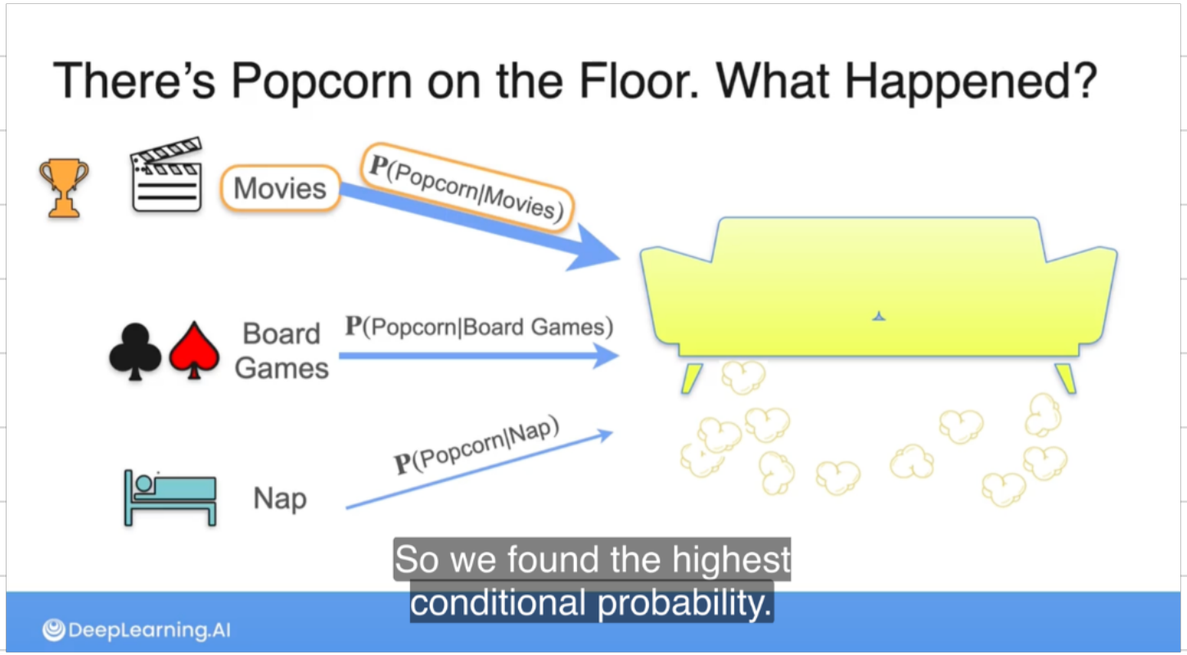
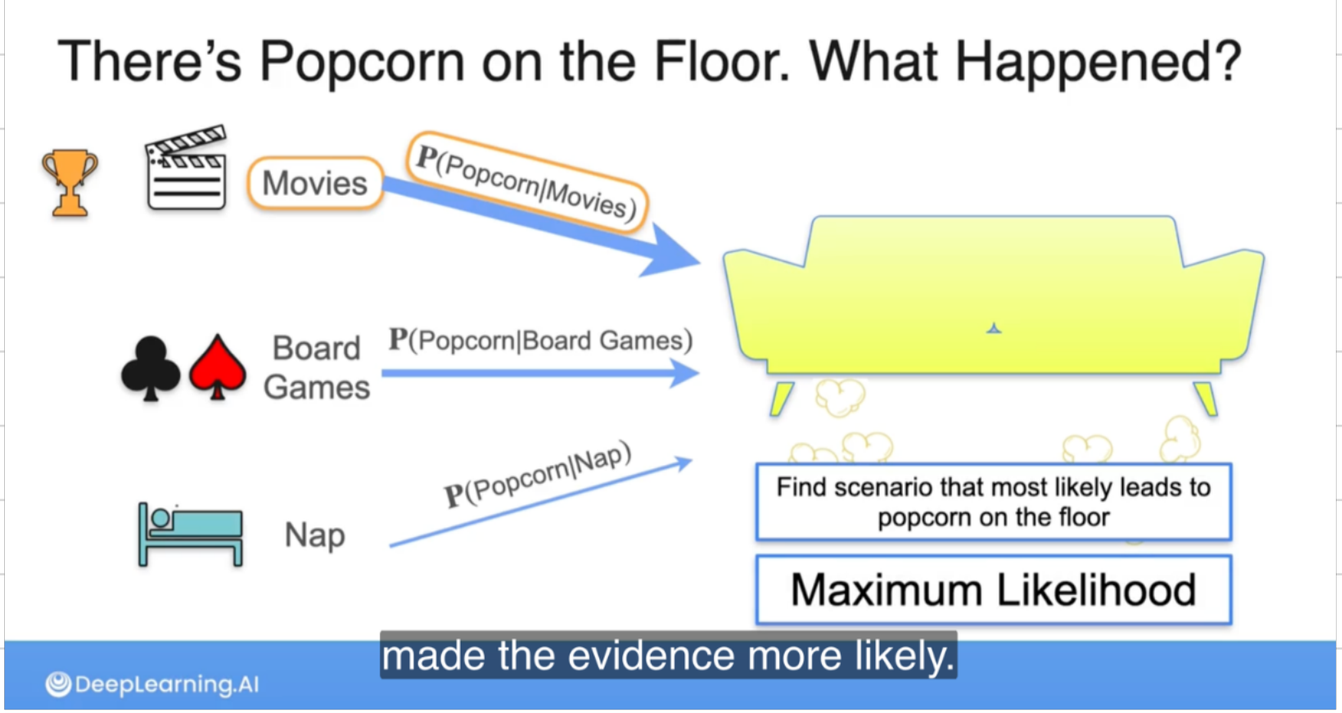
Day - 63, Feb 1, 2025 (Magh 19, 2081)

Point Estimation, Maximum Likelihood Estimation (MLE).

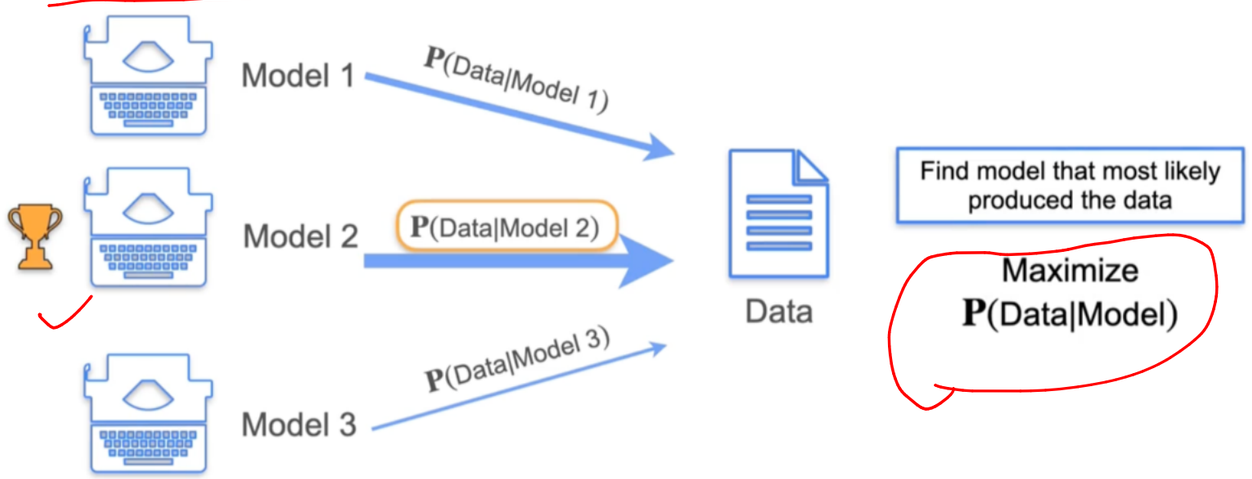
MLE is widely used in machine learning to train models, but the concept behind MLE is very simple. Imagine that you have some evidence of something and you want to find the scenario that may have led to that evidence so you pick out of all the possible scenarios, the one that created the evidence with the highest probability. Let me show you this with an example. Imagine that you walk into a living room and you see a bunch of popcorn lying on the floor next to the couch. Here's a question for you. Which one of these events is more likely to have happened? People watching a movie, people playing board games, or somebody taking a nap. Which one do you think was more likely to have happened? Well, let's try to look at what led to popcorn with the highest probability. The probability that popcorn was on the floor after watching movies is high. For board games, it's medium and for taking a nap, it's low because taking a nap does not produce popcorn in the floor. So we're going to go for whatever created the popcorn with higher probability, and that's movies. So we're going to infer that the most likely thing that happened is that people were watching movies. What we did was we maximize the conditional probability because there's a probability of popcorn given movies and that's high, then the probability of popcorn given board games, which is medium, and the probability of popcorn given nap, which is low. So we found the highest conditional probability. What we did, in other words, is to find the scenario that most likely leads to popcorn in the floor. This is called maximum likelihood. We picked the scenario that made the evidence more likely. This is actually what's done in machine learning many times. You have a bunch of data and several models that could have generated that data. What you do is you estimate the probability that we see this data given Model 1. The probability that we see this data given Model 2 and given Model 3 and whichever gives the highest probability is the model that we pick, the model that most likely produced the data. In other words, we're maximizing probability of data given model. Now why does linear regression fit in here? We're going to get into more details later. But to give you a vague idea, imagine that you have this as your data points and three possible models and imagine that we had a way to generate points based on a line, and the points would be generated close to the line. So there's a probability that the points appear based on Model 1, based on Model 2, and based on Model 3 and the model that gives these points the highest probability is the model that we're going to pick. But again, we're going to see this in more detail later.





So, in this case popcorn on the floor is due to watching movies event.

Maximum Likelihood

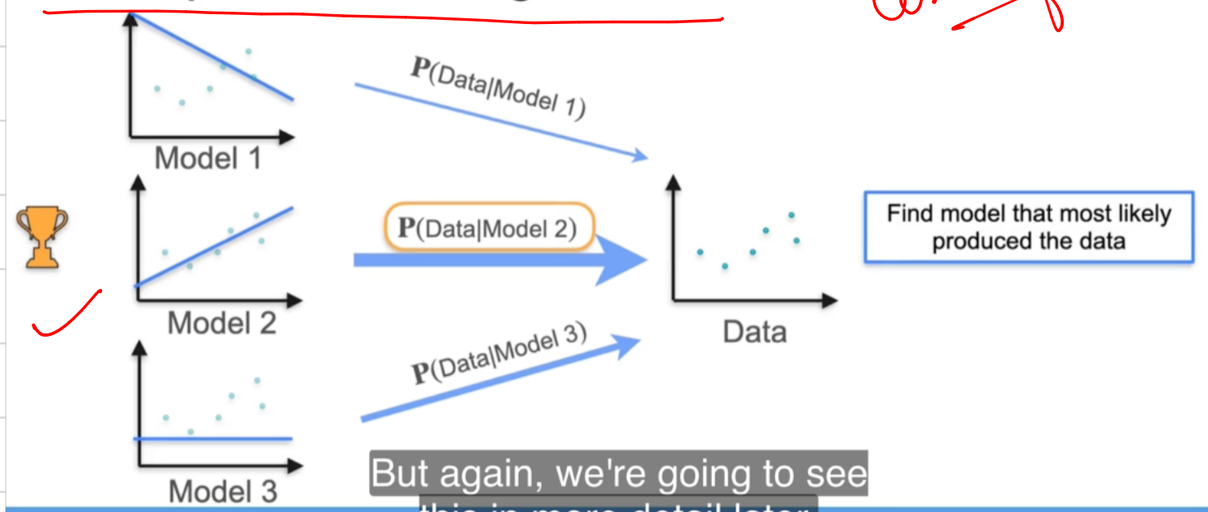


probability of data given model.

DeepLearning.AI

Example: Linear Regression

In the Case of LReg:



But again, we're going to see this in more detail later.

DeepLearning.AI

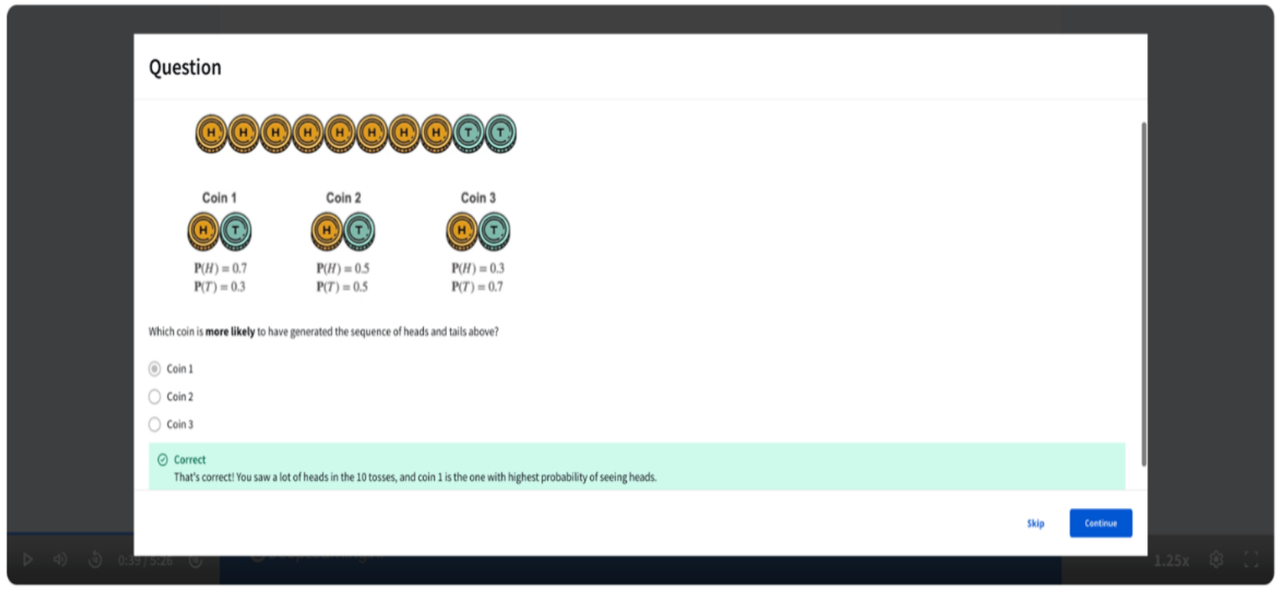
https://www.coursera.org/learn/machine-learning-probability-and-statistics/lecture/i4dJz/mle-bernoulli-example

Welcome To Colab... | Gmail | Dilli822 (Dilli Hang... | Convolutional Neu... | NLP - RNN, Visual... | Nepali Calendar 2... | Analytics | Home | Fifth Annual Nepal... | Google Docs | Other Favorites

coursera | Search in course | Search | English

Probability & Statistics for Machine Learning & Data Science > Week 3 > MLE: Bernoulli Example

Previous Next



MLE: Bernoulli Example

Save note

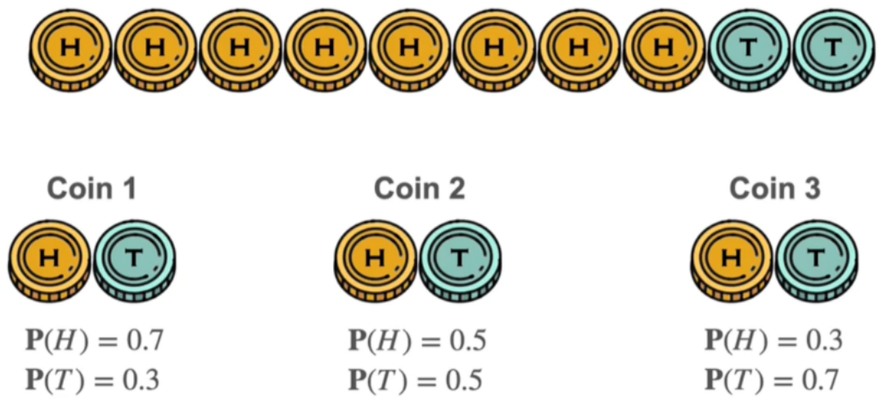
Transcript | Notes | Downloads

Let's go back to the coin example. Suppose that you toss a coin ten times and it lands in heads eight times, and in tails twice. Now you have three possible coins that you could have flipped to

Transcript language: English

coursera.org - To exit full screen, press esc

Maximum Likelihood: Bernoulli Example



10 tosses again?

DeepLearning.AI

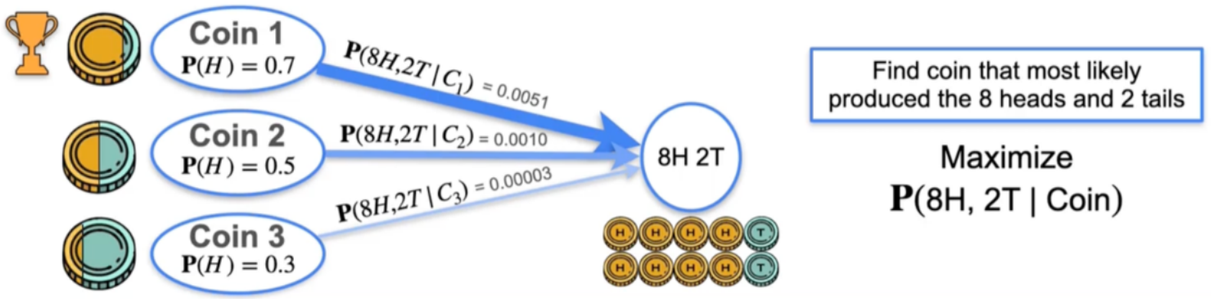
Maximum Likelihood: Bernoulli Example

Coin 1	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.3	0.3 = 0.0051
Coin 2	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5 = 0.0010
Coin 3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.30	0.7	0.7 = 0.00003

should probably go with coin one as the most likely one to have generated this.

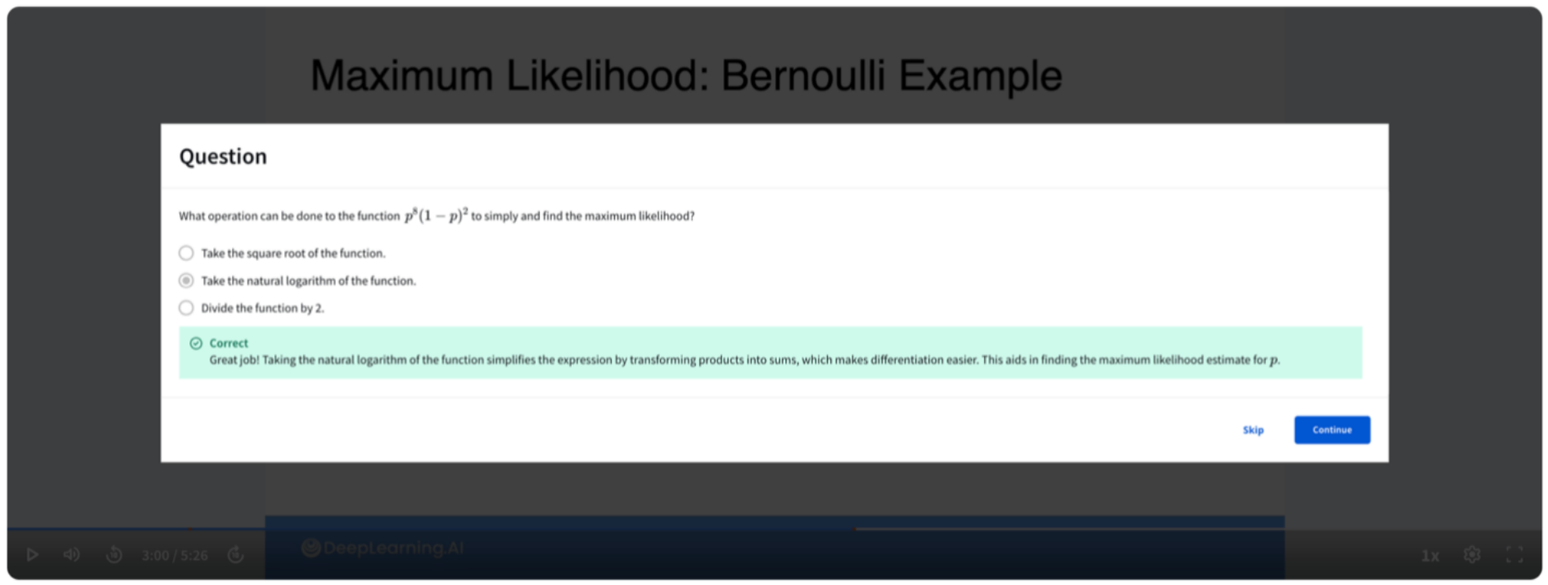
DeepLearning.AI

Maximum Likelihood: Bernoulli Example



DeepLearning.AI

Since log can convert the product terms into sum And make the larger value smaller.



MLE: Bernoulli Example

Maximum Likelihood: Bernoulli Example

$p = \mathbf{P}(H)$ Likelihood $L(p; 8H) = p^8(1-p)^2$ Function of p

You want p that maximizes the chances of seeing 8H

DeepLearning.AI

Maximum Likelihood: Bernoulli Example

$p = \mathbf{P}(H)$ Likelihood $L(p; 8H) = p^8(1-p)^2$ Function of p

You want p that maximizes the chances of seeing 8H

Log-likelihood $\ell(p; 8H) = \log\left((p^8(1-p)^2\right) = 8\log(p) + 2\log(1-p)$

DeepLearning.AI

Maximum Likelihood: Bernoulli Example

$p = \mathbf{P}(H)$ Likelihood $L(p; 8H) = p^8(1-p)^2$ Function of p

You want p that maximizes the chances of seeing 8H

Log-likelihood $\ell(p; 8H) = \log\left((p^8(1-p)^2\right) = 8\log(p) + 2\log(1-p)$

$$\frac{d}{dp}\left(8\log(p) + 2\log(1-p)\right) = \frac{8}{p} + \frac{2}{1-p}(-1) = 0 \rightarrow \hat{p} = \frac{8}{10}$$

DeepLearning.AI

General Purpose?

$\sum_{i=1}^{n} x_i$: # heads

$n - \sum_{i=1}^{n} x_i$: # tails

Maximum Likelihood: Bernoulli Example

n coins
k heads

$$\mathbf{X} = (X_1, \ldots, X_n)$$

$$X_i \overset{i.i.d}{\sim} Bernoulli(p)$$

Likelihood

$$L(p; \mathbf{x}) = P_p(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^{n} p_{X_i}(x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

If $x_i = 1, p^{[x_i]}(1-p)^{[1-x_i]} = p$
If $x_i = 0, p^{[x_i]}(1-p)^{[1-x_i]} = (1-p)$

$$\sum_{i=1}^{n} x_i = \text{\# heads} \qquad n - \sum_{i=1}^{n} x_i = \text{\# tails}$$

Now, if $X_i = 1$, $p^{[x_i]}(1-p)^{[1-x_i]} \Rightarrow p$

if $x_i = 0$, $p^{[x_i]}(1-p)^{[1-x_i]} \Rightarrow (1-p)$

DeepLearning.AI

Likelihood

$$L(p; \mathbf{x}) = P_p(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^{n} p_{X_i}(x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

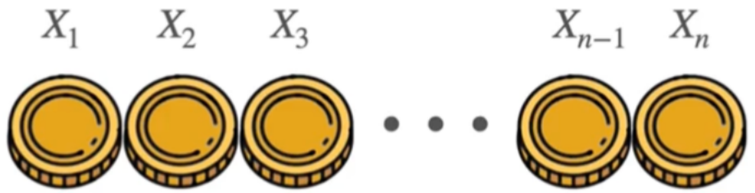
$$h(p; x) = P_p(X = x) = \prod_{i=1}^{n} p_{x_i}(x_i)$$

Log-likelihood

$$\ell(p; \mathbf{x}) = \log\left((p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}\right) = \left(\sum_{i=1}^{n} x_i\right)\log(p) + \left(n - \sum_{i=1}^{n} x_i\right)\log(1-p)$$

$$l(p; x) = \log\left((p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}\right)$$

Maximum Likelihood: Bernoulli Example



$$\mathbf{X} = (X_1, \ldots, X_n)$$

$$X_i \overset{i.i.d}{\sim} Bernoulli(p)$$

Likelihood

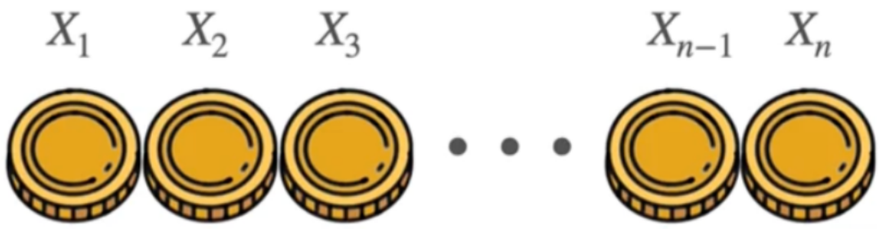
$$L(p; \mathbf{x}) = P_p(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^{n} p_{X_i}(x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\left(\sum_{i=1}^{n} x_i\right)}(1-p)^{\left(n-\sum_{i=1}^{n} x_i\right)}$$

Log-likelihood

$$\ell(p; \mathbf{x}) = \log\left((p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}\right) = \left(\sum_{i=1}^{n} x_i\right)\log(p) + \left(n - \sum_{i=1}^{n} x_i\right)\log(1-p)$$

DeepLearning.AI

Maximum Likelihood: Bernoulli Example



$$\mathbf{X} = (X_1, \ldots, X_n)$$

$$X_i \overset{i.i.d}{\sim} Bernoulli(p)$$

$$\ell(p; \mathbf{x}) = \log\left((p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}\right) = \left(\sum_{i=1}^n x_i\right)\log(p) + \left(n - \sum_{i=1}^n x_i\right)\log(1-p)$$

Find the maximum!

$$\frac{d}{dp}\ell(p; \mathbf{x}) = \frac{d}{dp}\left(\left(\sum_{i=1}^n x_i\right)\log(p) + \left(n - \sum_{i=1}^n x_i\right)\log(1-p)\right)$$

$$= \frac{\sum_{i=1}^n x_i}{p} + \frac{n - \sum_{i=1}^n x_i}{1-p}(-1) = 0 \quad \Rightarrow \quad \hat{p} = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}$$

DeepLearning.AI

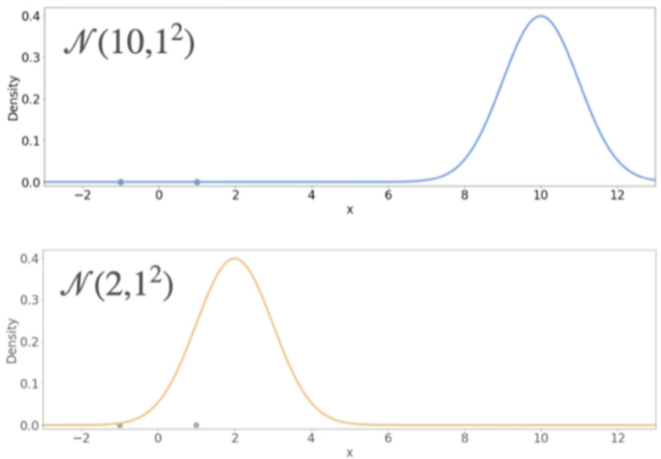
Bernoulli: Only one Event Both

So, $\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} \Rightarrow \bar{x}$

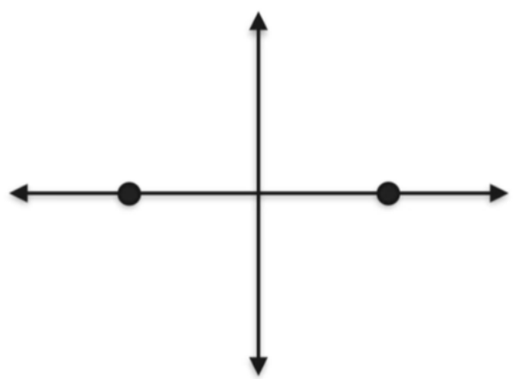
MLE: Bernoulli Example

Maximum Likelihood: Gaussian Example

Candidates



Observations

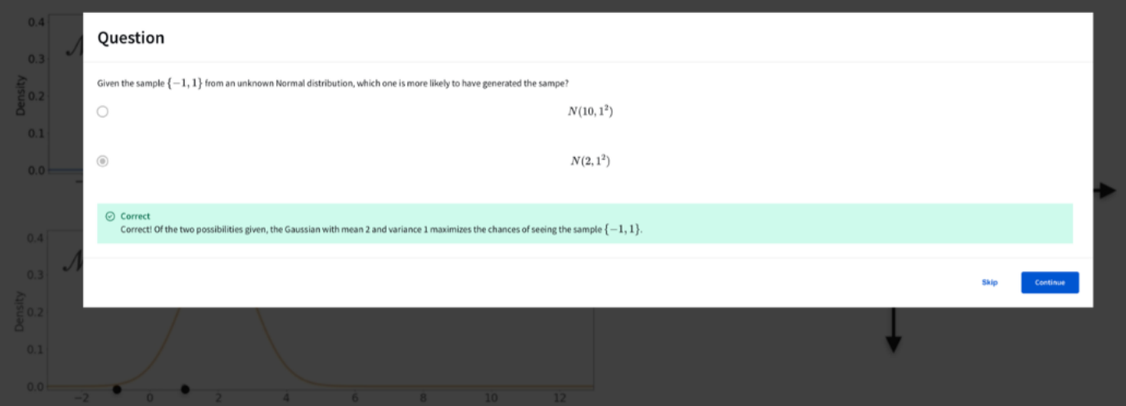


DeepLearning.AI

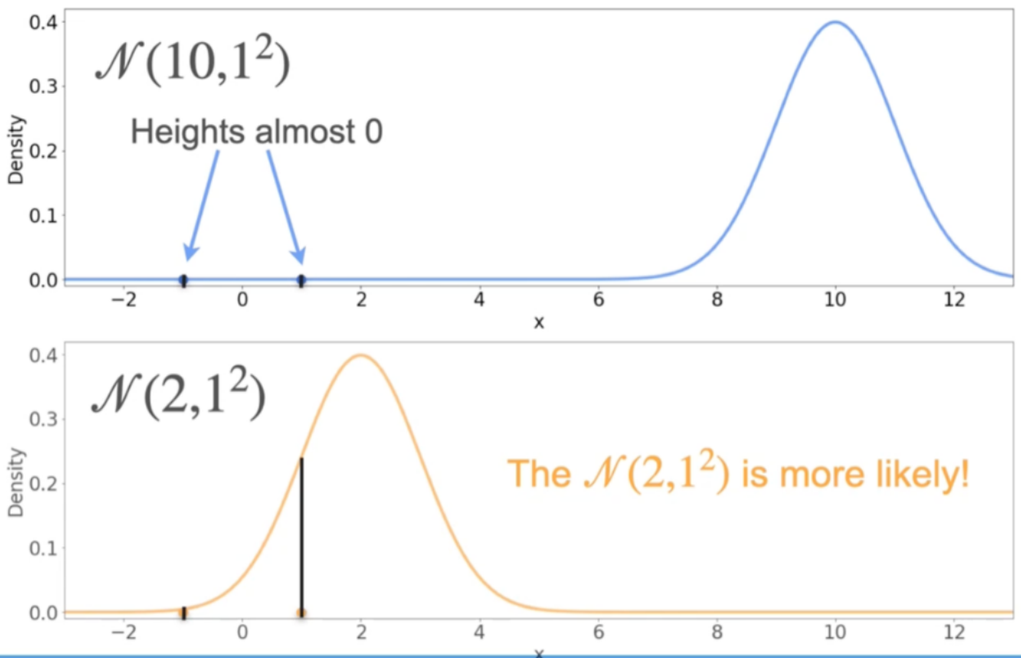
Maximum Likelihood: Gaussian Example

Candidates

Observations



DeepLearning.AI

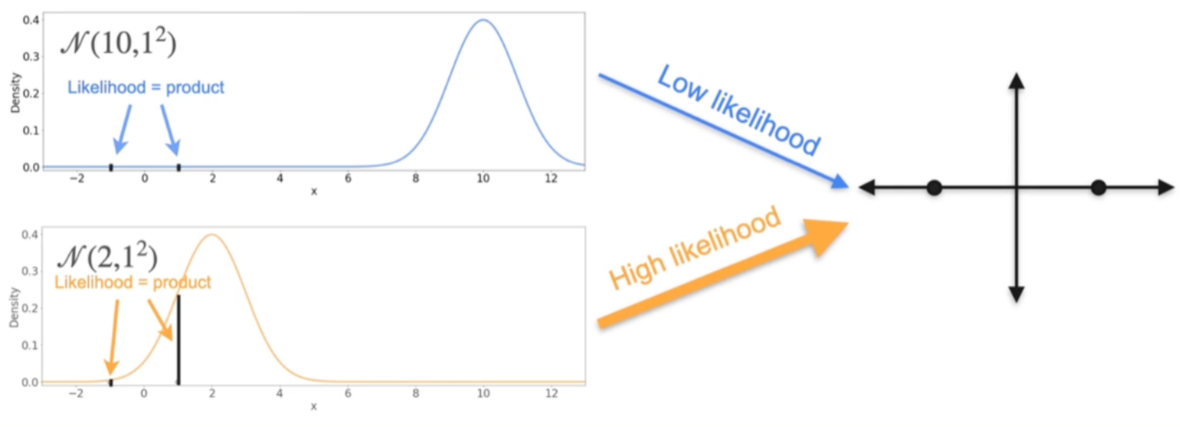


DeepLearning.AI

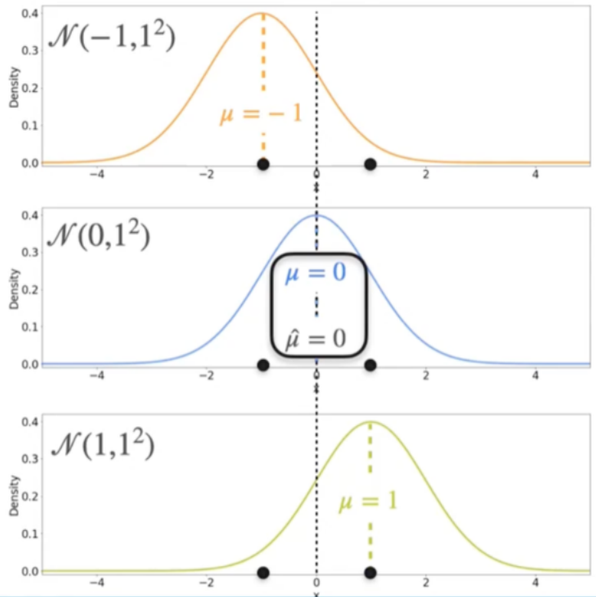
Maximum Likelihood: Gaussian Example

Candidates

Observations



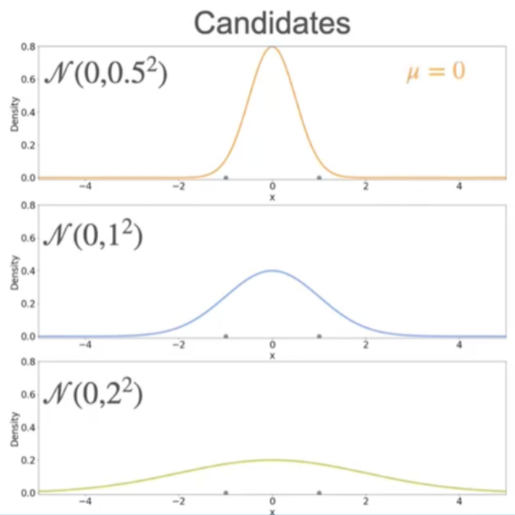
DeepLearning.AI

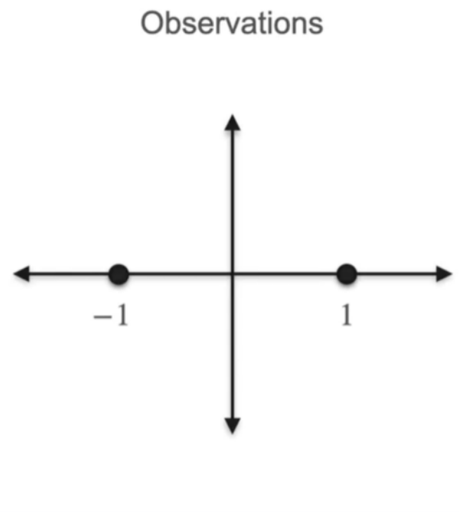


The best distribution is the one where the **mean** of the distribution is the **mean** of the sample

DeepLearning.AI

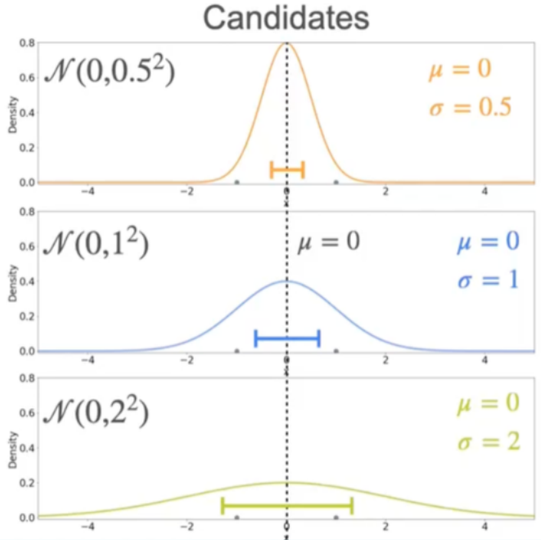
Gaussians With Three Different Variance

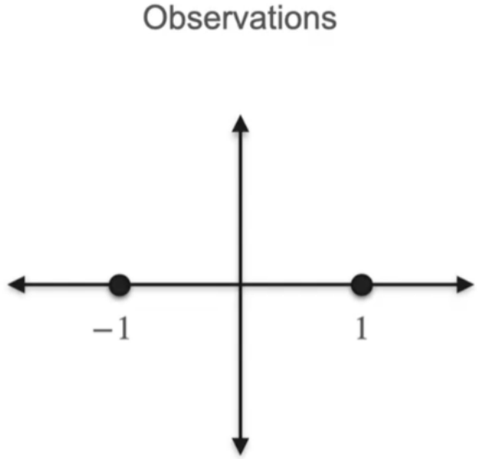




DeepLearning.AI

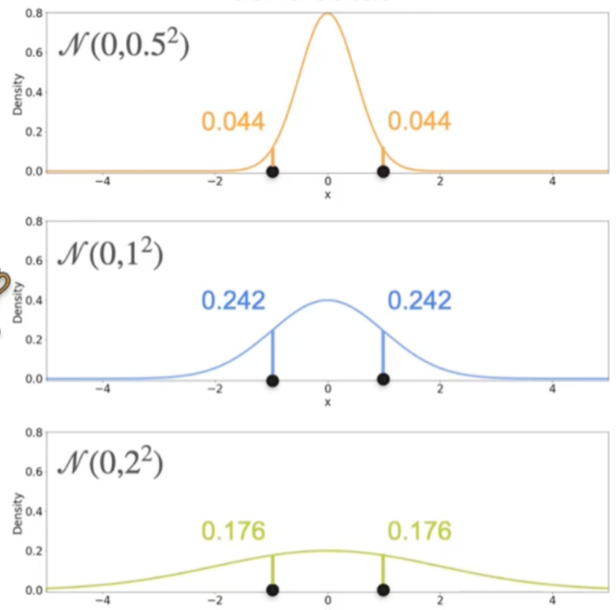
Gaussians With Three Different Variance





DeepLearning.AI

Candidates



Observations

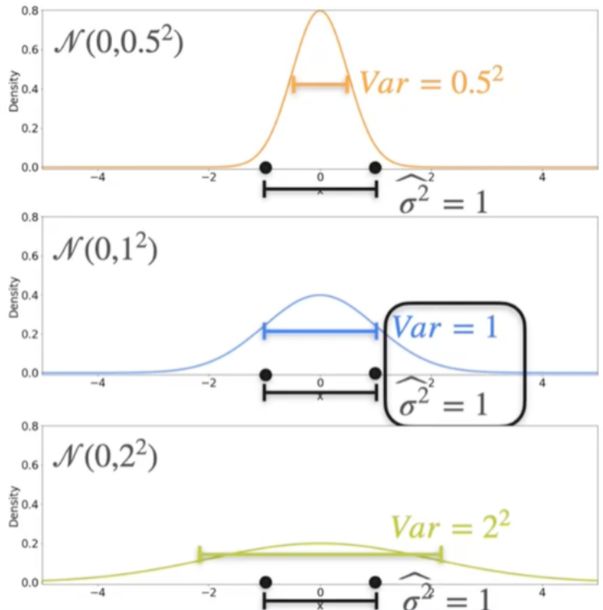
$0.044 \cdot 0.044 = 0.002$

The $\mathcal{N}(0,1^2)$ is more likely!

$0.242 \cdot 0.242 = 0.059$

$0.176 \cdot 0.176 = 0.031$

Candidates



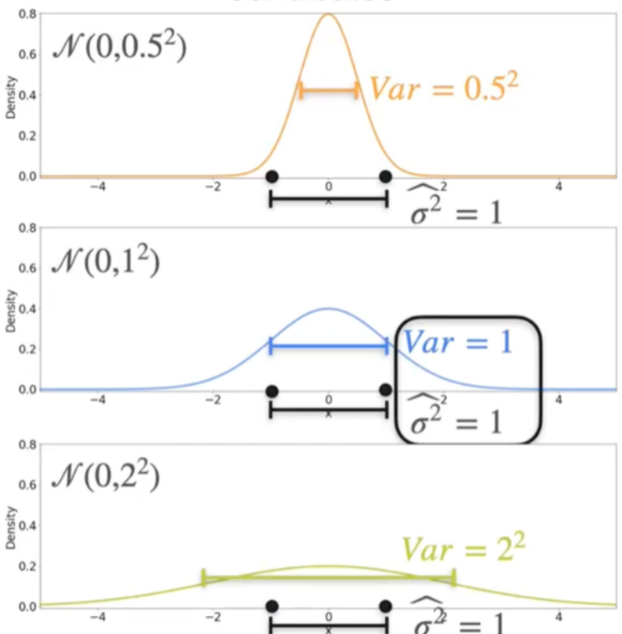
Observations

Variance of the observations

$$\widehat{\sigma^2} = \frac{1}{2}\left((0-1)^2 + (0+1)^2\right) = 1$$

DeepLearning.AI

Candidates



Observations

Variance of the observations

$$\widehat{\sigma^2} = \frac{1}{2}\left((0-1)^2 + (0+1)^2\right) = 1$$

The best distribution is the one where the **variance** of the distribution is the **variance** of the sample

DeepLearning.AI

MLE for Gaussian population

In the videos, you got an intuition of what the Maximum Likelihood Estimation (MLE) should look like for the mean and variance of a Gaussian population.

In this reading item, you will learn the derivation of both results.

Mathematical formulation

Suppose you have n samples $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ from a Gaussian distribution with mean μ and variance σ^2. This means that $X_i \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$.

If you want the MLE for μ and σ the first step is to define the likelihood. If both μ and σ are unknown, then the likelihood will be a function of these two parameters. For a realization of X, given by $x = (x_1, x_2, \ldots, x_n)$:

$$\begin{aligned} L(\mu, \sigma; \boldsymbol{x}) &= \prod_{i=1}^n f_{X_i}(x_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}} \\ &= \frac{1}{\left(\sqrt{2\pi}\right)^n \sigma^n} e^{-\frac{1}{2}\frac{\sum_{i=1}^n (x_i-\mu)^2}{\sigma^2}} \end{aligned}$$

Now all you have to do is find the values of μ and σ that maximize the likelihood $L(\mu, \sigma; \boldsymbol{x})$.

You might remember from the calculus course that one way to do this analytically is by taking the derivative of the Likelihood function and equating it to 0. The values of μ and σ that make the derivative zero, are the extreme points. In particular, for this case, they will be maximums.

Taking the derivative of the likelihood is a cumbersome procedure, because of all the products involved. However, there is a nice trick you can use to simplify things. Note that the logarithm function is always increasing, so the values that maximize $L(\mu, \sigma; \boldsymbol{x})$ will also maximize its logarithm. This is the **log-likelihood**, and it is defined as

$$\ell(\mu, \sigma) = \log(L(\mu, \sigma; \boldsymbol{x}))$$

The logarithm has the property of turning a product into a sum, this means that $\log(a \cdot b) = \log(a) + \log(b)$. This makes taking the derivative of the log-likelihood very straight forward. To get the simplest expression for the log-likelihood for a Gaussian population, you will also need the following properties of the logarithm:

$$\log(1/a) = -\log(a)$$

The logarithm has the property of turning a product into a sum, this means that $\log(a \cdot b) = \log(a) + \log(b)$. This makes taking the derivative of the log-likelihood very straight forward. To get the simplest expression for the log-likelihood for a Gaussian population, you will also need the following properties of the logarithm:

$$\log(1/a) = -\log(a)$$

and

$$\log(a^k) = k\log(a).$$

Putting it all together you get:

$$\begin{aligned} \ell(\mu, \sigma) &= \log\left(\frac{1}{\left(\sqrt{2\pi}\right)^n \sigma^n} e^{-\frac{1}{2}\frac{\sum_{i=1}^n (x_i-\mu)^2}{\sigma^2}}\right) \\ &= -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2}\frac{\sum_{i=1}^n (x_i-\mu)^2}{\sigma^2} \end{aligned}$$

Now to find the MLE for μ and σ, all there is left to do is take the partial derivatives of the log-likelihood, and equate them to zero.

For the partial derivative with respect to μ note that the first two terms do not involve μ, so you get:

$$\begin{aligned} \frac{\partial}{\partial \mu}\ell(\mu, \sigma) &= -\frac{1}{2}\frac{\sum_{i=1}^n 2(x_i-\mu)}{\sigma^2}(-1) \\ &= \frac{1}{\sigma^2}\left(\sum_{i=1}^n x_i - \sum_{i=1}^n \mu\right) = \frac{1}{\sigma^2}\left(\sum_{i=1}^n x_i - n\mu\right) \end{aligned}$$

Now, for the partial derivative with respect to σ you get that

$$\frac{\partial}{\partial \sigma}\ell(\mu, \sigma) = -\frac{n}{\sigma} - \frac{1}{2}\left(\sum_{i=1}^n (x_i-\mu)^2\right)(-2)\frac{1}{\sigma^3} = -\frac{n}{\sigma} + \left(\sum_{i=1}^n (x_i-\mu)^2\right)\frac{1}{\sigma^3}$$

The next step is equating this to 0 to find the estimates for μ and σ. Let's begin with the partial derivative with respect to μ:

$$\frac{\partial}{\partial \mu}\ell(\mu, \sigma) = \frac{1}{\sigma^2}\left(\sum_{i=1}^n x_i - n\mu\right) = 0$$

First, observe that since $\sigma > 0$, the only option is that $\sum_{i=1}^n x_i - n\mu = 0$. Simple algebraic manipulations show that the MLE for μ has to be

$$\hat{\mu} = \frac{\sum_{i=1}^n x_i}{n} = \overline{x},$$

which is the sample mean.

Next, find the value of σ that achieves $\frac{\partial}{\partial \sigma}\ell(\mu, \sigma) = 0$:

which is the sample mean.

Next, find the value of σ that achieves $\frac{\partial}{\partial\sigma}\ell(\mu,\sigma)=0$:

$$\frac{\partial}{\partial\sigma}\ell(\mu,\sigma)=-\frac{n}{\sigma}+\left(\sum_{i=1}^{n}(x_i-\mu)^2\right)\frac{1}{\sigma^3}=0$$

In this case, first note that since $\sigma>0$ you can simplify the expression to

$\frac{\partial}{\partial\sigma}\ell(\mu,\sigma)=-n+(\sum_{i=1}^{n}(x_i-\mu)^2)\,\frac{1}{\sigma^2}=0$

Also, you can replace μ by its estimate $\hat{\mu}=\overline{x}$, because yuo want both partial derivatives to be 0 at the same time. You get

$\frac{\partial}{\partial\sigma}\ell(\mu,\sigma)=-n+(\sum_{i=1}^{n}(x_i-\overline{x})^2)\,\frac{1}{\sigma^2}=0$

This gives you

$\sigma^2=\frac{\sum(x_i-\overline{x})^2}{n}$,

so the MLE for the standard deviation is

$$\hat{\sigma}=\sqrt{\frac{\sum(x_i-\overline{x})^2}{n}}$$

This expression tells you that the MLE for the standard deviation of a Gaussian population is the square root of the average squared difference between each sample and the sample mean. This expression is very similar to the one you learnt in Week 2 for the sample standard deviation. The only difference is the normalizing constant: for the MLE you have $1/n$ while for the sample standard deviation you use $1/(n-1)$.

A final comment: formally, what you just did was the derivation of the critical point. To make it all complete, you would need to show that these are the coordinates of a maximum point (and not a minimum or saddle point). However, this proof would require a little bit more complicated math and we will skip it here.

A simple example

Now, let's see how this looks like with an example. Suppose you are interested on distribution of heights of 18 year olds in the US. You have the following 10 measurements:

66.75	70.24	67.19	67.09	63.65
64.64	69.81	69.79	73.52	71.74

You get

$\frac{\partial}{\partial\sigma}\ell(\mu,\sigma)=-n+(\sum_{i=1}^{n}(x_i-\overline{x})^2)\,\frac{1}{\sigma^2}=0$

This gives you

$\sigma^2=\frac{\sum(x_i-\overline{x})^2}{n}$,

so the MLE for the standard deviation is

$$\hat{\sigma}=\sqrt{\frac{\sum(x_i-\overline{x})^2}{n}}$$

This expression tells you that the MLE for the standard deviation of a Gaussian population is the square root of the average squared difference between each sample and the sample mean. This expression is very similar to the one you learnt in Week 2 for the sample standard deviation. The only difference is the normalizing constant: for the MLE you have $1/n$ while for the sample standard deviation you use $1/(n-1)$.

A final comment: formally, what you just did was the derivation of the critical point. To make it all complete, you would need to show that these are the coordinates of a maximum point (and not a minimum or saddle point). However, this proof would require a little bit more complicated math and we will skip it here.

A simple example

Now, let's see how this looks like with an example. Suppose you are interested on distribution of heights of 18 year olds in the US. You have the following 10 measurements:

66.75	70.24	67.19	67.09	63.65
64.64	69.81	69.79	73.52	71.74

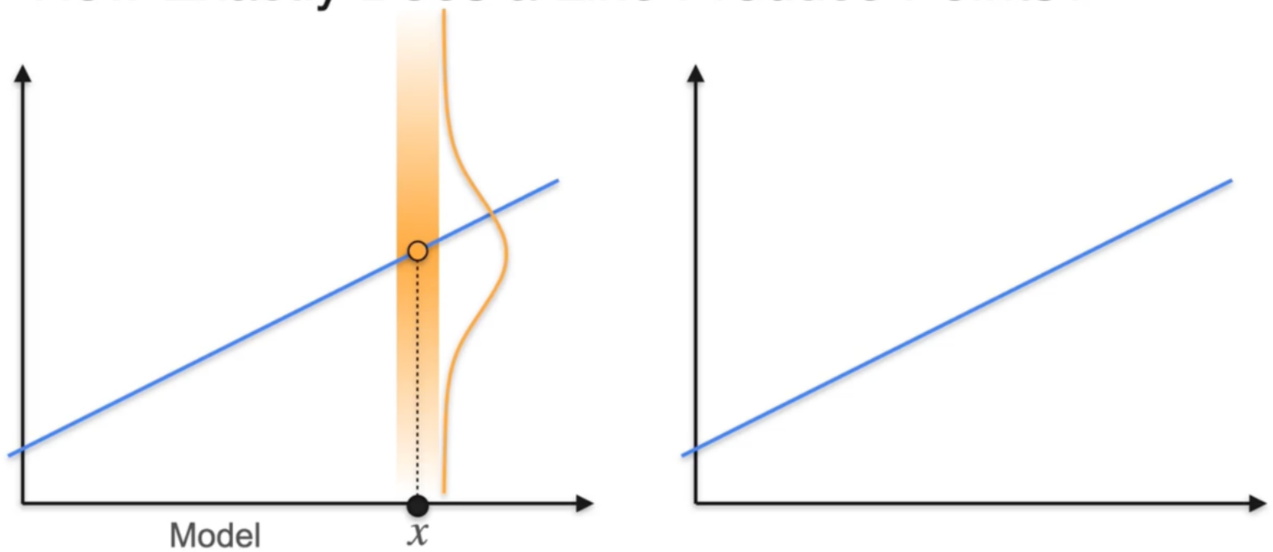
Each measurement is supposed to come from a Gaussian distribution with unknown parameters μ and σ. The MLE estimation for the parameters with this samples are

$$\hat{\mu}=\frac{66.75+70.24+67.19+67.09+63.65+64.64+69.81+69.79+73.52+71.74}{10}=68.442$$

and

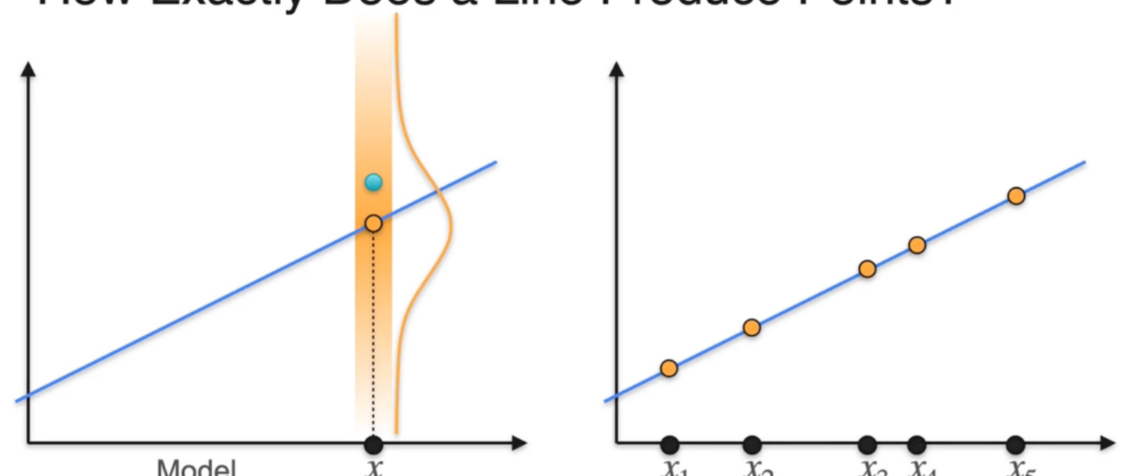
$$\begin{aligned}\hat{\sigma}&=\sqrt{\frac{1}{10}\left(\begin{array}{l}(66.75-68.442)^2+(70.24-68.442)^2+(67.19-68.442)^2+(67.09-68.442)^2+\\(63.65-68.442)^2+(64.64-68.442)^2+(69.81-68.442)^2+(69.79-68.442)^2+\\(73.52-68.442)^2+(71.74-68.442)^2\end{array}\right)}\\&=2.954\end{aligned}$$

How Exactly Does a Line Produce Points?



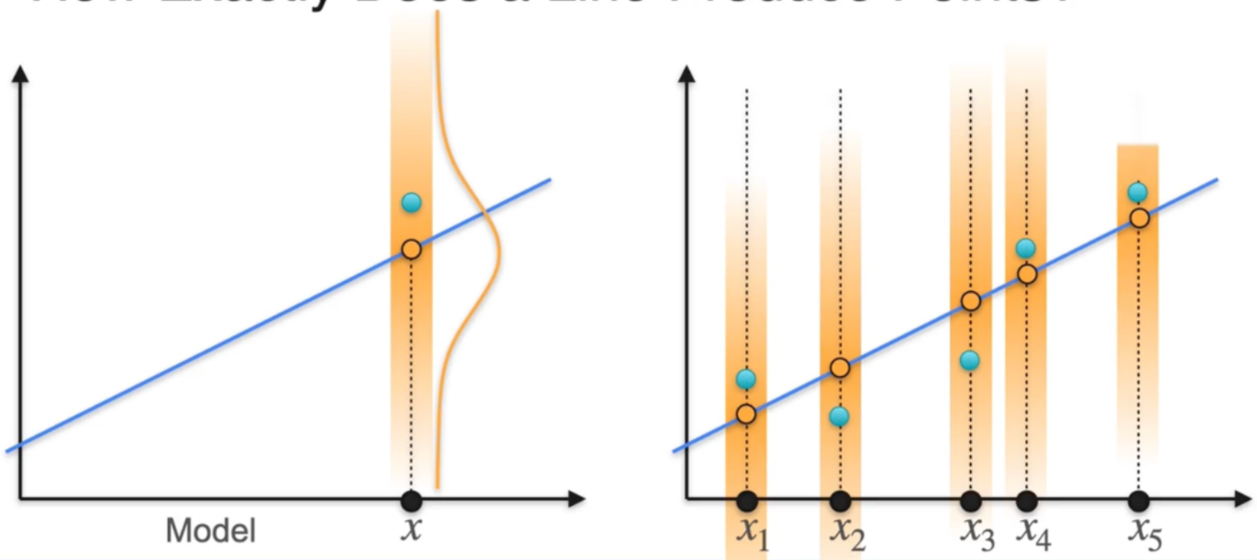
DeepLearning.AI

How Exactly Does a Line Produce Points?



DeepLearning.AI

How Exactly Does a Line Produce Points?



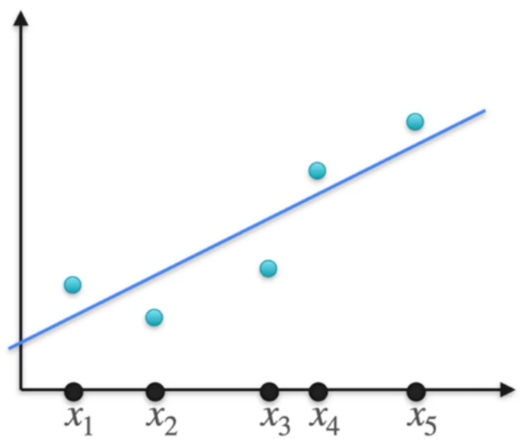
DeepLearning.AI

Linear Regression

Line that best produced the points

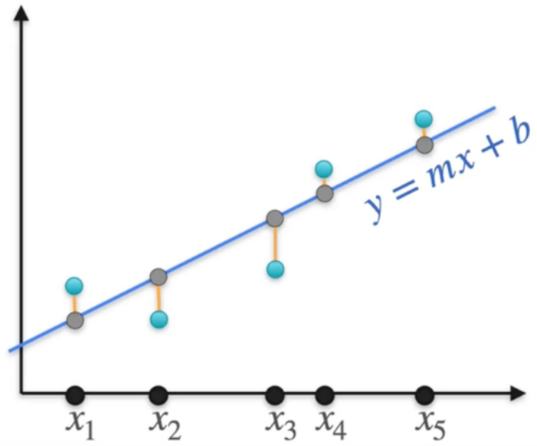
How?

Line that best fits the data (linear regression)



DeepLearning.AI

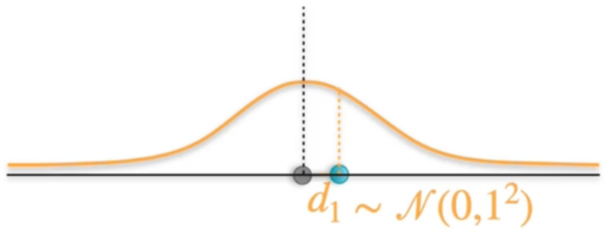
Linear Regression and Likelihood

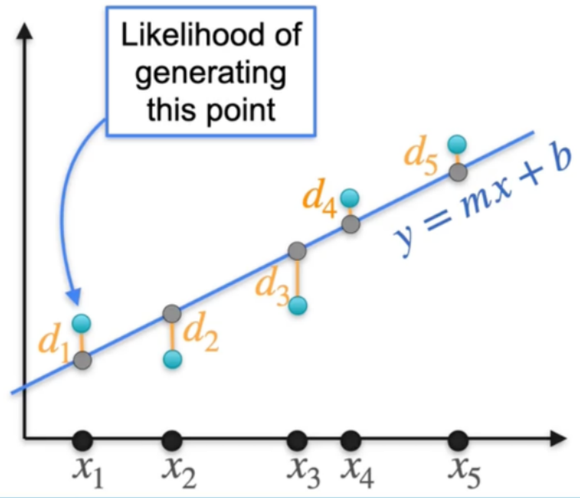


DeepLearning.AI

Linear Regression and Likelihood

Likelihood:

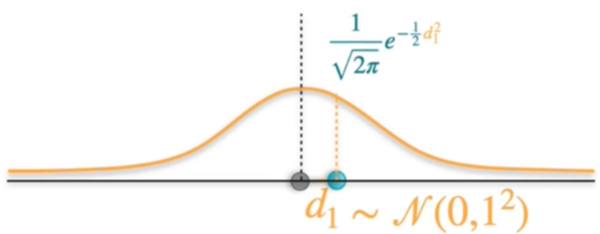


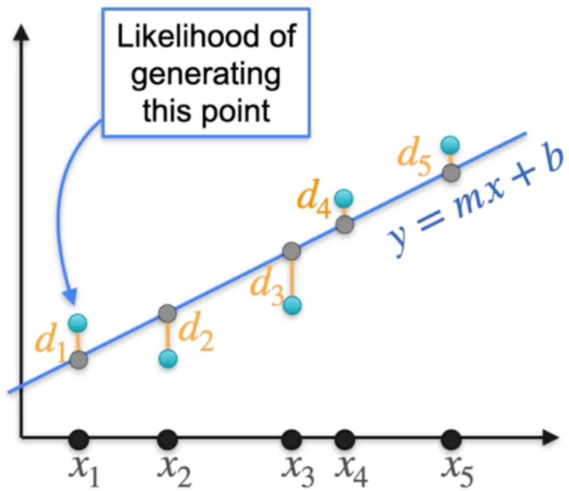


DeepLearning.AI

Linear Regression and Likelihood

Likelihood:





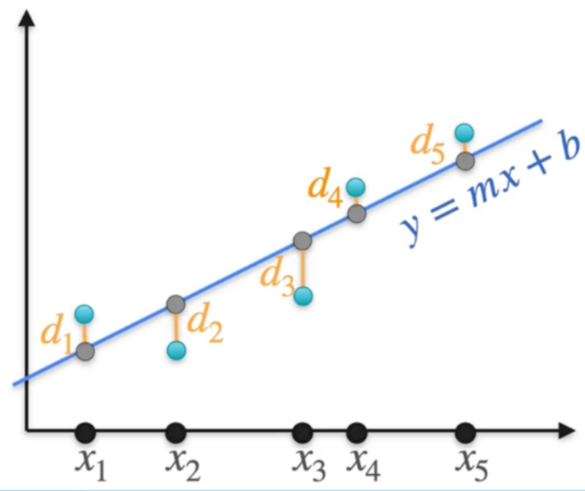
DeepLearning.AI

Linear Regression and Likelihood

Likelihood:

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d_1^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d_2^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d_3^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d_4^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d_5^2}$$

Maximize



DeepLearning.AI

Linear Regression and Likelihood

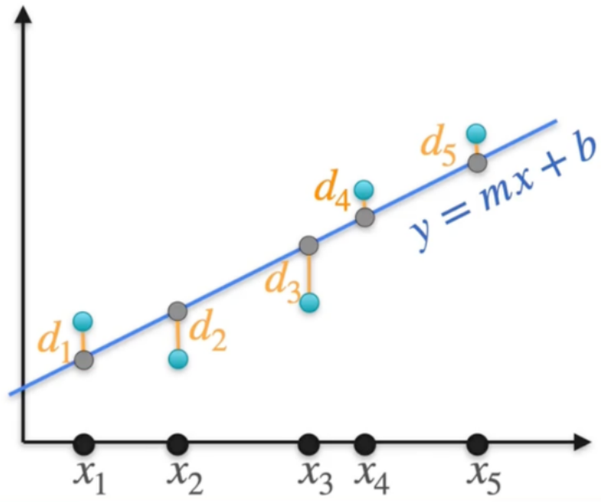
Likelihood:

$$\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}d_1^2}\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}d_2^2}\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}d_3^2}\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}d_4^2}\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}d_5^2}$$

Maximize

$$e^{-\frac{1}{2}d_1^2}\cdot e^{-\frac{1}{2}d_2^2}\cdot e^{-\frac{1}{2}d_3^2}\cdot e^{-\frac{1}{2}d_4^2}\cdot e^{-\frac{1}{2}d_5^2}$$

$$e^{-\frac{1}{2}(}$$



DeepLearning.AI

Linear Regression and Likelihood

Likelihood:

$$\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}d_1^2}\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}d_2^2}\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}d_3^2}\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}d_4^2}\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}d_5^2}$$

Maximize

$$e^{-\frac{1}{2}d_1^2}\cdot e^{-\frac{1}{2}d_2^2}\cdot e^{-\frac{1}{2}d_3^2}\cdot e^{-\frac{1}{2}d_4^2}\cdot e^{-\frac{1}{2}d_5^2}$$

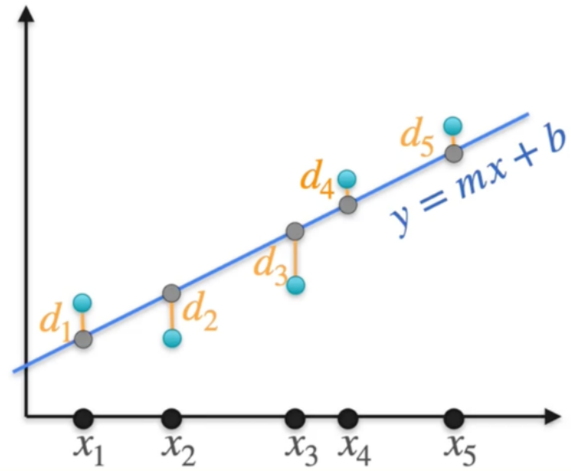
$$e^{-\frac{1}{2}(d_1^2+d_2^2+d_3^2+d_4^2+d_5^2)}$$

Minimize

$$d_1^2+d_2^2+d_3^2+d_4^2+d_5^2$$

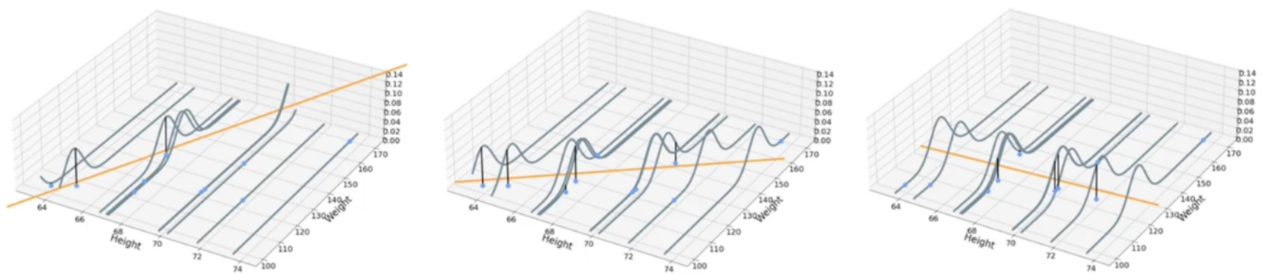
Linear regression!

Least squares error!



DeepLearning.AI

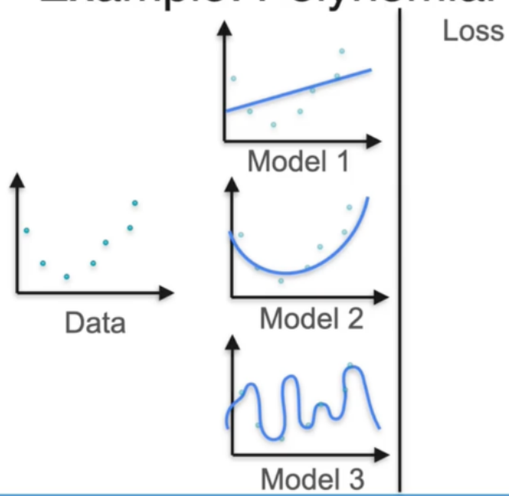
Picking the Right Model



Model 1:
Likelihood $= 4.91 \cdot 10^{-260}$

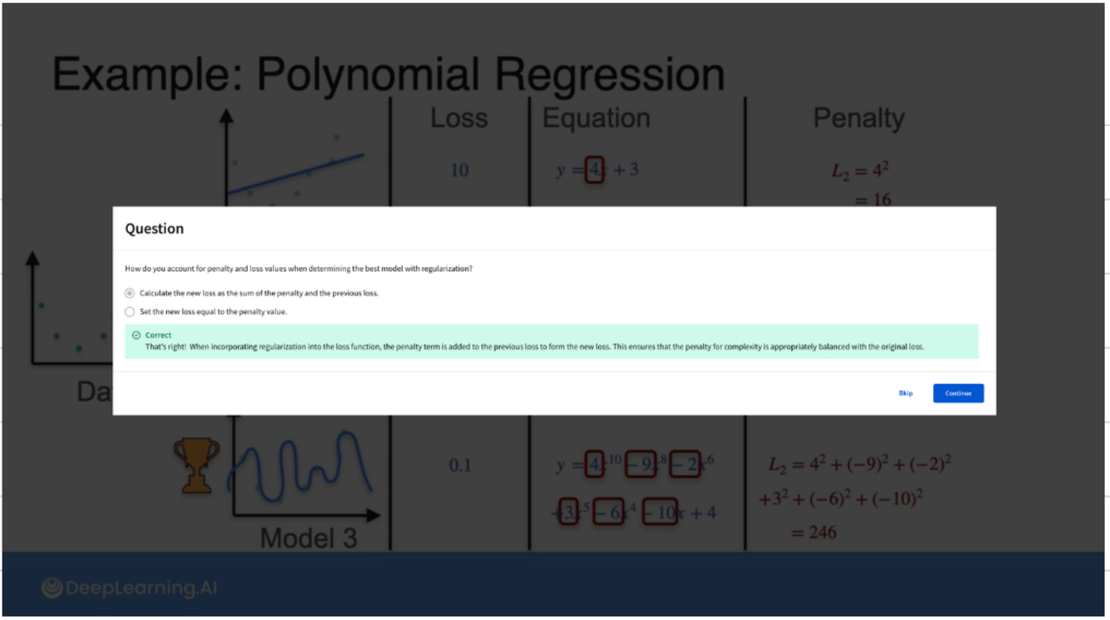
DeepLearning.AI

Example: Polynomial Regression

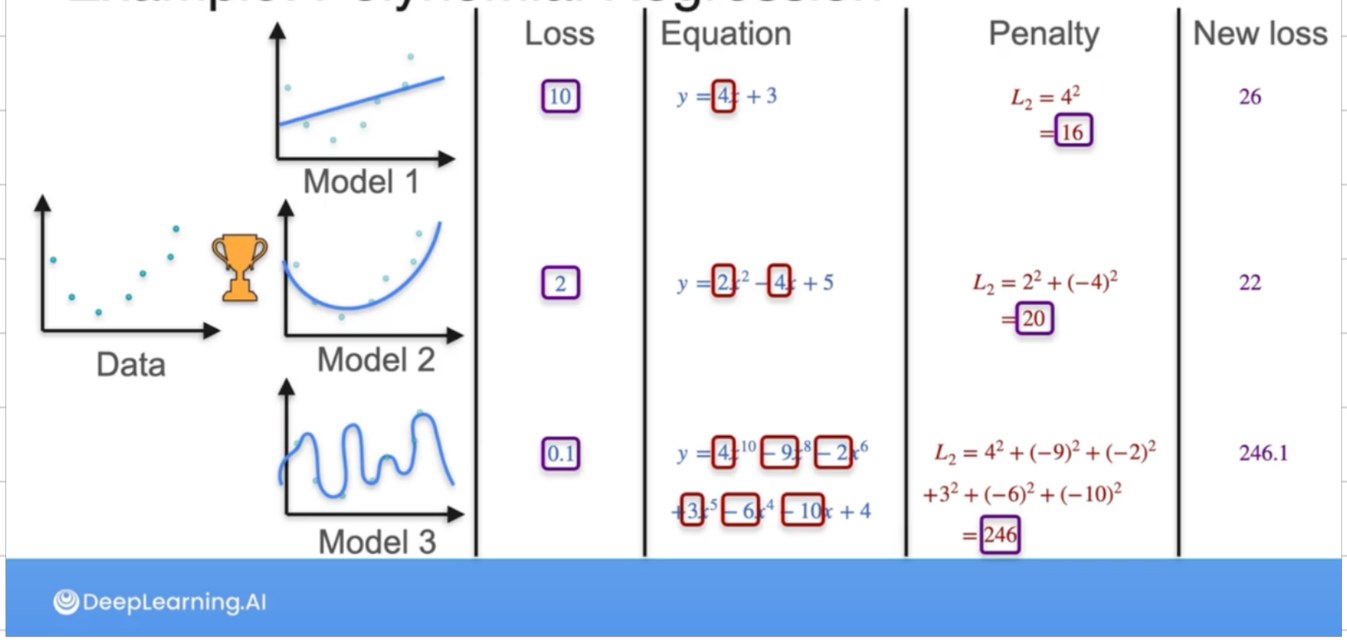


DeepLearning.AI

Example: Polynomial Regression



Example: Polynomial Regression



Regularization Term

Model: $y = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$

Log-loss: $\ell\ell$

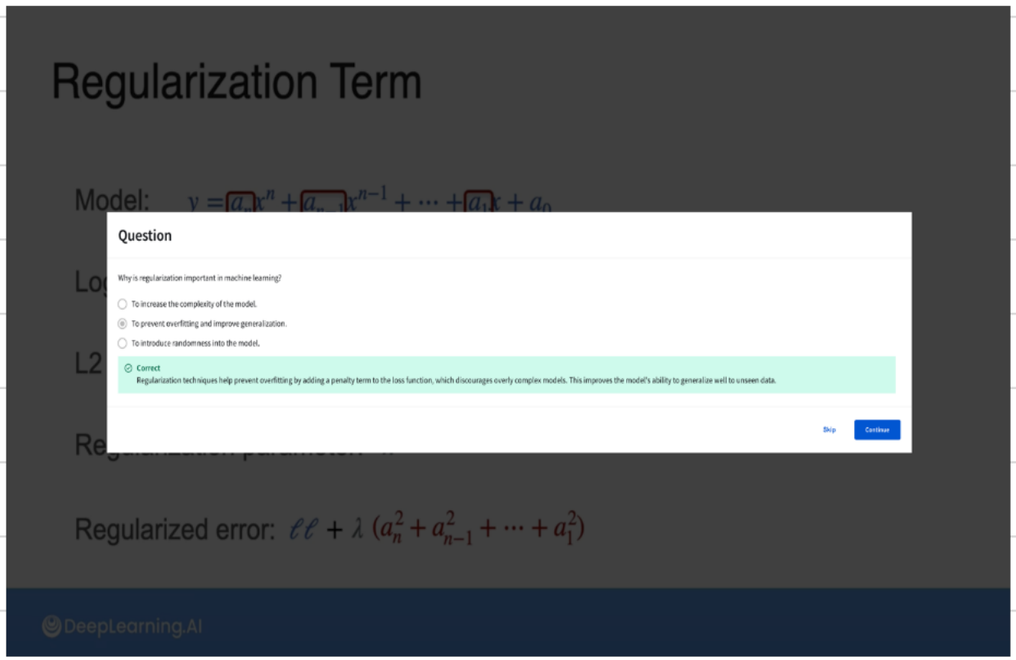
L2 Regularization Error: $a_n^2 + a_{n-1}^2 + \cdots + a_1^2$

Regularization parameter: λ

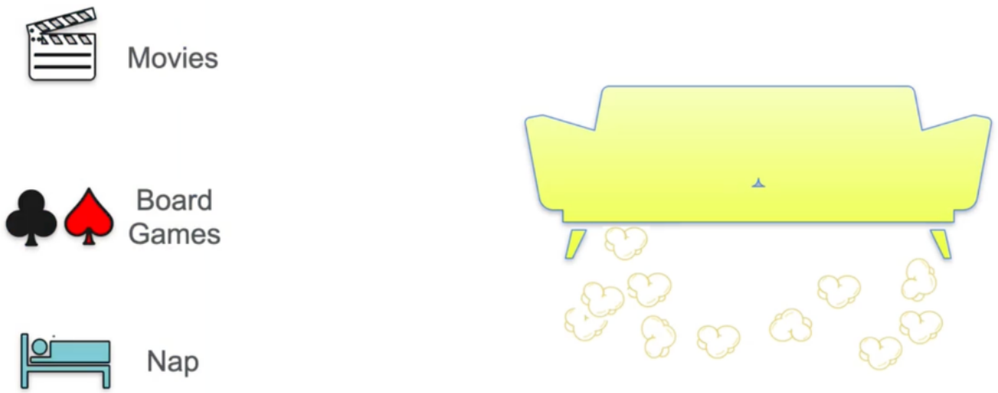
Regularized error: $\ell\ell + \lambda\ (a_n^2 + a_{n-1}^2 + \cdots + a_1^2)$

DeepLearning.AI

Regularization Term

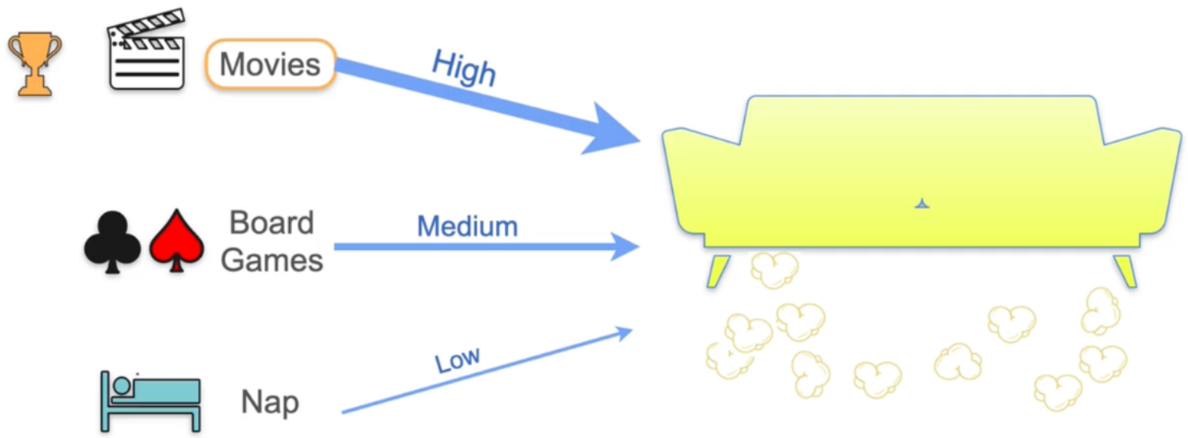


There's Popcorn on the Floor. What Happened?



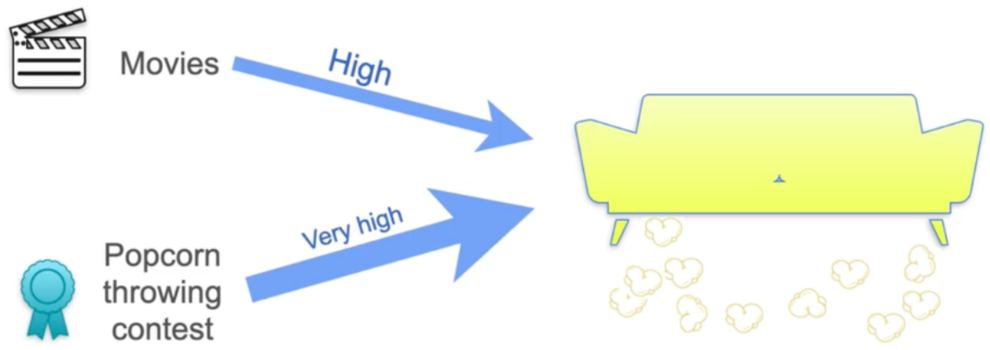
DeepLearning.AI

There's Popcorn on the Floor. What Happened?



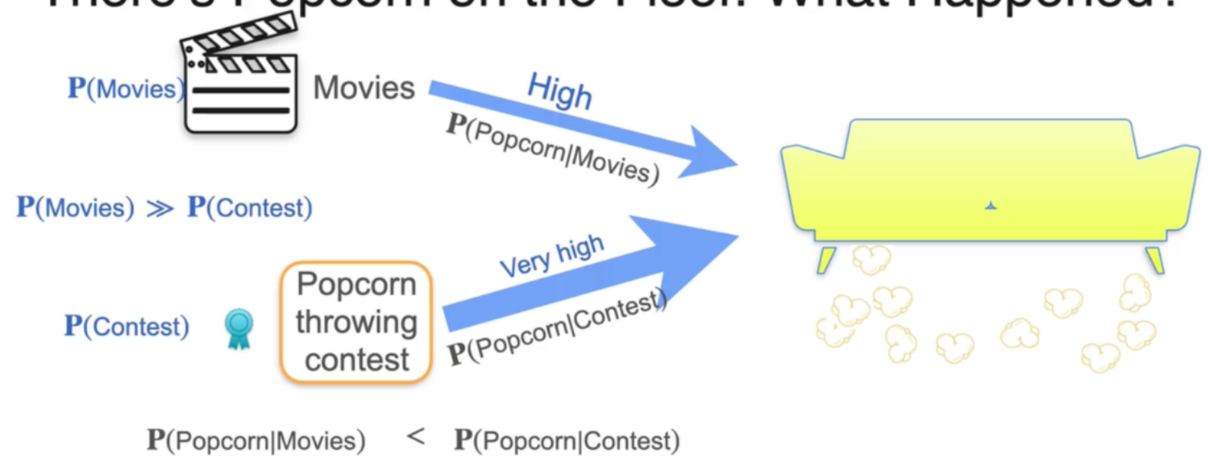
DeepLearning.AI

There's Popcorn on the Floor. What Happened?



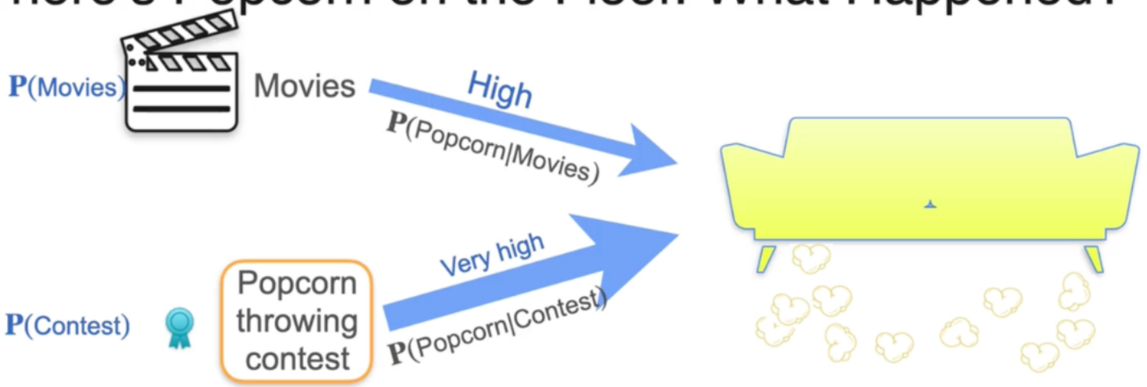
DeepLearning.AI

There's Popcorn on the Floor. What Happened?



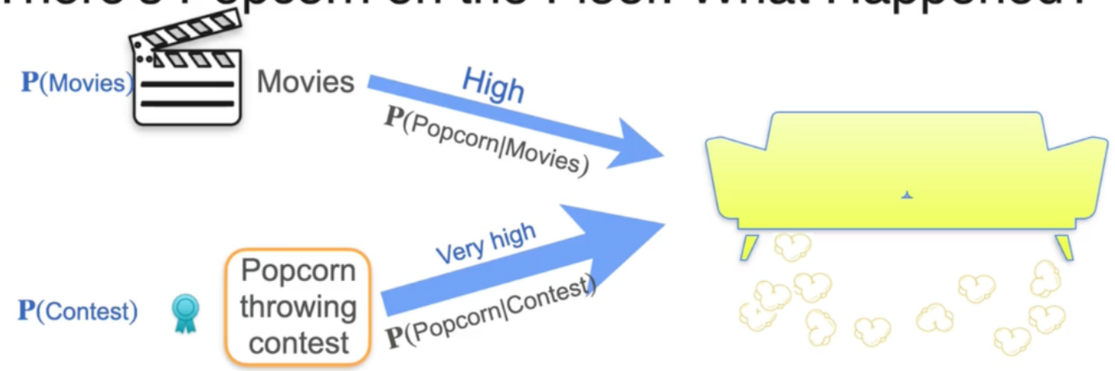
DeepLearning.AI

There's Popcorn on the Floor. What Happened?



$$\mathbf{P}(\text{Popcorn}|\text{Movies})\,\mathbf{P}(\text{Movies}) > \mathbf{P}(\text{Popcorn}|\text{Contest})\,\mathbf{P}(\text{Contest})$$

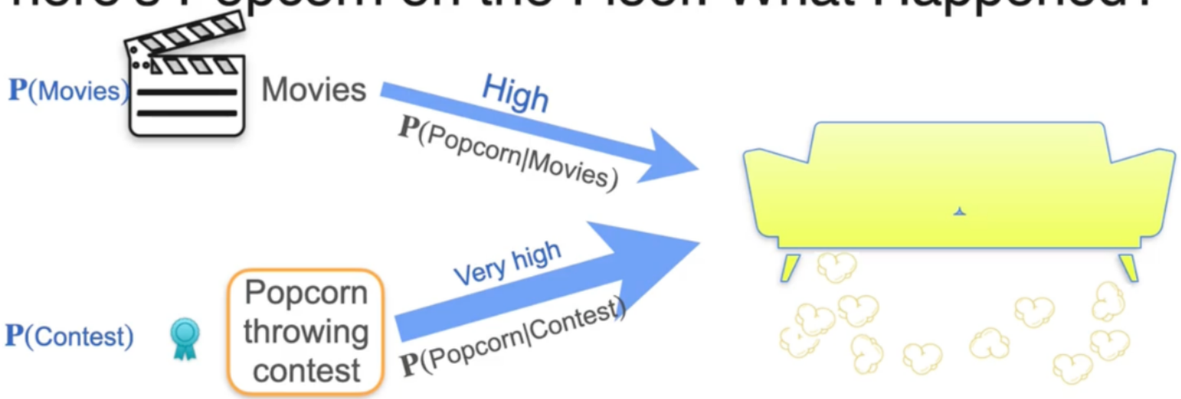
There's Popcorn on the Floor. What Happened?



$$\mathbf{P}(\text{Popcorn}|\text{Movies})\,\mathbf{P}(\text{Movies}) > \mathbf{P}(\text{Popcorn}|\text{Contest})\,\mathbf{P}(\text{Contest})$$

$$\mathbf{P}(A \mid B)\mathbf{P}(B) = \mathbf{P}(A \cap B)$$

There's Popcorn on the Floor. What Happened?

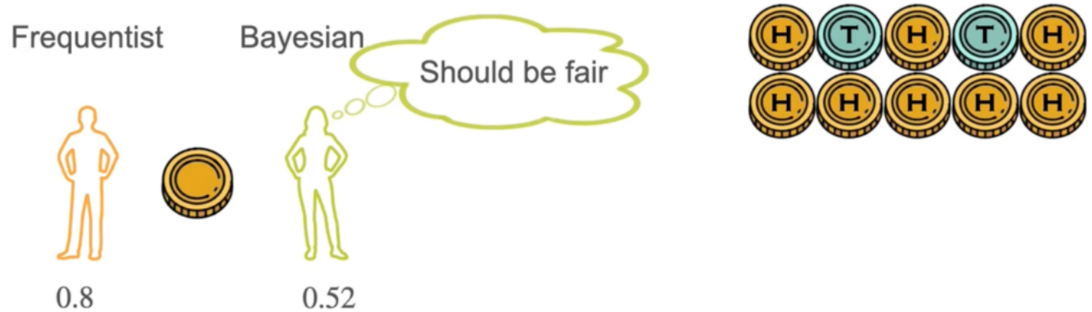


$$\mathbf{P}(\text{Popcorn}|\text{Movies})\,\mathbf{P}(\text{Movies}) > \mathbf{P}(\text{Popcorn}|\text{Contest})\,\mathbf{P}(\text{Contest})$$

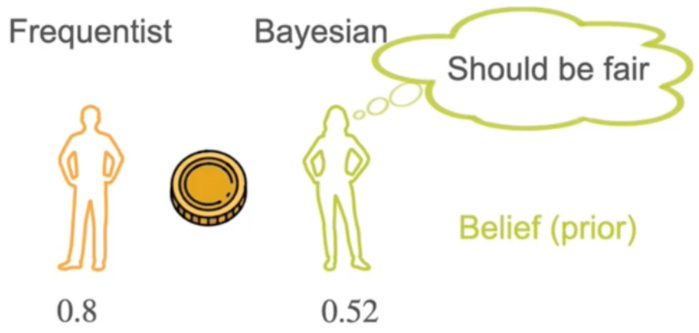
$$\mathbf{P}(A \mid B)\mathbf{P}(B) = \mathbf{P}(A \cap B)$$

$$\mathbf{P}(\text{Popcorn} \cap \text{Movie}) > \mathbf{P}(\text{Popcorn} \cap \text{Contest})$$

Frequentists vs. Bayesians



Frequentists vs. Bayesians



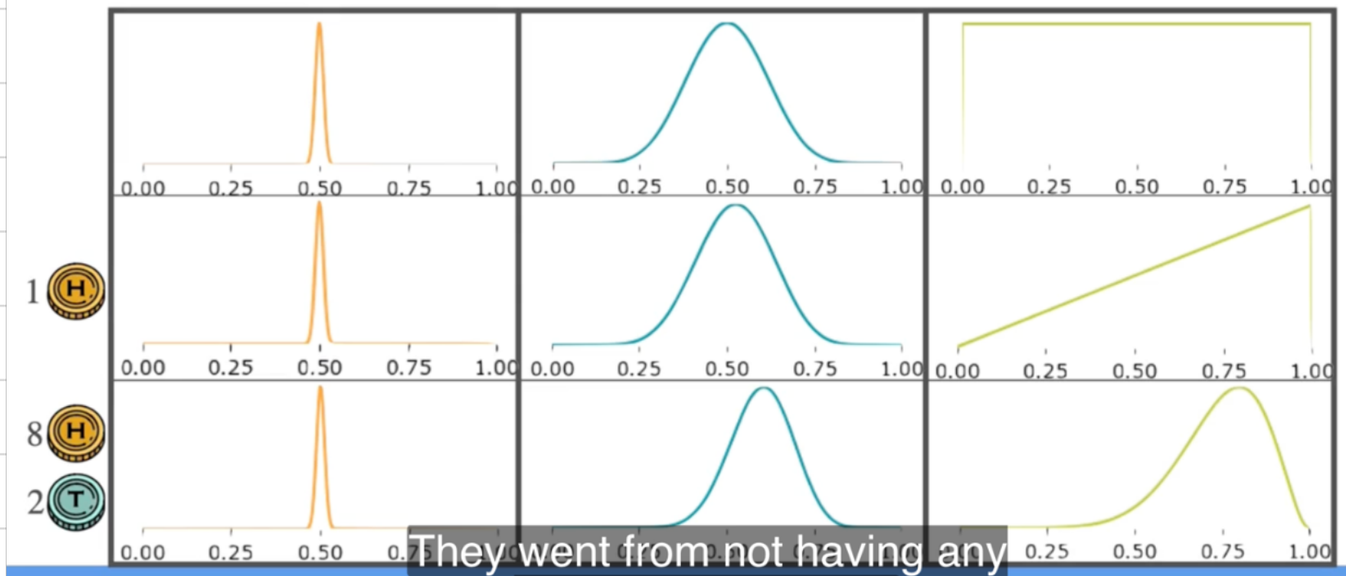
DeepLearning.AI

Frequentist Vs. Bayesian Statistics

Frequentists	Bayesians
• Probabilities represent long term frequency of events	• Probabilities represent the degree of belief (or certainty)
• Concept of Likelihood	• Concept of Prior
• Goal: Find the model that most likely generated the observed data	• Goal: update prior belief based on observations

DeepLearning.AI

Updating Your Beliefs



They went from not having any information to being

DeepLearning.AI

Source:

Coursera Probability & Statistics for Data Science and Machine learning