec{y} + \vec{\beta}^T X^T W X \vec{\beta} - \vec{y}^T W X \vec{\beta}\right]$$

$$= -X^T W^T \vec{y}$$

Knowledge Check: Regression - Course 2

aws training and certification

Now, $J(\beta) = \sum_{i=1}^{N} \omega_i \left(y_i - \vec{x_i} \cdot \vec{\beta}\right)^2 = \left(\vec{y} - X\vec{\beta}\right)^T W(\vec{y} - X\vec{\beta})$

What does the expression $\frac{\partial J}{\partial \beta}$ evaluate to?

a. $-2X^T\vec{y} + 2X^TX\vec{\beta}$

b. $-X^TW\vec{y} + X^TWX\vec{\beta}$

c. $-2X^T\vec{y} + 2X^TWX\vec{\beta}$

d. $-2X^TW\vec{y} + 2X^TWX\vec{\beta}$

$$\frac{\partial J}{\partial \beta} = \frac{\partial}{\partial \beta}\left(\vec{y}^T - \vec{\beta}^T X^T\right)\left(W\vec{y} - WX\vec{\beta}\right)$$

$$= \frac{\partial}{\partial \beta}\left[\underbrace{\vec{y}^T W \vec{y}}_{0} - \vec{\beta}^T X^T W \vec{y} + \vec{\beta}^T X^T W X \vec{\beta} - \vec{y}^T W X \vec{\beta}\right]$$

$$= -X^T W^T \vec{y} - X^T W \vec{y} + 2X^T W X \vec{\beta}$$

$$= -2X^T W \vec{y} + 2X^T W X \vec{\beta}$$

Knowledge Check: Regression - Course 2

aws training and certification

Now, $J(\beta) = \sum_{i=1}^{N} \omega_i \left(y_i - \vec{x_i} \cdot \vec{\beta}\right)^2 = \left(\vec{y} - X\vec{\beta}\right)^T W(\vec{y} - X\vec{\beta})$

What does the expression $\frac{\partial J}{\partial \beta}$ evaluate to?

a. $-2X^T\vec{y} + 2X^TX\vec{\beta}$

b. $-X^TW\vec{y} + X^TWX\vec{\beta}$

c. $-2X^T\vec{y} + 2X^TWX\vec{\beta}$

d. $-2X^TW\vec{y} + 2X^TWX\vec{\beta}$

$$\frac{\partial J}{\partial \beta} = \frac{\partial}{\partial \beta}\left(\vec{y}^T - \vec{\beta}^T X^T\right)\left(W\vec{y} - WX\vec{\beta}\right)$$

$$= \frac{\partial}{\partial \beta}\left[\underbrace{\vec{y}^T W \vec{y}}_{0} - \vec{\beta}^T X^T W \vec{y} + \vec{\beta}^T X^T W X \vec{\beta} - \vec{y}^T W X \vec{\beta}\right]$$

$$= -X^T W^T \vec{y} - X^T W \vec{y} + 2X^T W X \vec{\beta}$$

$$= -2X^T W \vec{y} + 2X^T W X \vec{\beta}$$

Knowledge Check: Regression - Course 1

aws training and certification

Given the following three data points, please calculate (by hand) the closed form solution for linear regression and predict the value of y when x = 2

a. $Y = -0.5x + 0.3$

b. $Y = 0.5x + 0.3$

c. $Y = 0.5x - 0.3$

d. $Y = -0.5x - 0.3$

x	y
-1	0.8
0	0.3
1	-0.2

Knowledge Check: Regression - Course 1

aws training and certification

Given the following three data points, please calculate (by hand) the closed form solution for linear regression and predict the value of y when x = 2

a. $Y = -0.5x + 0.3$

b. $Y = 0.5x + 0.3$

c. $Y = 0.5x - 0.3$

d. $Y = -0.5x - 0.3$

x	y
-1	0.8
0	0.3
1	-0.2

$\bar{x} = 0$

$\bar{y} = \frac{0.8 + 0.3 - 0.2}{3}$

$= 0.3$

$\boxed{\hat{y} = ax + b}$

$a = \frac{Cov(X,Y)}{var(X)}$

$= \frac{(-1-0)(0.8-0.3) + 0 + (1-0)(-0.2-0.3)}{1^2 + 0^2 + 1^2}$

$= -1/2$

$b = \bar{y} - a\bar{x}$

$= 0.3 - (-1/2)(0)$

$= 0.3$

Knowledge Check: Regression - Course 1

aws training and certification

Which one of the following is a linear function?

a. $y = 0.2x_1 + 2.0$

b. $y = \sin(0.2x_1 + 2.0)$

c. $y = 2x_1^2 + 5z_2^2$

d. $y = \log(x_1 + x_2)$

e. $y = \sin(x_1) - \cos(x_2)$

Knowledge Check: Regression - Course 1

aws training and certification

Does high R^2 always imply a good fit?

a. True

b. False

aws training and certification

Certificate of Completion

Hem Bahadur Gurung

Has successfully completed

Linear and Logistic Regression

Director, Training and Certification	8 1/2 hours	10 September, 2021
	Duration	Completion Date