AWS Course - Linear and Logistic Regression

https://www.aws.training/Details/eLearning?id=26599

aws training and certification

For this course on linear and logistic regression we join Naumaan Nayyar, AWS Applied Scientist, for discussions on techniques used to make predictions in machine learning. This course focuses specifically on linear models for regression, least squares error, maximum likelihood estimate, regularization, logistic regression, empirical loss minimization, and gradient-based optimization methods.

OLS - Least Squares Minimization

Lesson 2 of 12

Linear Regression via Least Squares Minimization

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

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

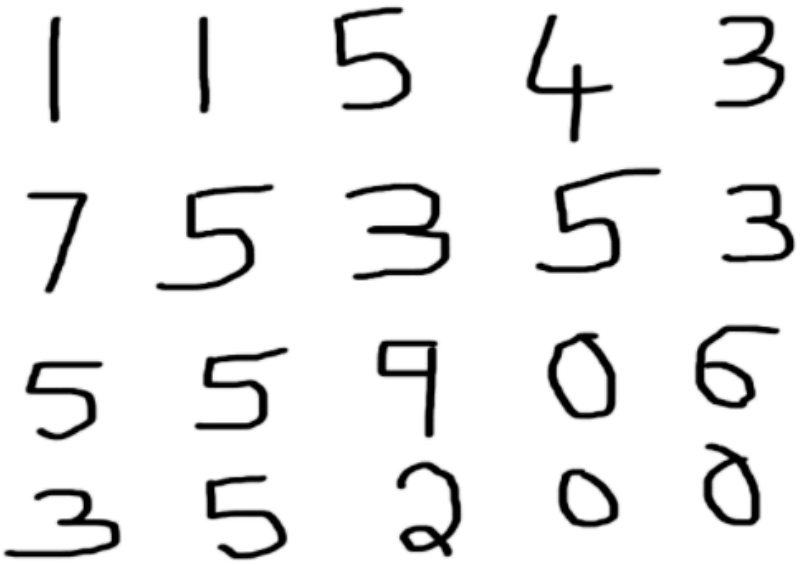
Why use linear models?

- Interpretable
- Low complexity
- Scalable

MNIST Example

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.*



THE CHALLENGE | A NEURAL NETWORK SOLUTION

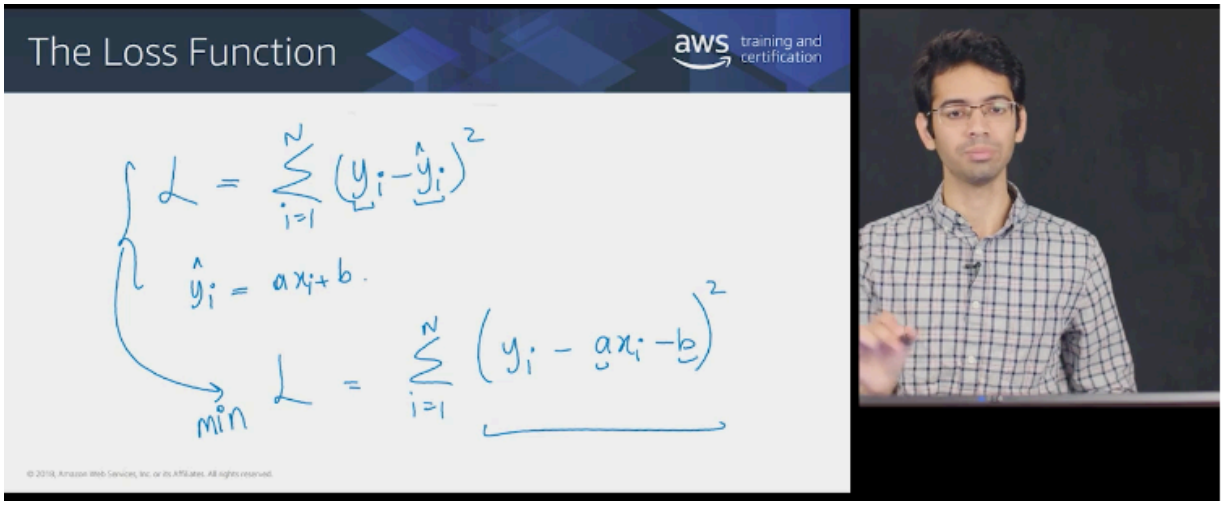
Uses advanced regularization techniques via a neural network

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

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		

Definition of 1-Dimension Ordinary Least Squares (OLS)

Minimize loss function to get the values of a and b



To minimize, take derivative and set to 0

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 and b

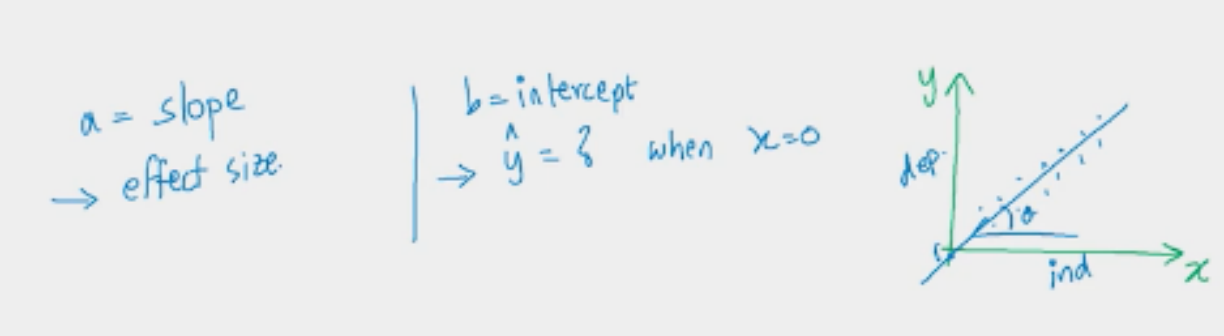
a – slope of the line (size of relationship between X and Y)
b – intercept

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

Interpretation of L

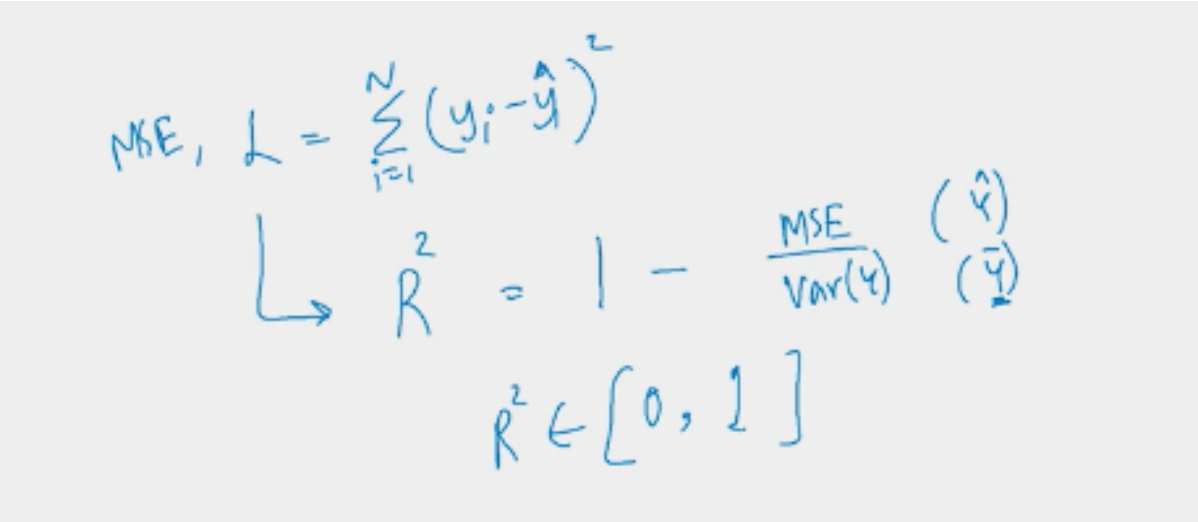
In general, $R^2 \epsilon [0,1]$
Let's observe the model performance with the help of R^2 which is defined as. Keep in mind, L = loss function.

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



L => Loss function
Represented by the R square = 1 - (Mean Square Error) / (variance of y)

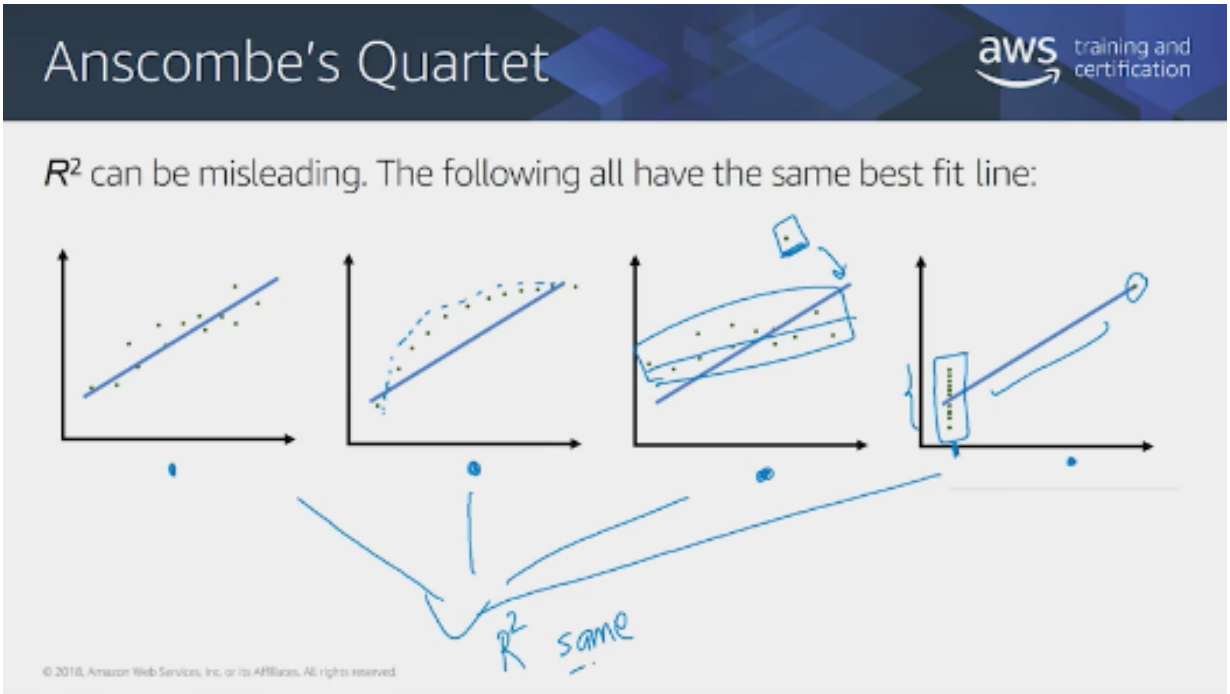
Captures how good is our model compared to a simple baseline model, where predictions are “average” values
R2 should be between 0 and 1
The higher, the better the model fits



Anscombe's Quartet

Anscombe's Quartet

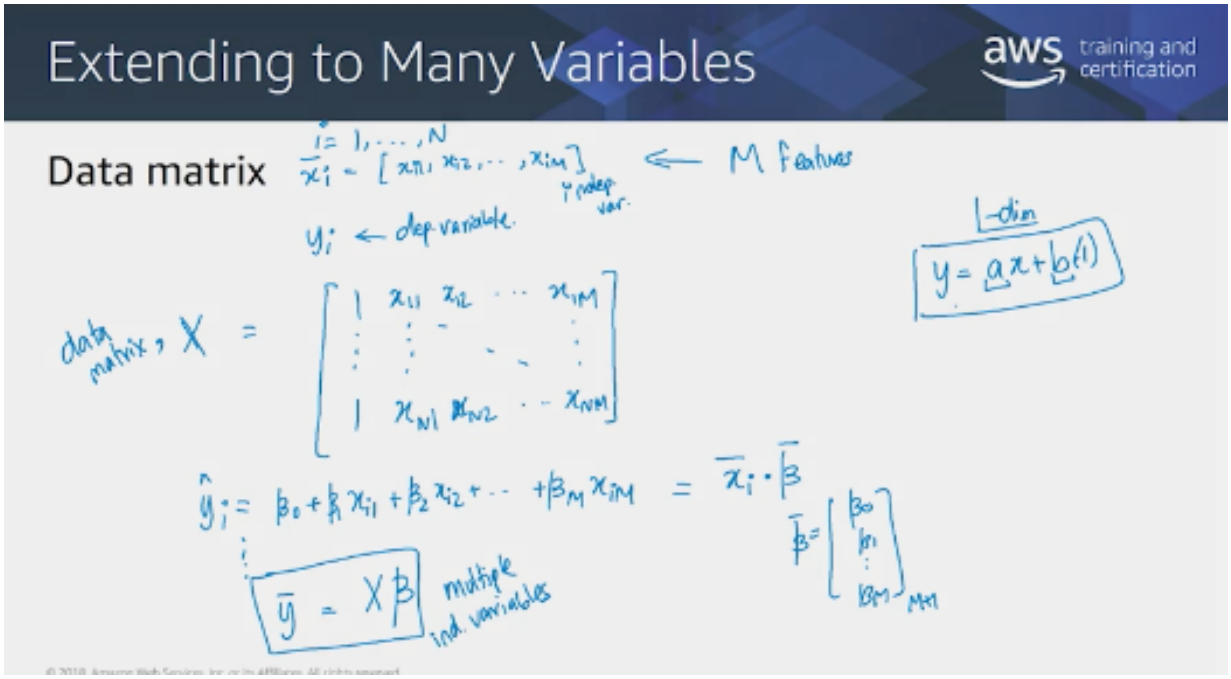
Now we'll use a series of Anscombe's Quartet examples to discover why the R^2 value is not always the best fit to predict y.

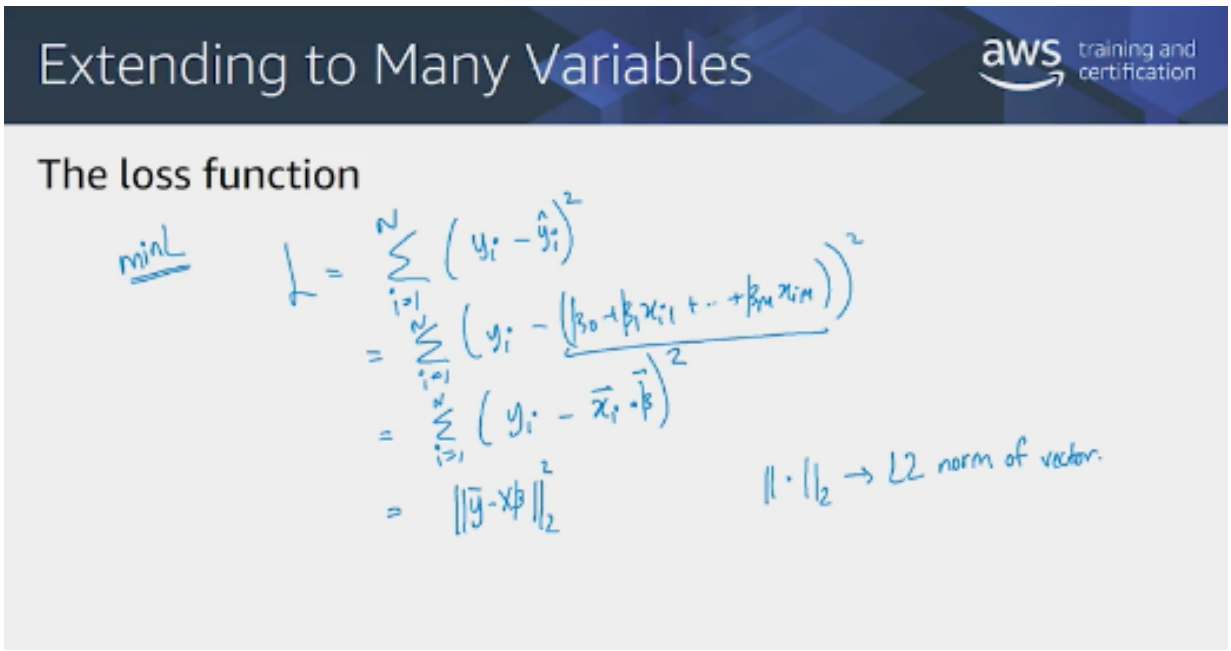


=> we should not use R2 only as a way to evaluate model performance

Multivariate Ordinary Least Squares (OLS)

In this video we discuss transforming one parameter that is dependent on OLS and loss function equations, to suit a higher number of input parameters xi.





Optimizing Loss Function for Multivariate

Optimization

This video shows how to minimize the loss function by using matrix calculations and making the first order derivative value zero.

Ordinary Least Squares (OLS) Pros and Cons

Here, we'll discuss some of the advantages of the OLS method to predict data and its fail cases.

OLS Pros

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

OLS Cons

- Influenced by outliers
- X^TX need not exist

=> need columns of X (or features) to be lineary independent

Lesson 3 of 12

Knowledge Check 1

Q1

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.

x	y
-1	0.8
0	0.3
1	-0.2

- (●) A. Y = -0.5x + 0.3
- ( ) B. Y = 0.5x + 0.3
- ( ) C. Y = 0.5x - 0.3
- ( ) D. Y = -0.5x - 0.3

Correct

Q2

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. $y = \sin(x_1) - \cos(x_2)$

Q2

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. $y = \sin(x_1) - \cos(x_2)$

Correct

Q3

Does high R^2 always imply a good fit?

No - R2 should not be used alone
A high value is a good indicator but some models can fit poorly because of Outliers
Example: Anscombe quartet
Correct

R2 only tells us how much of the variance in the data is explained
It does not tell us weather the model is a good fit for the data or not

Lesson 4 of 12

Linear Regression: A Probabilistic Approach

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?

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

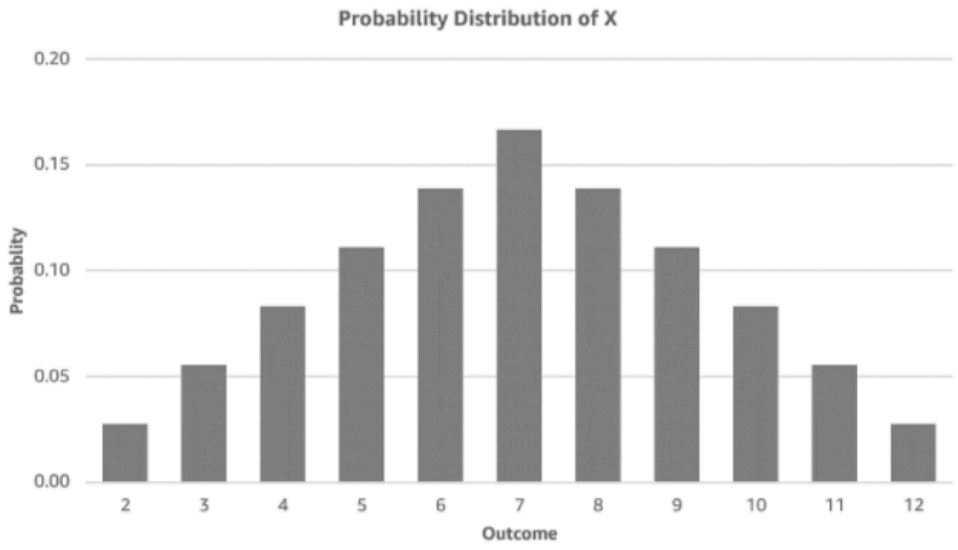
$$\hat{\theta} = \underbrace{arg}_{\theta} \underbrace{max\ L(\theta; Y)}_{\theta \text{ that maximizes L}}$$

Continuous Probability

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

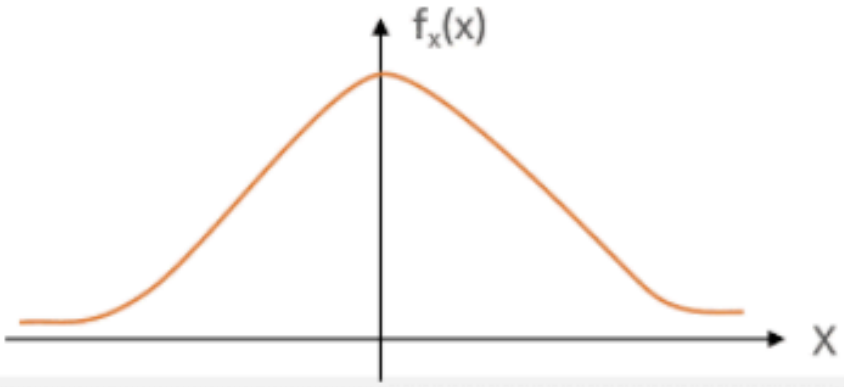
Discrete Random Variables:

Sum of two dice



Continuous Random Variable:

Probability density function

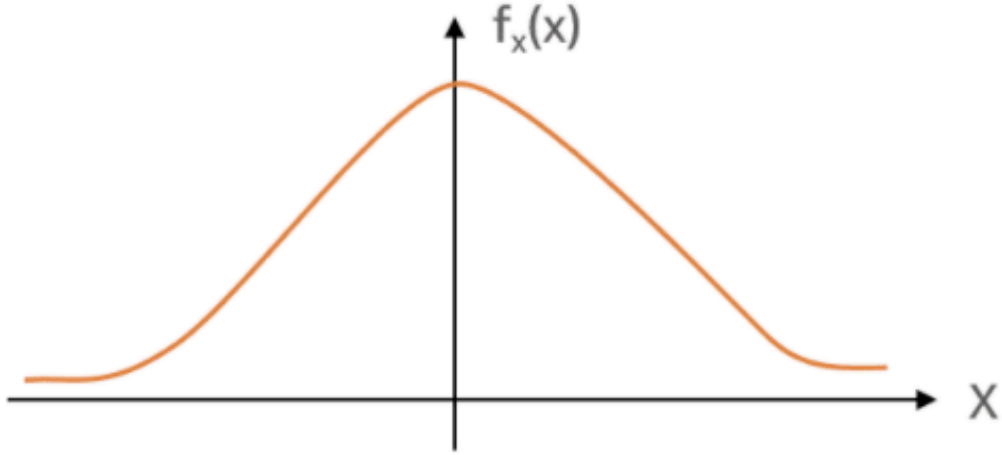


$$X \sim Gaussian\left(\mu, \sigma^2\right) \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

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

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

Gaussian Random Variables

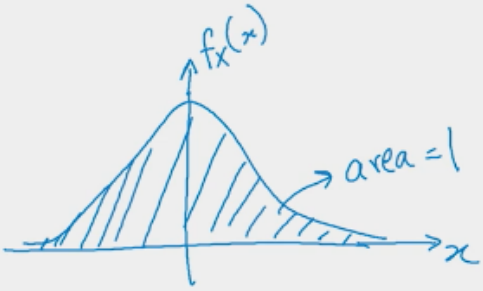


$$X \sim Gaussian\left(\mu, \sigma^2\right) \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

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

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

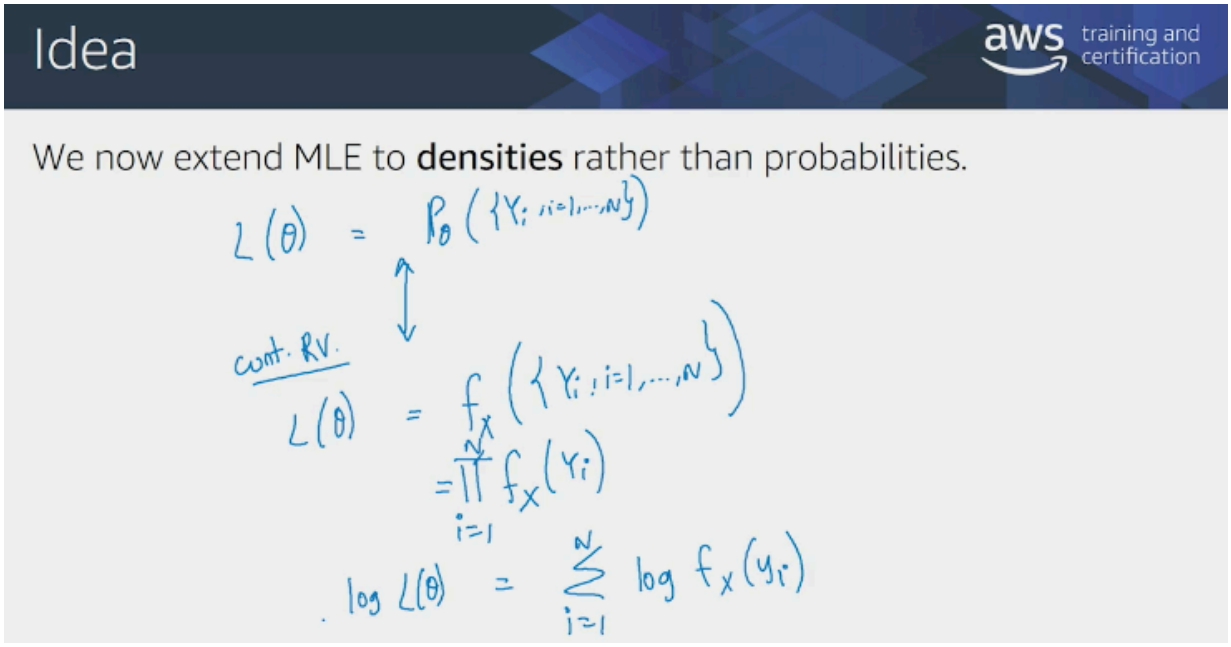
μ ↑ mean

σ ↑ standard deviation

(σ^2 ← variance)

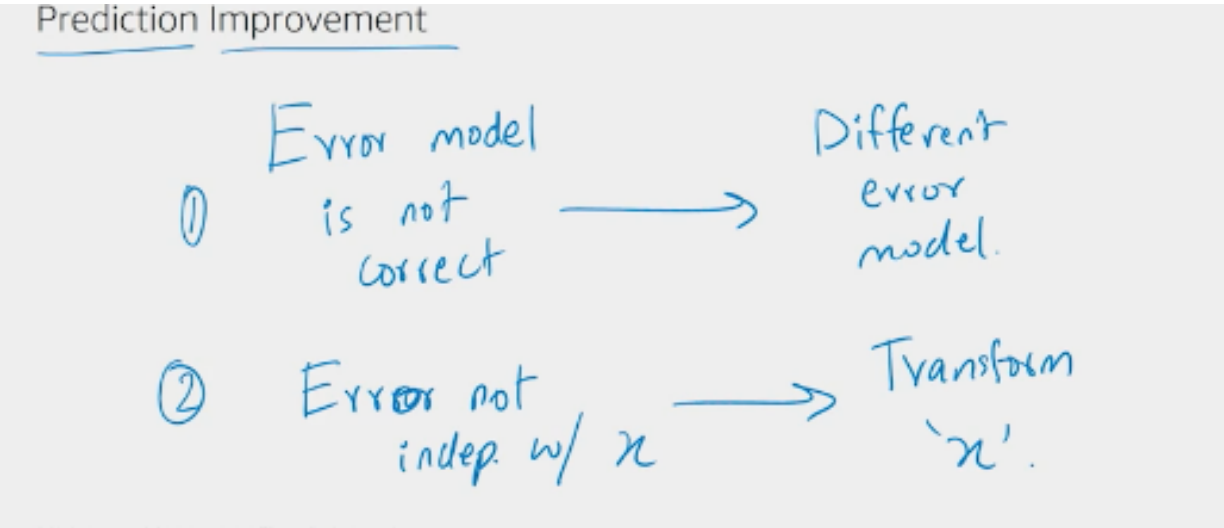
Gaussian Maximum Likelihood Estimate (MLE)

In this video, we'll explore how to derive MLE for the mean of the Gaussian continuous random variable. We'll also learn how maximizing MLE and minimizing loss function have the same result.



Improvements in Maximum Likelihood Estimate (MLE)

In this last video of Linear Regression: A Probabilistic Approach, you will note improvements in the MLE model by observing its performance in fail case scenarios (like Anscombe's Quartet) of MSE. Let's take a look at some of these improvements.



Q1

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?

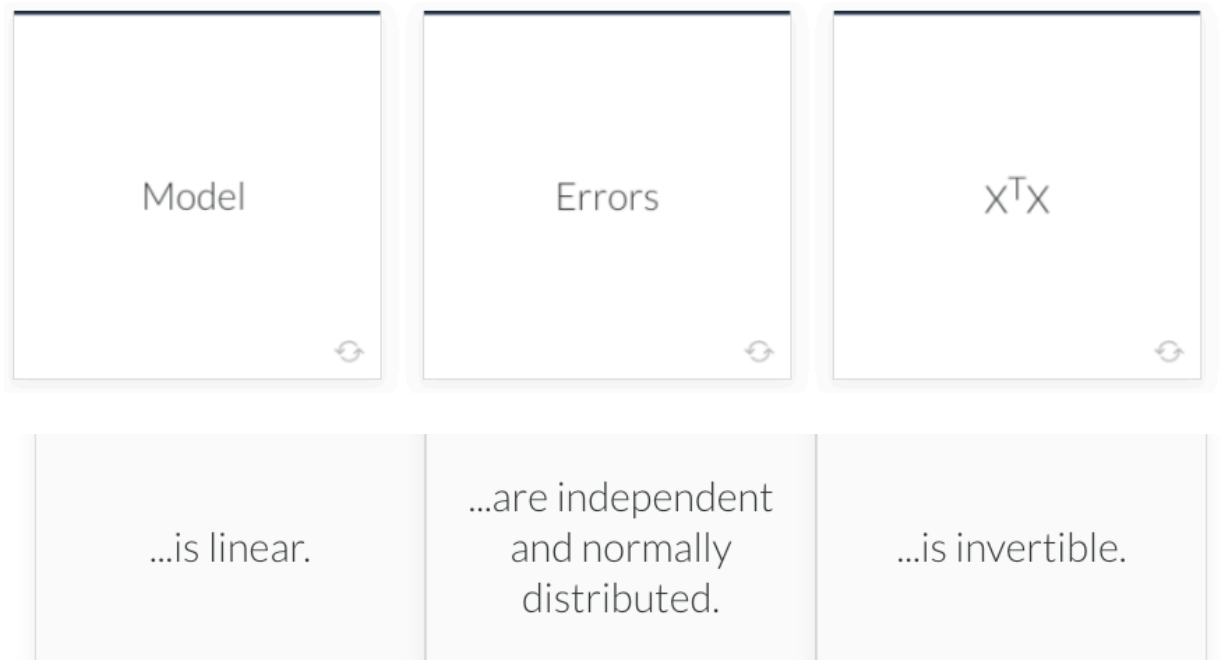
- A. 1/6
- B. 3/5
- C. 2/5
- D. 5/6

Lesson 6 of 12

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

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:



XtX: Columns of X are lineray independent

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

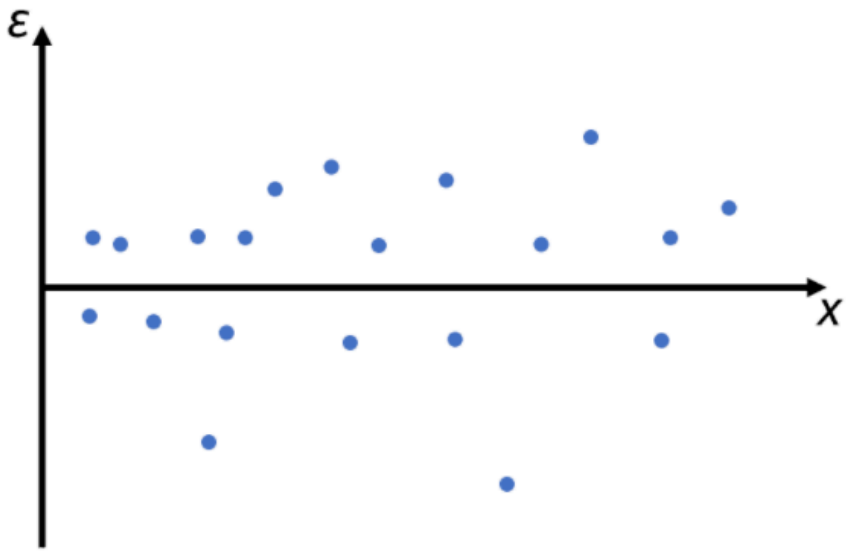
$$\epsilon_i = y_i(x) - \hat{y}_i$$

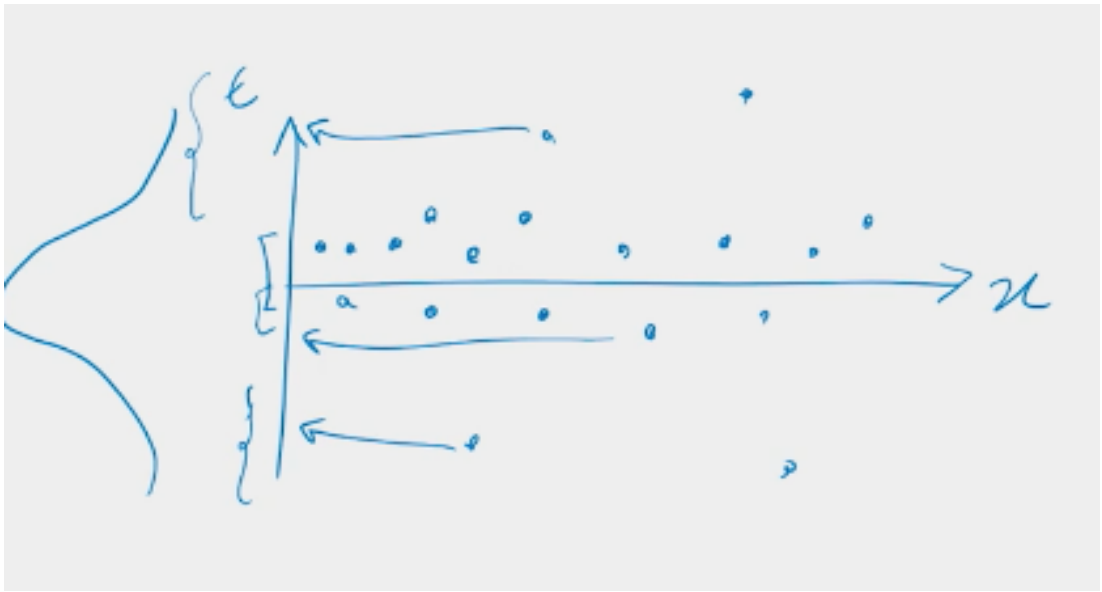
Assumptions:

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

A good residual plot

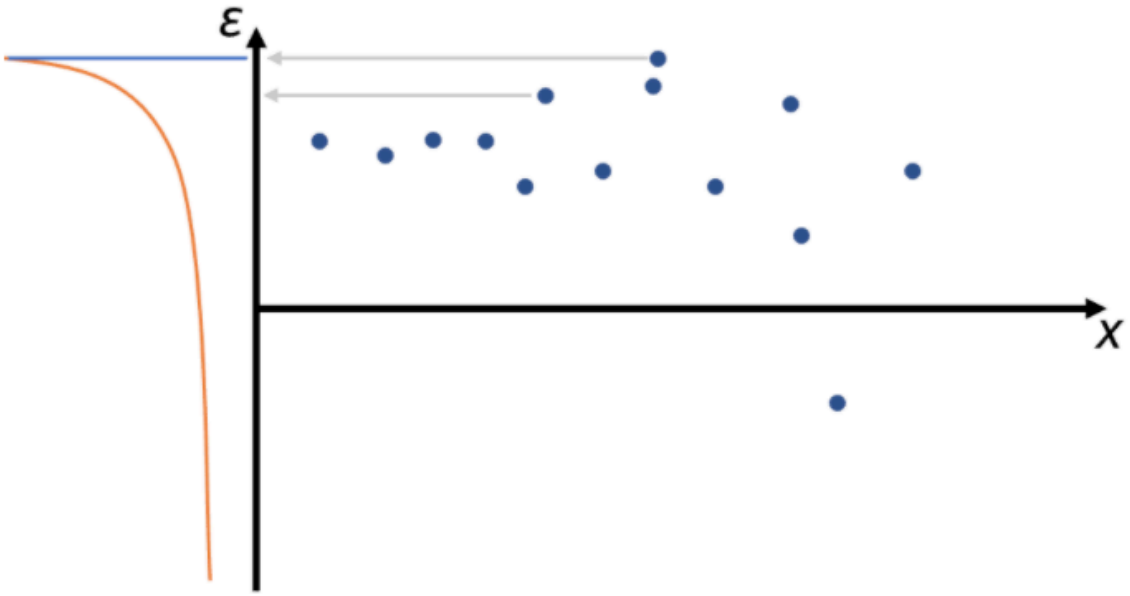
Every creative endeavor requires that you take risks. If you try and don't succeed, you've still learned something. You're not failing. You're discovering what doesn't work.





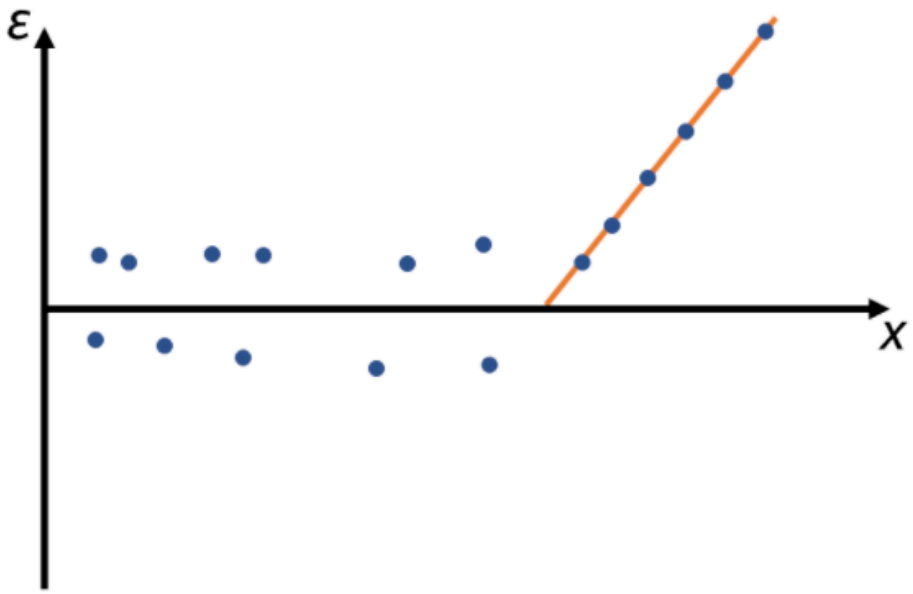
A bad residual plot

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.



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:

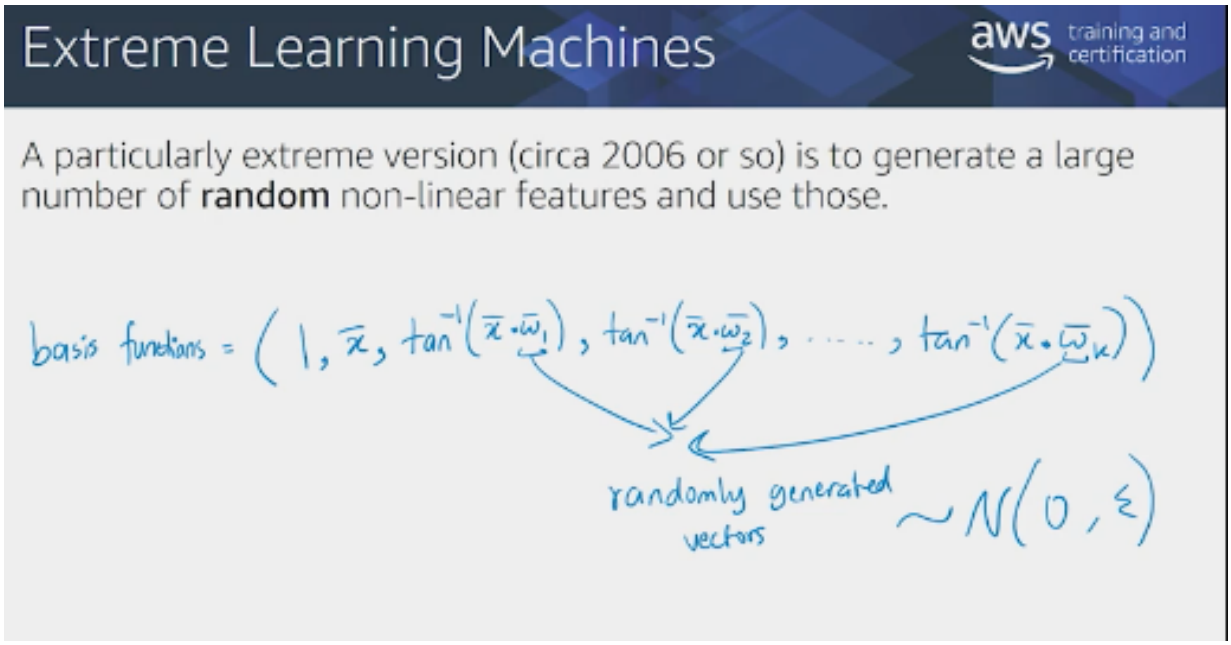
① polynomial $= \left(1, x, x^2, x^3, \dots, x^K\right)$

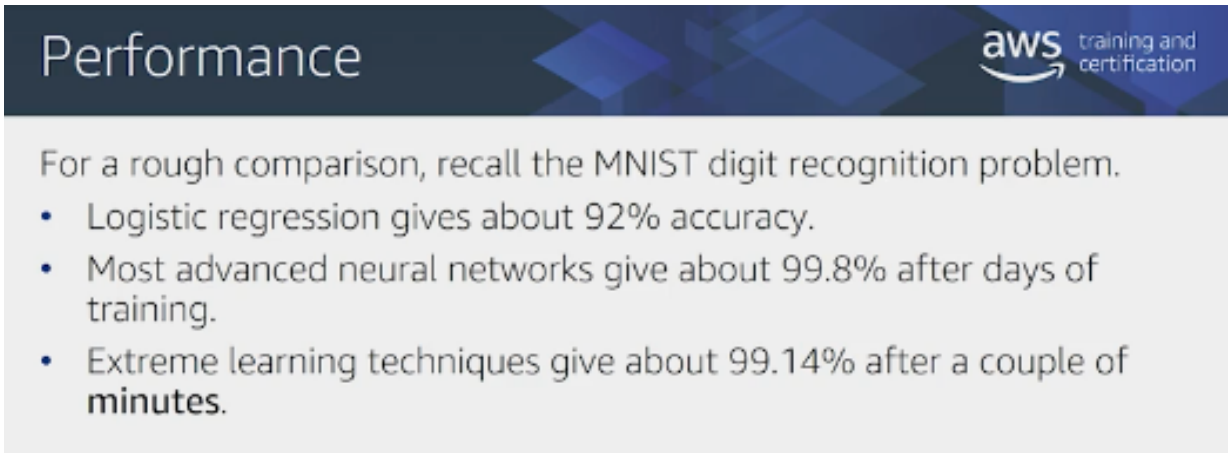
② periodic $= \left(1, \sin(x), \cos(x), \sin(2x), \cos(2x), \dots, \sin(kx), \cos(kx)\right)$

③ kernel method $= \left(1, e^{-\|x-x_1\|^2}, e^{-\|x-x_2\|^2}, \dots, e^{-\|x-x_N\|^2}\right)$

Extreme Learning Machines

Extreme learning machines generate a huge number of random non-linear features and performance comparisons with logistic regression models and advanced neural networks. This video explores extreme learning machines.





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.

Overfitting

You may encounter a model overfitting issue with non-linear regression features because of a higher degree of coefficients. In this video, we discuss using regularization to overcome that issue.

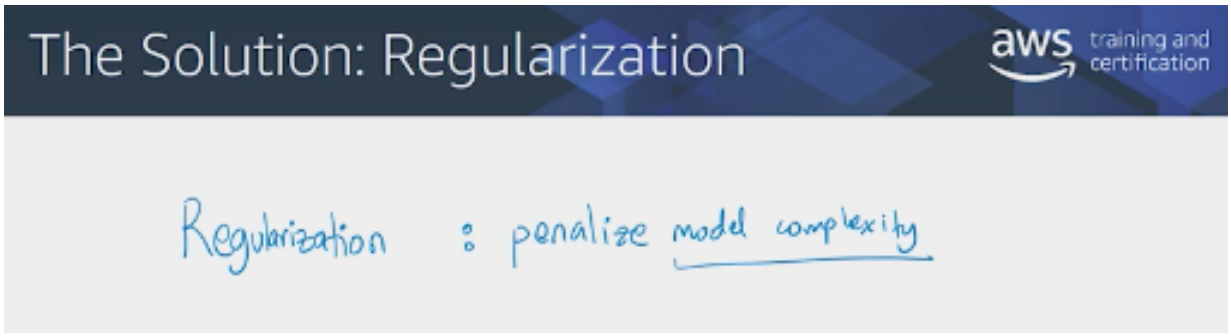
As we grow the polynomial order, the model may start to capture variation between variables too much

Polynomial Regression of Different Degrees

aws training and certification

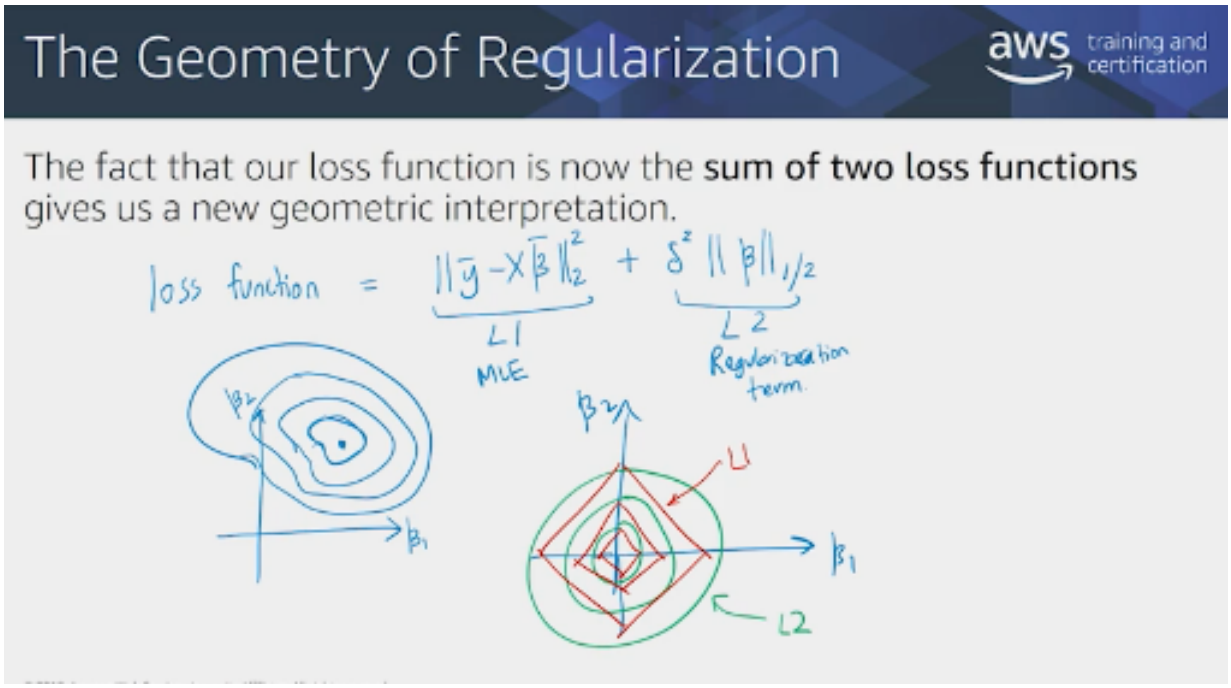
The more complex a model is, the more likely it may overfit

=> Regularization = penalize complexity



Regularization

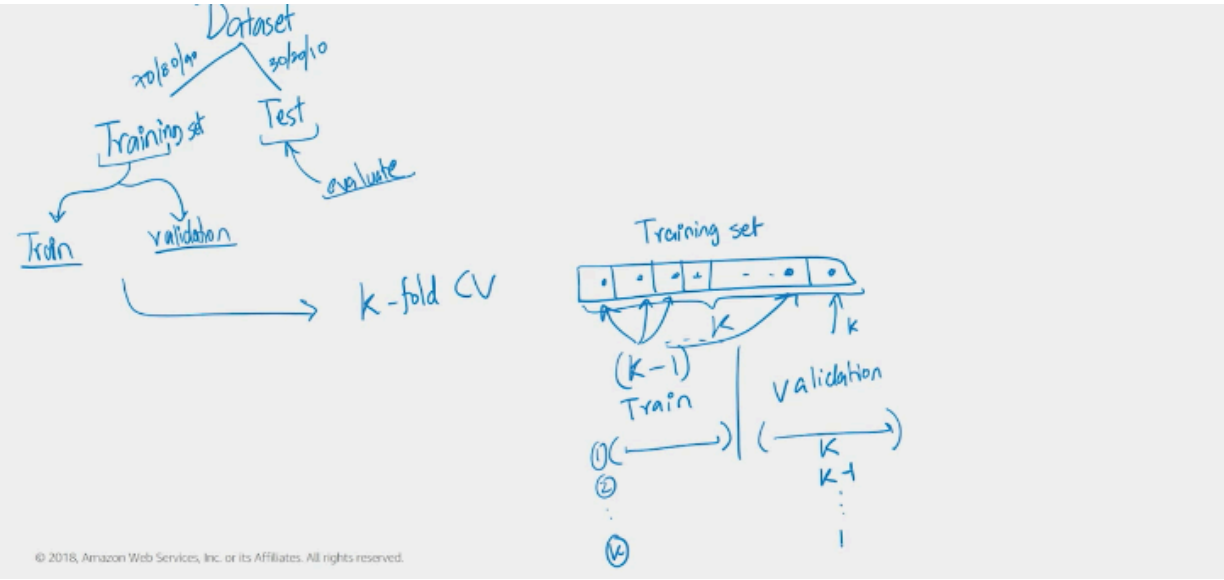
Earlier when you computed the coefficients of linear model (ß) using MSE or MLE, you did not opt any model for ß. This was because you were unaware of the coefficients and didn't know of a model to represent them. Now, we'll explore regularization from the Bayesian point of view, learn how to obtain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



Cross-Validation

The δ that you see in ridge regression and lasso regression isn't really something that machines can learn on their own. This video shows you how to address this challenge using the cross-validation method.

hyper parameter is not learnt automatically by a model
Idea of cross validation:
- take data set and split in 2 sets: training and test set



Alternative Interpretation of Regularization

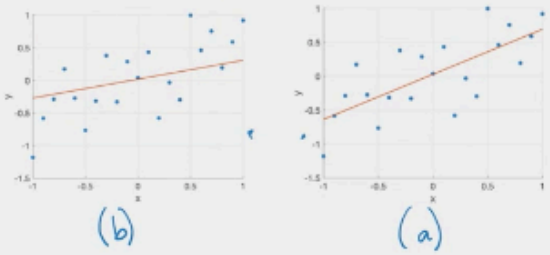
This video provides an overview of the regularization process, and various regularization techniques (ridge, lasso, and elastic regressions).

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



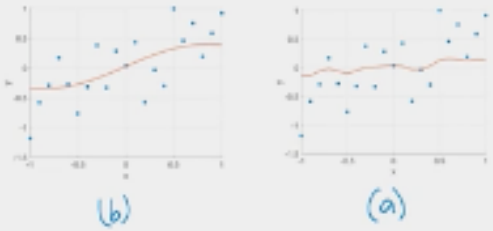
Straight line. Radiant Bayesian Function don't model linear models
Higher penalty in part 2 => slope closer to 0

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



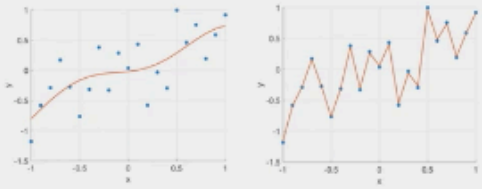
C and d can be eliminated as they are linear
greater value of lambda gives us smoother lines

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



B and d represent linear relationship -> eliminate
higher values of lamda give us smoother function with less oscillation

Lesson 8 of 12

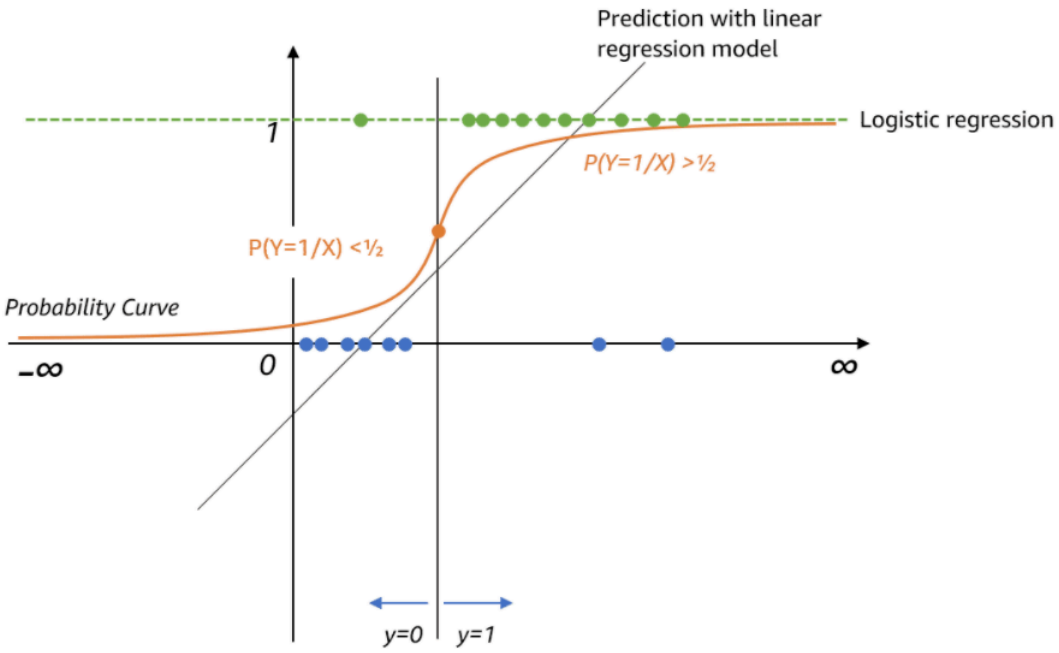
Classification and Logistic Regression

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.

Logistic Regression

In this video, we discuss the continuity and unbounded issues that resulted in the OLS models. We also cover logistic regression based on probability estimates, and the link function which helps users transition into logistic regression. Take a moment to review the image below before watching the video.



Increasing function logistic regression

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).

Example:

Logistic problem:

Take Linear model $\vec{Y} = X\vec{\beta}$

Apply link model (transformation) $\hat{y}_i = \Phi(\vec{x}_i . \vec{\beta})$

Finding full-log likelihood:

Maximum likelihood function

$$L = \prod_{i=1}^{N} P(Y = y_i | \vec{x}_i)$$

$$\{log(L)\}_{max} = \sum_{i=1}^{N} [y_i \, log \, \Phi(\vec{x}_i . \vec{\beta}) + (1 - y_i) \, log(1 - \Phi(\vec{x}_i . \vec{\beta})]$$

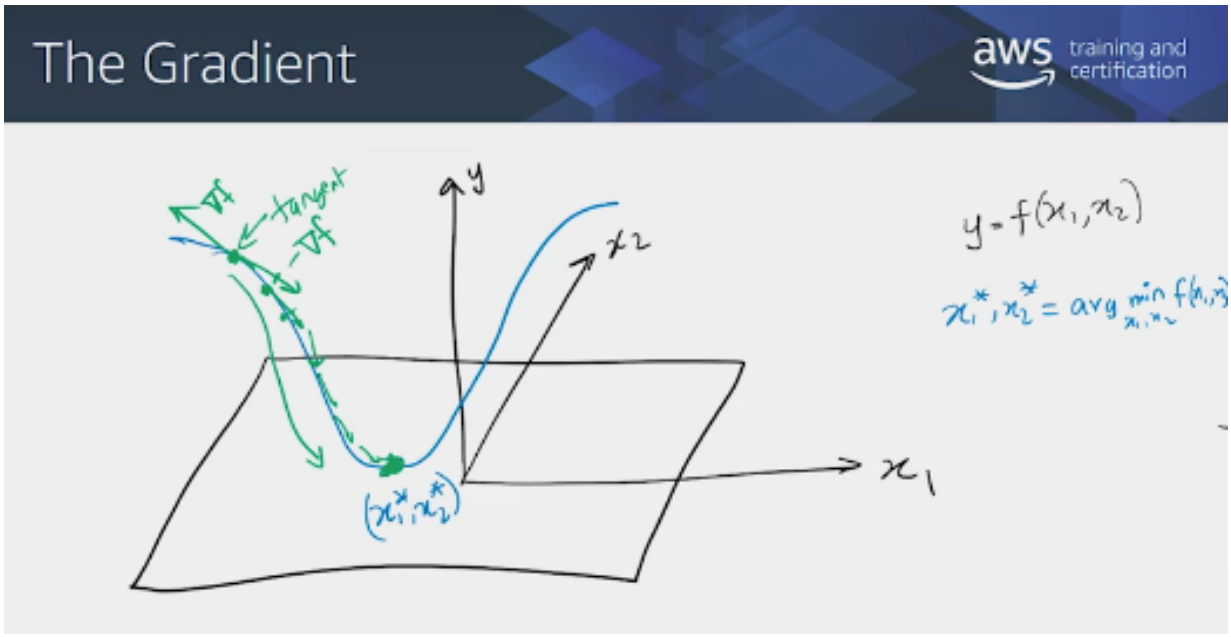
Like OLS method here we find $\{-log(L)\}_{min}$

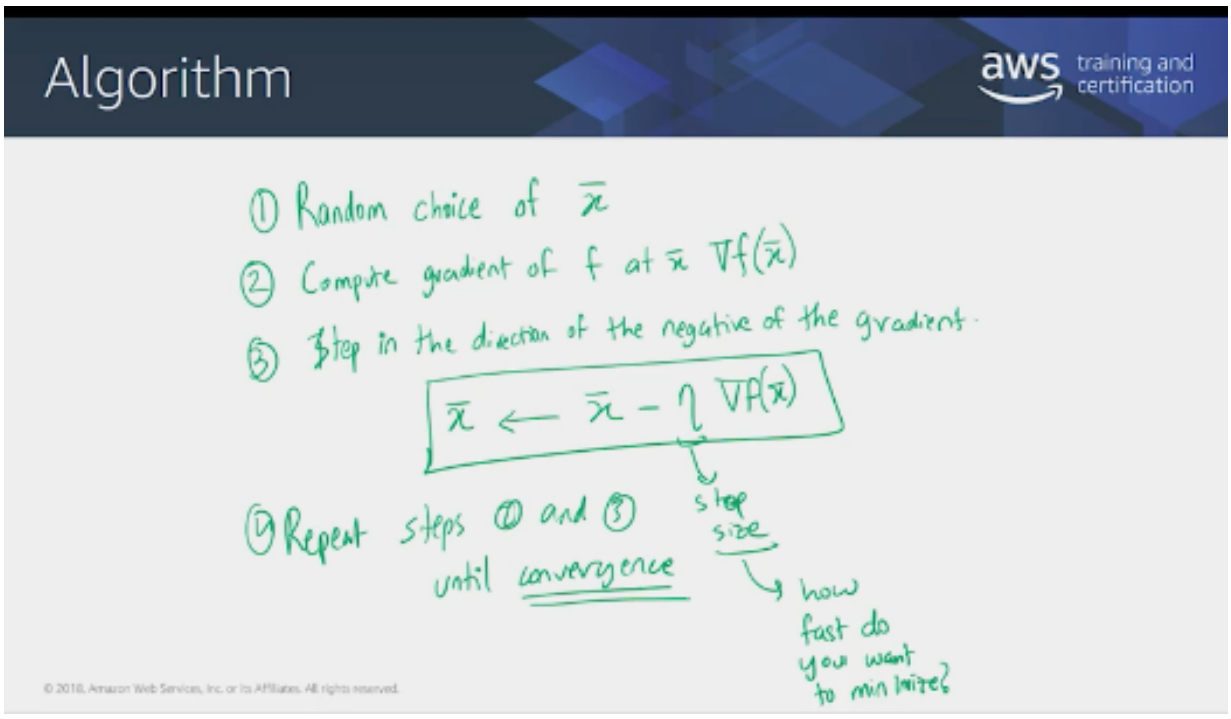
Minimization:

Minimizing $\{-log(L)\}_{min}$ doesn't have any explicit solutions so we have to minimize it numerically

Batch/Full Gradient Descent

In this video, we explore how to geometrically understand reaching the minimum of a function in the gradient descent method by moving in the – ∇f direction.





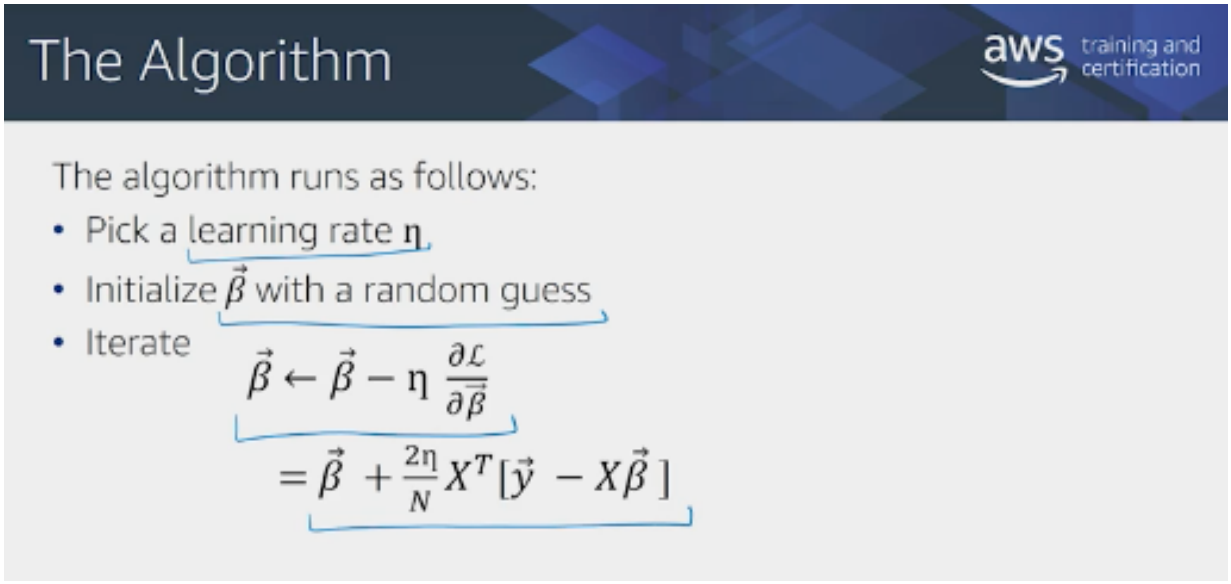
Gradient descend step size => how quickly we want to learn, how fast e want to

approach the minima

- too low: can take a lot of time to get to the minima
- Too high: chance we can overshoot the minima

Gradient Descent for Ordinary Least Squares (OLS)

Now you will find the gradient descent for OLS by computing the numerical minimization of the mean log-likelihood, and then formulating the learning algorithm with the learning rate η.



Gradient Descent for Logistic Regression

In this video, we'll learn how to compute the gradient descent of loss function for logistic regression. We'll do this by finding separate gradients for individual terms in the loss function, and applying final derivatives.

All Terms

aws training and certification

$$\mathcal{L}(\vec{\beta}) = -\sum_i [y_i \log(\phi(\vec{x}_i \cdot \vec{\beta})) + (1 - y_i) \log(1 - \phi(\vec{x}_i \cdot \vec{\beta}))]$$

$$\frac{\partial \mathcal{L}(\vec{\beta})}{\partial \vec{\beta}} = -\sum_i \left[y_i x_i^T \left(1 - \phi(\vec{x}_i \cdot \vec{\beta})\right) + (1 - y_i)(-x_i^T)\, \phi(\vec{x}_i \cdot \vec{\beta}) \right]$$

$$= -\sum_i \left[y_i x_i^T - y_i x_i^T \phi(\vec{x}_i \cdot \vec{\beta}) + x_i^T y_i \phi(\vec{x}_i \cdot \vec{\beta}) - x_i^T \phi(\vec{x}_i \cdot \vec{\beta}) \right]$$

$$= -\sum_i x_i^T \left(y_i - \phi(\vec{x}_i \cdot \vec{\beta})\right)$$

$$= -X^T \left(\vec{y} - \Phi(X\beta)\right)$$

$$\Phi \rightarrow \begin{bmatrix} \phi_1 \\ \vdots \\ \phi_N \end{bmatrix}$$

Summary of Calculations

Here, you'll review the gradient descent equation of logistic regression and define the features model based on the normalized logistic gradient.

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**.

The Algorithm

aws training and certification

The algorithm runs as follows. Pick a learning rate n.

- Initialize $\vec{\beta}$ with a random guess.
- Iterate

$$\vec{\beta} \leftarrow \vec{\beta} - \eta \frac{\partial \mathcal{L}}{\partial \vec{\beta}} = \vec{\beta} + \frac{\eta}{N} X^T [\vec{y} - \Phi(X\vec{\beta})].$$

logistic regression using GD

Comparison with Ordinary Least Squares (OLS)

In this video we compare the expressions of gradient descent for OLS and gradient descent for logistic regression, which are similar and have almost the same kind of regularization.

Regularization

aws training and certification

OLS: Regularization, $L = \|\vec{y} - X\vec{\beta}\|_2^2 + \delta^2 \|\beta\|_{2,1}^{2/1}$

Logistic: Reg, $L = \mathcal{L}(\beta) + \delta^2 \|\beta\|_{2,1}^{2/1}$

(L2) Ridge ← uniform penalization

logistic loss function

(L1) LASSO ← dimensionality reduction

Beyond Two Types

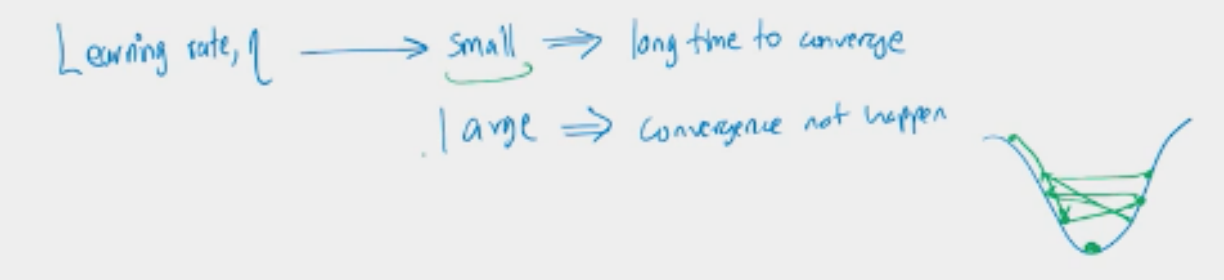
In general, you will encounter more than two classifications of data. In this video, we make the regression suit a higher number of classifications with the help of SoftMax function, and then compare logistic regression and SoftMax in terms of their probabilities.

Summary

This video includes a summary of what was covered in this section. You should now have an understanding of regression models and practical tricks for using training algorithms.

Importance of:
Learning Rate

- small values => take a long time to converge
- large values => can make convergence not happen at al



=> Important to monitor this loss function during gradient descent and make sure it goes down.
=> also important to normalize properly => divide by N (features)

Gradient descent applies to each framework the same: OLS, Logistic or SoftMax

Lesson 10 of 12

Optimization Techniques

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.

Gradient Descent Algorithm

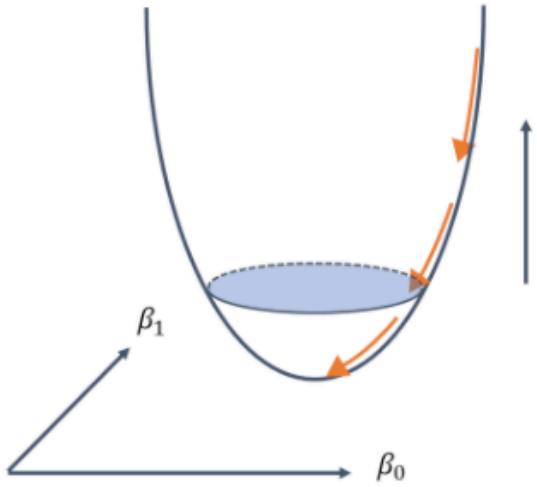
- works by initially generating random parameters, and generating a fixed step size

In each iteration, recompute values of coefficient parameters
In direction of steepest descent
When gradient = 0 => finished covering

Gradient descent

Generate a random parameter $\vec{\beta}$

Iterate: $\vec{\beta} \leftarrow \vec{\beta} - \eta \frac{\partial L}{\partial \vec{\beta}}$



Other Optimization Methods

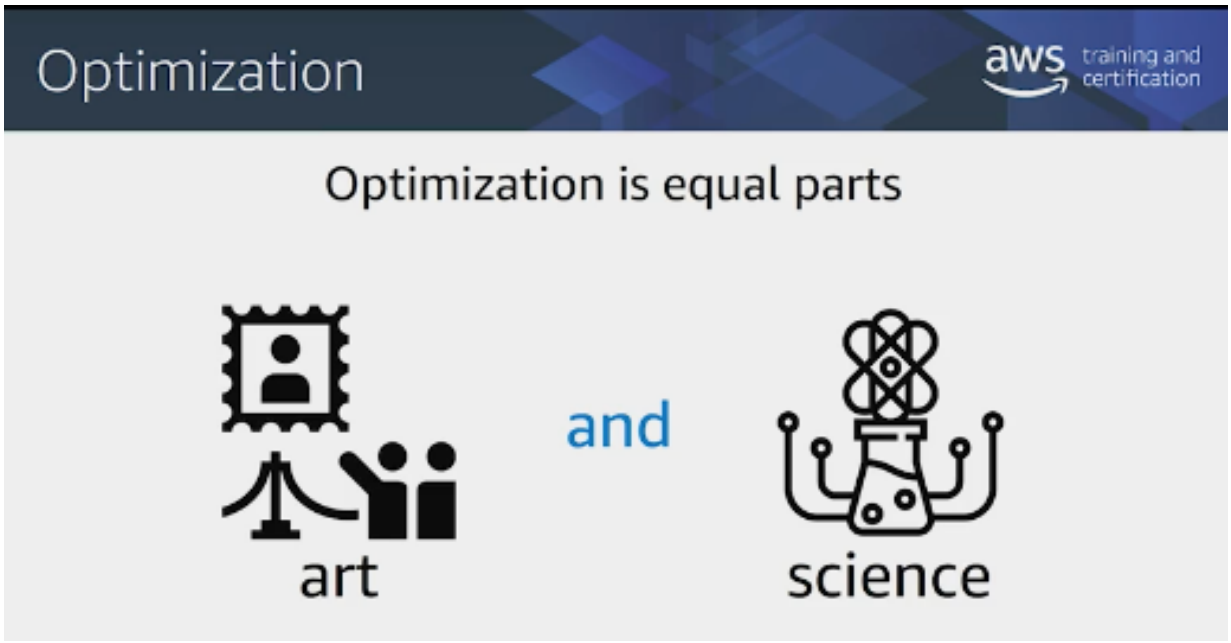
In this video, we briefly discuss other optimization models that are reliant on the gradient descent technique but have a modification.

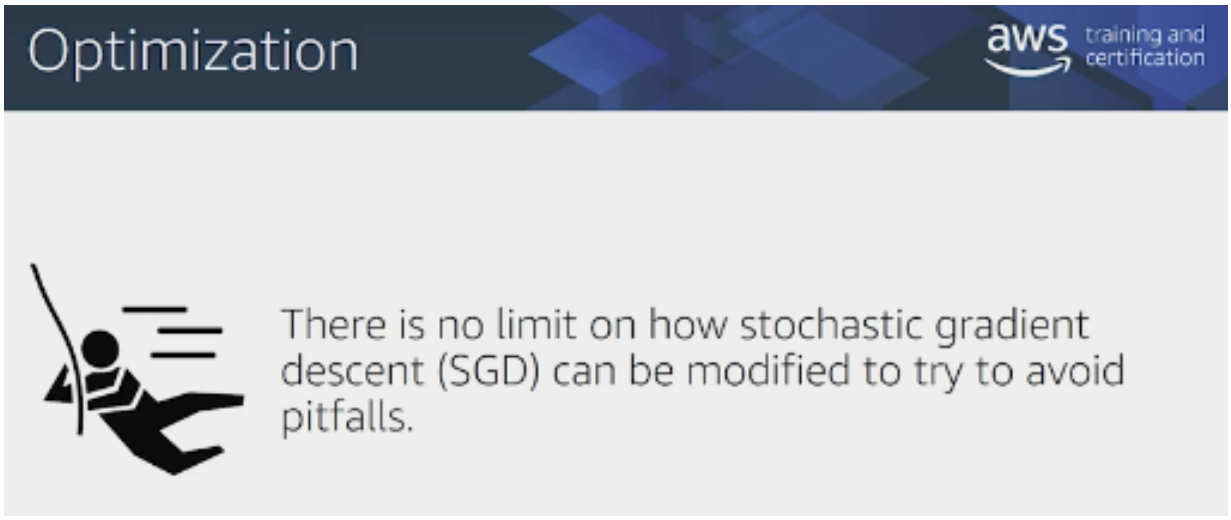
Optimization techniques based on gradient descent:

- 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.

Optimization in Practice

You may wonder which optimization method should be used for a given data set. There are no set guidelines available to help answer this question; instead, you will need to practice various optimization techniques for different data sets, comparing the results, and learning what works best for your unique business challenges.





Optimization

aws training and certification

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

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.