Module-1

Math, numpy, Pytorch

Acknowledgement

Thanks to the following courses and corresponding researchers for making their teaching/research material online

- Mike X Cohen's Master Deep Learning course (Udemy)
- AI Tools such as Chat GPT, Grok, etc
- Introduction to Deep Learning, University of Illinois at Urbana-Champaign
- Introduction to Deep Learning, Carnegie Mellon University
- Convolutional Neural Networks for Visual Recognition, Stanford University
- Natural Language Processing with Deep Learning, Stanford University
- And Many More .....

1

Spectral Theories in Mathematics

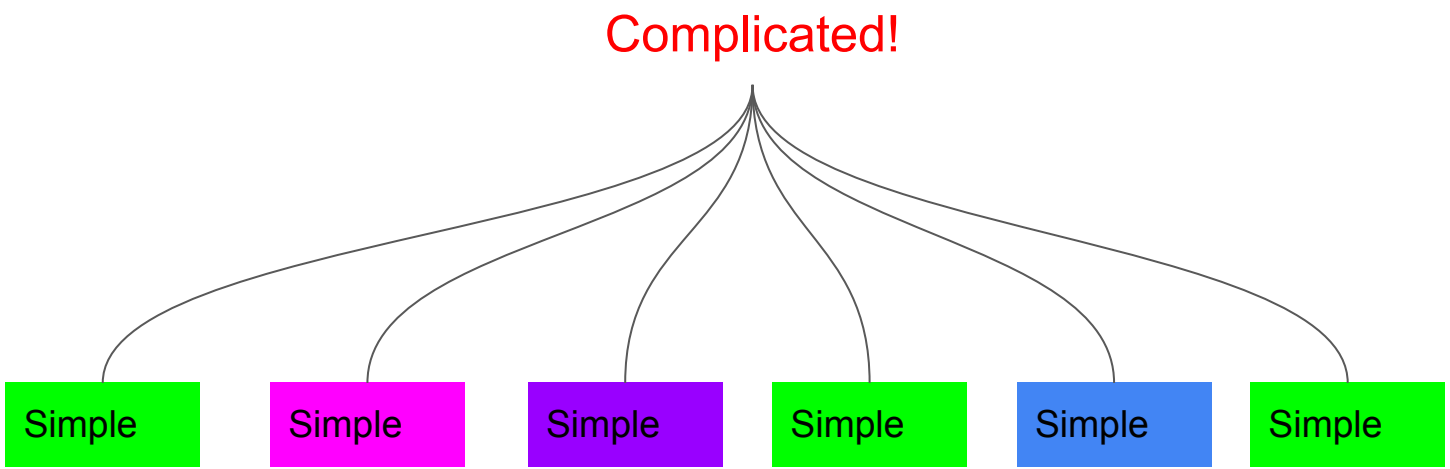
Spectral theories in mathematics

In this submodule, you will explore:

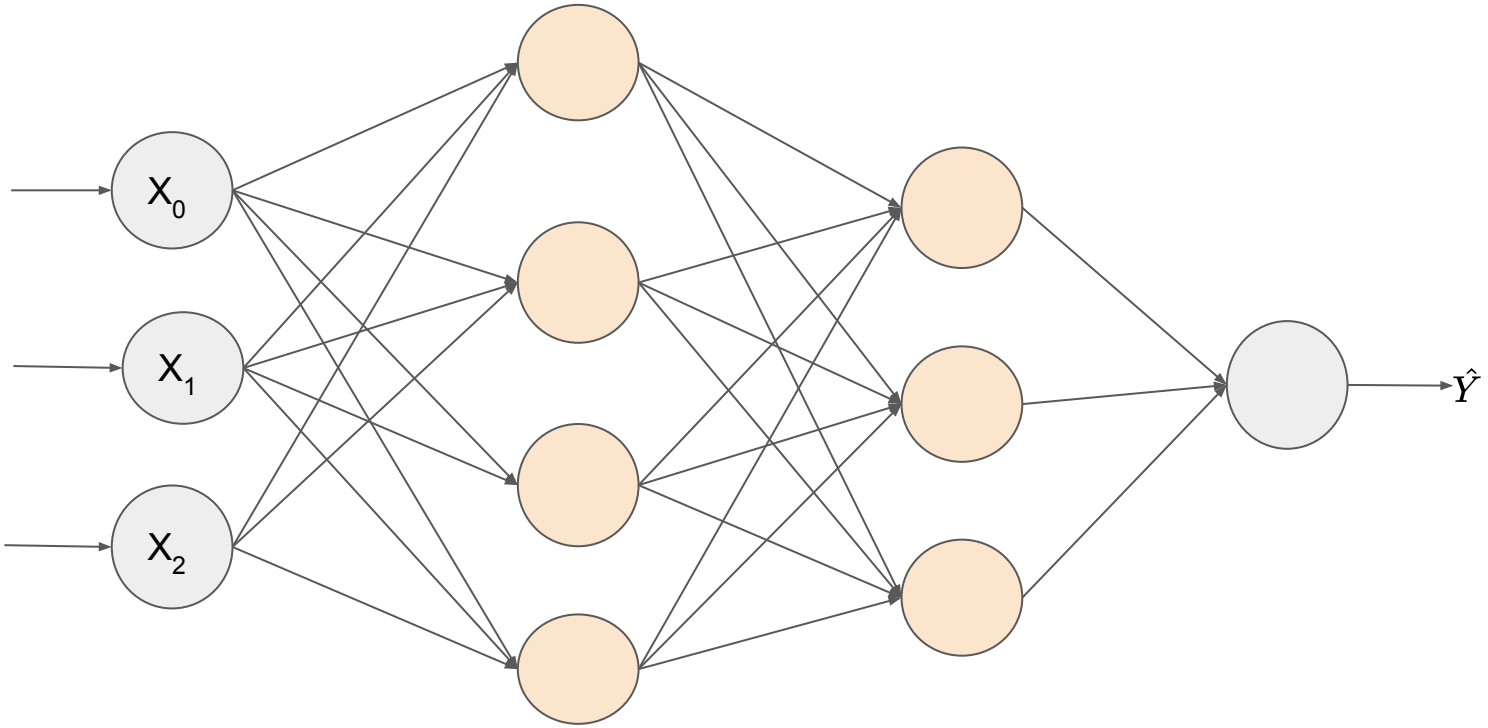
- The meaning and significance of **spectral theory**
- How **artificial neural networks** relate to spectral theory
- The distinction between **"complicated"** and **"complex"** systems
- Why **deep learning** is simultaneously **easy, complicated, and complex**!

General idea of spectral theories in mathematics

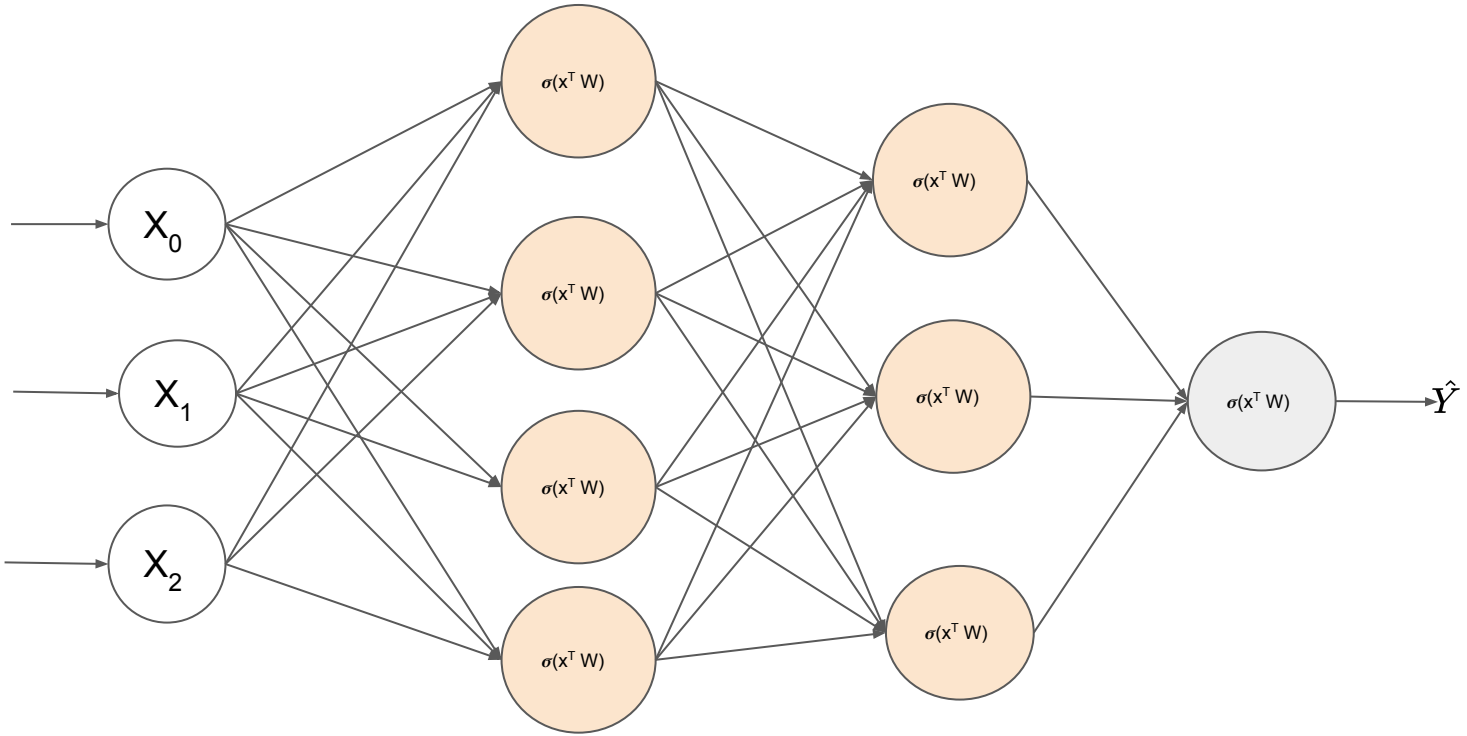
- Spectral theory involves breaking down complex structures into simpler, individual components, making them easier to analyze and understand



Spectral theories and deep learning



Spectral theories and deep learning



Complicated vs. Complex

Complicated

- Lots of Parts
- Linear or few nonlinearities
- Intuitive and /or understandable
- Example: Car, Phone, SVD

Complicated

- Few parts
- Many nonlinearities
- Counter-or unintuitive, difficult or impossible to understand
- Example: biology, conway's game of life

Complicated vs. Complex

$$\frac{dx}{dt} = \sigma(y - x)$$

$$\frac{dy}{dt} = x(\sigma - z) - y$$

$$\frac{dz}{dt} = xy - \beta z$$

https://en.wikipedia.org/wiki/Lorenz_system

Deep Learning: simple, complicated, complex

- Simple because it's made of simple, easy math(mostly)
- Complicated because it contains many many parts
- Complex because of myriad nonlinearities and because it is unintuitive and difficult to understand.

Spectral theories in mathematics

In this submodule, you have learned:

- The meaning of "spectral theory" and its relevance.
- How artificial neural networks relate to spectral theory.
- The distinction between "complicated" and "complex."
- Why deep learning is simultaneously easy, complicated, and complex!

Quiz-1

1. What is the core idea of spectral theory?
   a. Breaking down large problems into smaller, manageable components
   b. Ranking of data
   c. Sorting of data
   d. None of the above

2. Why is deep learning considered simple at a fundamental level?
   a. Because it relies on two basic operations: dot product and non-linearity
   b. Because it only uses dot products
   c. Because it only involves non-linear functions
   d. None of the above

3. Why is deep learning considered complicated?
   a. It consists of many many parts
   b. Due to the presence of non-linearity
   c. Because it only uses dot products
   d. None of the above

4. Why is deep learning considered complex?
   a. It is unintuitive and difficult to interpret
   b. It is difficult to understand
   c. It involves many interconnected components
   d. None of the above

Quiz-1

1. What is the core idea of spectral theory?
   a. Breaking down large problems into smaller, manageable components
   b. Ranking of data
   c. Sorting of data
   d. None of the above

2. Why is deep learning considered simple at a fundamental level?
   a. Because it relies on two basic operations: dot product and non-linearity
   b. Because it only uses dot products
   c. Because it only involves non-linear functions
   d. None of the above

3. Why is deep learning considered complicated?
   a. It consists of many many parts
   b. Due to the presence of non-linearity
   c. Because it only uses dot products
   d. None of the above

4. Why is deep learning considered complex?
   a. It is unintuitive and difficult to interpret
   b. It is difficult to understand
   c. It involves many interconnected components
   d. None of the above

2

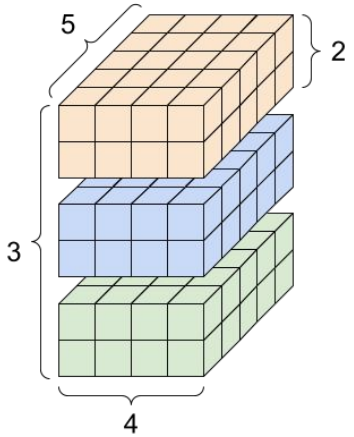
Terms and datatypes in math and computers

Terms and objects in math and computers

In this submodule, you will learn

- Some important terms in linear algebra and data storage.
- “Types” of numbers and variables

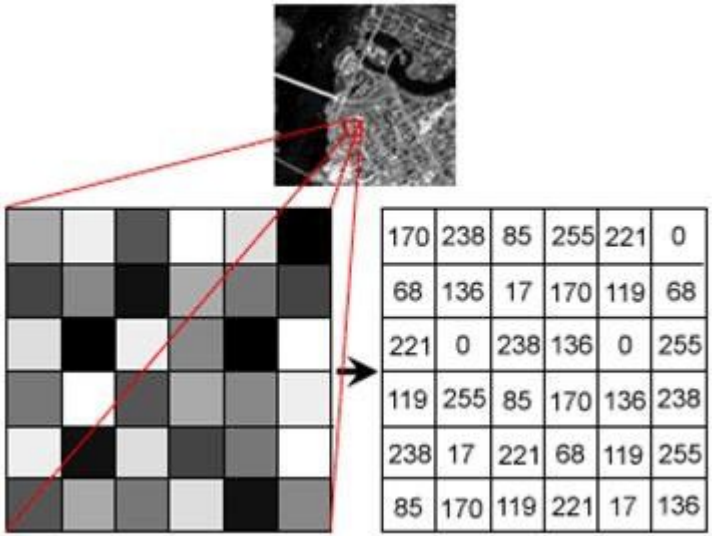
Linear algebra terminology

Object ⇨ 7 $\begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$ $\begin{bmatrix} -1 & 0 & 2 & 8 \\ 0 & 1 & 4 & 4 \\ 1 & 4 & 9 & 1 \end{bmatrix}$ 

Name ⇨ “Scalar” “Vector” “Matrix” “Tensor”

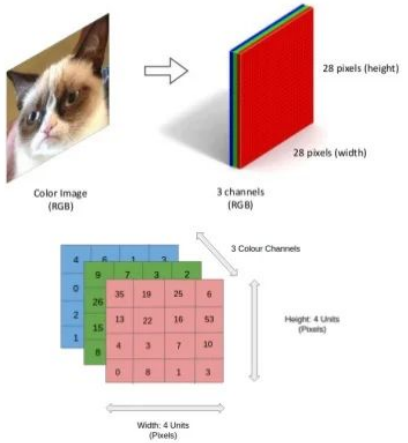
Storing images on computers

Grayscale image as matrix



170	238	85	255	221	0
68	136	17	170	119	68
221	0	238	136	0	255
119	255	85	170	136	238
238	17	221	68	119	255
85	170	119	221	17	136

color image is 3rd-order tensor



Disambiguation of "Data Type"

“Data type” in computer science	“Data type” in statistics
Format of data storage	Category of data
Implications: Operations, storage space	Implications: Appropriate statistical procedures
Examples: floating-point, boolean, string	Examples: Categorical, numerical, ordinal, ratio

Data types: disambiguation

"Data type" in computer science

Format of data storage

Implications: Operations, storage space

Examples: floating-point, boolean, string

4

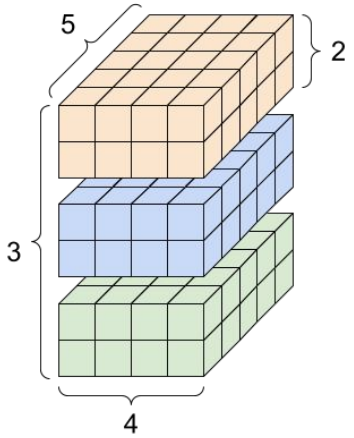
4.0

[1,2,3]

np.array([1,2,3])

torch.Tensor([1,2,3])

Linear algebra terminology

Object	7	$\begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$	$\begin{bmatrix} -1 & 0 & 2 & 8 \\ 0 & 1 & 4 & 4 \\ 1 & 4 & 9 & 1 \end{bmatrix}$	
Math	“Scalar”	“Vector”	“Matrix”	“Tensor”
Numpy	“Array”	“Array”	“ND array”	“ND array”
Pytorch	“Tensor”	“Tensor”	“Tensor”	“Tensor”

Terms and objects in math and computers

In this submodule, you have learned

- Some important terms in linear algebra and data storage
- Types of numbers and variables.

Quiz-2

1. What is a scalar in mathematical terms?
   a. A matrix with one row
   b. A single numerical value
   c. A 1D array
   d. A two-element tuple

2. Which of the following is an example of a 1-dimensional data structure?
   a. Matrix
   b. Scalar
   c. Vector
   d. Tensor

3. Which data type represents multi-dimensional arrays used in deep learning?
   a. Matrix
   b. Scalar
   c. Vector
   d. Tensor

4. What does the term “datatype” refer to in computing?
   a. Size of the memory
   b. Kind of value a variable holds
   c. Name of the variable
   d. Number of bits processed

5. Which of the following is not a numeric data type?
   a. Integer
   b. Float
   c. String
   d. Double

Quiz-2

1. What is a scalar in mathematical terms?
   a. A matrix with one row
   b. A single numerical value
   c. A 1D array
   d. A two-element tuple

2. Which of the following is an example of a 1-dimensional data structure?
   a. Matrix
   b. Scalar
   c. Vector
   d. Tensor

3. Which data type represents multi-dimensional arrays used in deep learning?
   a. Matrix
   b. Scalar
   c. Vector
   d. Tensor

4. What does the term “datatype” refer to in computing?
   a. Size of the memory
   b. Kind of value a variable holds
   c. Name of the variable
   d. Number of bits processed

5. Which of the following is not a numeric data type?
   a. Integer
   b. Float
   c. String
   d. Double

3

Converting reality to number

Converting reality to numbers

In this submodule, you will learn

- How to represent real-world outcomes using numbers
- The difference between “dummy-coding” and “one-hot encoding”

Two types of reality

Continuous

- Numeric
- Many (possibly infinite distinct values)
- Examples: height, exam scores, income, review score

Categorical

- Discrete
- Limited (typically a few distinct values)
- Examples: landscape (sea vs. mountain), picture identity (cat or dot), disease diagnosis

Representing categorical data

Dummy-Coding

- 0 or 1 (false or true)
- Creates a single vector
- Examples: exam (pass/fail), house(sold/market), fraud detection

One-hot encoding

- 0 or 1 per category
- Creates a matrix
- Examples: image recognition, hand-written letter recognition

Dummy-coding

Reality	y
Pass	1
Pass	1
Fail	0

$$Y= \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$$

One-hot encoding

Genre	History	Scifi	Kids
Y1	0	1	0
Y2	0	0	1
Y3	1	0	0

Y= $\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$

Converting reality to numbers

In this submodule, you have learned

- How to represent real-world outcomes using numbers
- The difference between “dummy coding” and “one-hot encoding”

Quiz-3

1. In machine learning, converting categorical variables like "red", "green", "blue" into numbers is called:
   a. Labeling
   b. One-hot encoding
   c. Normalizing
   d. Sampling

2. What kind of variables are converted to numbers using encoding schemes like label encoding or one-hot encoding?
   a. Continuous variables
   b. Categorical variables
   c. Boolean variables
   d. Time series data

3. Which real-world object is best represented as a time series when converting to numbers?
   a. A photograph
   b. A weather log recorded hourly
   c. A paragraph of text
   d. A pie chart

4. What is the main challenge in converting reality to numbers?
   a. Data storage
   b. Preserving interpretability and context
   c. Code optimization
   d. Network latency

5. Which of the following is an example of converting qualitative data to quantitative data?
   a. Assigning 1 to "Yes" and 0 to "No
   b. Taking an average temperature
   c. Counting number of words in a text
   d. Multiplying two features

Quiz-3

1. In machine learning, converting categorical variables like "red", "green", "blue" into numbers is called:
   a. Labeling
   b. One-hot encoding
   c. Normalizing
   d. Sampling

2. What kind of variables are converted to numbers using encoding schemes like label encoding or one-hot encoding?
   a. Continuous variables
   b. Categorical variables
   c. Boolean variables
   d. Time series data

3. Which real-world object is best represented as a time series when converting to numbers?
   a. A photograph
   b. A weather log recorded hourly
   c. A paragraph of text
   d. A pie chart

4. What is the main challenge in converting reality to numbers?
   a. Data storage
   b. Preserving interpretability and context
   c. Code optimization
   d. Network latency

5. Which of the following is an example of converting qualitative data to quantitative data?
   a. Assigning 1 to "Yes" and 0 to "No
   b. Taking an average temperature
   c. Counting number of words in a text
   d. Multiplying two features

4

Vector and matrix transpose

Vector and matrix transpose

In this submodule, you will learn

- How to interpret and use the transpose operation

Transposing a vector

$$\begin{bmatrix} 1 \\ 0 \\ 2 \\ 5 \\ -2 \end{bmatrix}^T = \begin{bmatrix} 1 & 0 & 2 & 5 & -2 \end{bmatrix}^T = \begin{bmatrix} 1 \\ 0 \\ 2 \\ 5 \\ -2 \end{bmatrix}$$

Transposing a matrix

$$\begin{bmatrix} 1 & 5 \\ 0 & 6 \\ 2 & 8 \\ 5 & 3 \\ -2 & 0 \end{bmatrix}^T = \begin{bmatrix} 1 & 0 & 2 & 5 & -2 \\ 5 & 6 & 8 & 3 & 0 \end{bmatrix}^T = \begin{bmatrix} 1 & 5 \\ 0 & 6 \\ 2 & 8 \\ 5 & 3 \\ -2 & 0 \end{bmatrix}$$

P_Code

A_Code

Vector and matrix transpose

In this submodule, you have learned

- How to interpret and use the transpose operation

Quiz-4

1. The transpose of the transpose of a matrix A is:
   a. Always the identity matrix
   b. The inverse of A
   c. A itself
   d. Negative of A

2. What is the result of (αA)^T , where α α is a scalar and A is a matrix?
   a. α^TA
   b. αA^T
   c. αA^T/α
   d. None of the above

3. Which of the following is always true about a vector $\vec{v}$?
   a. $\vec{v}$ ^T $\vec{v}$ is a matrix
   b. $\vec{v}$ ^T $\vec{v}$ is a scalar
   c. $\vec{v}$ ^T $\vec{v}$ is always 0
   d. $\vec{v}$ ^T $\vec{v}$ is a vector

4. For two vectors $\vec{a}$, $\vec{b}$ ∈ R^n a , the expression $\vec{a}$ ^T $\vec{b}$ is equivalent to:
   a. Matrix multiplication
   b. Vector addition
   c. Dot product
   d. Outer product

5. If A is an n × n symmetric matrix, then:
   a. A^T=−A
   b. A^T=A
   c. A^T not =A
   d. A^T=0

Quiz-4

1. The transpose of the transpose of a matrix A is:
   a. Always the identity matrix
   b. The inverse of A
   c. A itself
   d. Negative of A

2. What is the result of (αA)^T , where α α is a scalar and A is a matrix?
   a. α^TA
   b. αA^T
   c. αA^T/α
   d. None of the above

3. Which of the following is always true about a vector $\vec{v}$?
   a. $\vec{v}$ ^T $\vec{v}$ is a matrix
   b. $\vec{v}$ ^T $\vec{v}$ is a scalar
   c. $\vec{v}$ ^T $\vec{v}$ is always 0
   d. $\vec{v}$ ^T $\vec{v}$ is a vector

4. For two vectors $\vec{a}$, $\vec{b}$ ∈ R^n a , the expression $\vec{a}$ ^T $\vec{b}$ is equivalent to:
   a. Matrix multiplication
   b. Vector addition
   c. Dot product
   d. Outer product

5. If A is an n × n symmetric matrix, then:
   a. A^T=−A
   b. A^T=A
   c. A^T not =A
   d. A^T=0

5

Dot product

The dot product

In this submodule, you will learn

- Various notations for the dot product
- How to compute the dot product in vectors and matrices
- Why the dot product is so important in human civilization

Notations for and definition of the dot product

$$\alpha = a.b =< a, b \geq a^T b = \sum_{i=1}^{n} a_i b_i$$

Dot product

$$V = \begin{bmatrix} 1 \\ 0 \\ 2 \\ 5 \\ -2 \end{bmatrix} \qquad W = \begin{bmatrix} 2 \\ 8 \\ -6 \\ 1 \\ 0 \end{bmatrix}$$

$V^T W = 1*2+0*8+2*(-6)+5*1+(-2)*0 = -5$

Dot product

V

$$\begin{bmatrix} 1 \\ 0 \\ 2 \\ 5 \\ -2 \end{bmatrix}$$

W

$$\begin{bmatrix} 2 \\ 8 \\ -6 \end{bmatrix}$$

$V^TW=1*2+0*8+2*(-6)+5*?+(-2)*?$

Dot product in 2D

0	3	2
-3	-3	1
1	0	2

1	0	6
2	-1	0
5	1	4

0*1+3*0+2*6+-3*2+-3*-1+1*0+1*5+0*1+2*4=22

P-Code

A-Code

Interpretation of the dot product

A single number that reflects the commonalities between two objects (vectors, matrices, tensors, signals, images)

Applications of the dot product

The dot product is the computational backbone for many operations

- Statistics: Correlation, least-squares, entropy, PCA
- Signal processing: Fourier transform, filtering
- Science: Geometry, physics, mechanics
- Linear Algebra: Projection, transformations, multiplication
- Deep Learning: Convolution, matrix multiplication, Gram matrix(used in style transfer)

The dot product

In this submodule, you have learned

- Various notations for the dot product.
- How to compute the dot product in vectors and matrices
- Why the dot product is so important in human civilization

Quiz-4

1. What is the dot product of vectors a = [1 , 2] and b = [3 , 4]?
   a. 7
   b. 10
   c. 11
   d. 8

2. The result of the dot product of two vectors is a:
   a. Vector
   b. Matrix
   c. Scalar
   d. Tensor

3. If two vectors are perpendicular (orthogonal), their dot product is:
   a. 1
   b. 0
   c. Undefined
   d. Equal to their magnitudes

4. If $\vec{a}$ = [2 , − 1 , 3] and $\vec{b}$ = [4 , 0 , − 2] what is $\vec{a} \cdot \vec{b}$?
   a. 8
   b. 2
   c. -2
   d. 10

5. What is the dot product of a vector with itself, i.e., $\vec{a} \cdot \vec{a}$?
   a. Zero
   b. Square of its magnitude
   c. Its transpose
   d. Identity

Quiz-4

1. What is the dot product of vectors a = [1 , 2] and b = [3 , 4]?
   a. 7
   b. 10
   c. 11
   d. 8

2. The result of the dot product of two vectors is a:
   a. Vector
   b. Matrix
   c. Scalar
   d. Tensor

3. If two vectors are perpendicular (orthogonal), their dot product is:
   a. 1
   b. 0
   c. Undefined
   d. Equal to their magnitudes

4. If $\vec{a}$ = [2 , − 1 , 3] and $\vec{b}$ = [4 , 0 , − 2] what is $\vec{a} \cdot \vec{b}$?
   a. 8
   b. 2
   c. -2
   d. 10

5. What is the dot product of a vector with itself, i.e., $\vec{a} \cdot \vec{a}$?
   a. Zero
   b. Square of its magnitude
   c. Its transpose
   d. Identity

5

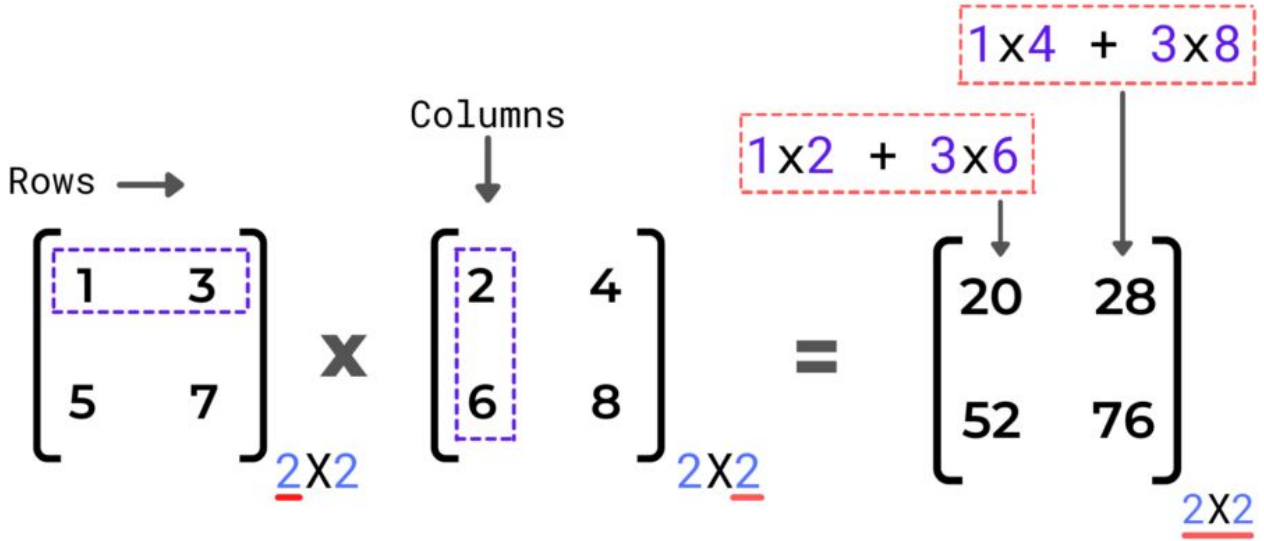
Matrix Multiplication

Matrix multiplication

In this submodule you will learn about

- How to refer to matrix sizes
- The rule for matrix multiplication validity
- One of the ways to conceptualize and implement matrix multiplication

Matrix Sizes:

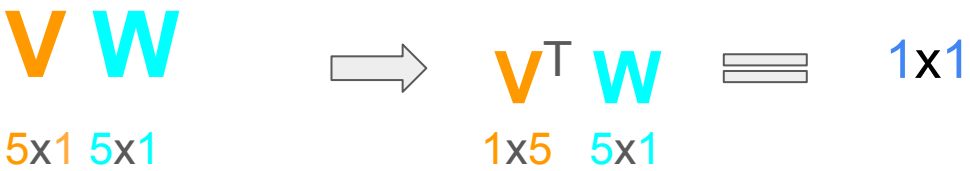


the resulting matrix will have the same number of rows as the first matrix and the same number of columns as the second matrix.

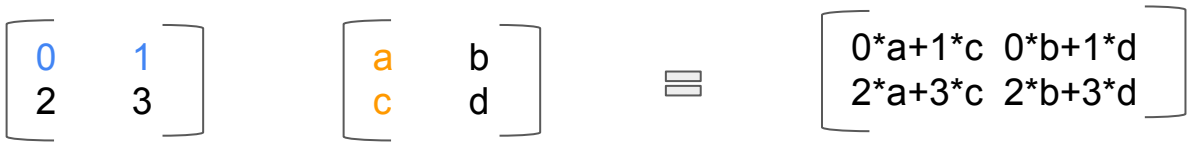
Matrix multiplication:Examples

A B **B A** **C A** **C^T A**

5x2 2 x 7 **2 x 7 5x2** **5 x 7 5x2** **7 x 5 5x2**


V W
5x1 5x1
V^T W
1x5 5x1
1x1

Matrix multiplication as ordered dot products



P-Code

A-Code

Matrix Multiplication

In this submodule, you have learned

- How to refer to matrix sizes
- The rule for matrix multiplication validity
- One of the ways to conceptualize and implement matrix multiplication.

Quiz-5

1. What is the condition for two matrices A and B to be multiplied?
   a. Number of columns in A = Number of rows in B
   b. Number of rows in A = Number of columns in B
   c. Number of columns in A = Number of columns in B
   d. A and B must be square matrices

2. A is a 2×3 matrix and B is a 3×4 matrix, what is the size of the product matrix AB?
   a. 3×2
   b. 2×3
   c. 3×4
   d. 2×4

3. Is matrix multiplication commutative? (i.e., is AB=BA)
   a. Always true
   b. Sometimes true
   c. Never true
   d. Only when A and B are identity matrices

4. What is the result of multiplying a matrix A with the identity matrix I ?
   a. A zero matrix
   b. Matrix A remains unchanged
   c. Transpose of A
   d. Inverse of A

5. If A is a 3×2 matrix and B is a 2×3 matrix, what will be the size of AB?
   a. 2×2
   b. 3×3
   c. 3×2
   d. 2×3

Quiz-5

1. What is the condition for two matrices A and B to be multiplied?
   a. Number of columns in A = Number of rows in B
   b. Number of rows in A = Number of columns in B
   c. Number of columns in A = Number of columns in B
   d. A and B must be square matrices

2. A is a 2×3 matrix and B is a 3×4 matrix, what is the size of the product matrix AB?
   a. 3×2
   b. 2×3
   c. 3×4
   d. 2×4

3. Is matrix multiplication commutative? (i.e., is AB=BA)
   a. Always true
   b. Sometimes true
   c. Never true
   d. Only when A and B are identity matrices

4. What is the result of multiplying a matrix A with the identity matrix I ?
   a. A zero matrix
   b. Matrix A remains unchanged
   c. Transpose of A
   d. Inverse of A

5. If A is a 3×2 matrix and B is a 2×3 matrix, what will be the size of AB?
   a. 2×2
   b. 3×3
   c. 3×2
   d. 2×3

6

Softmax

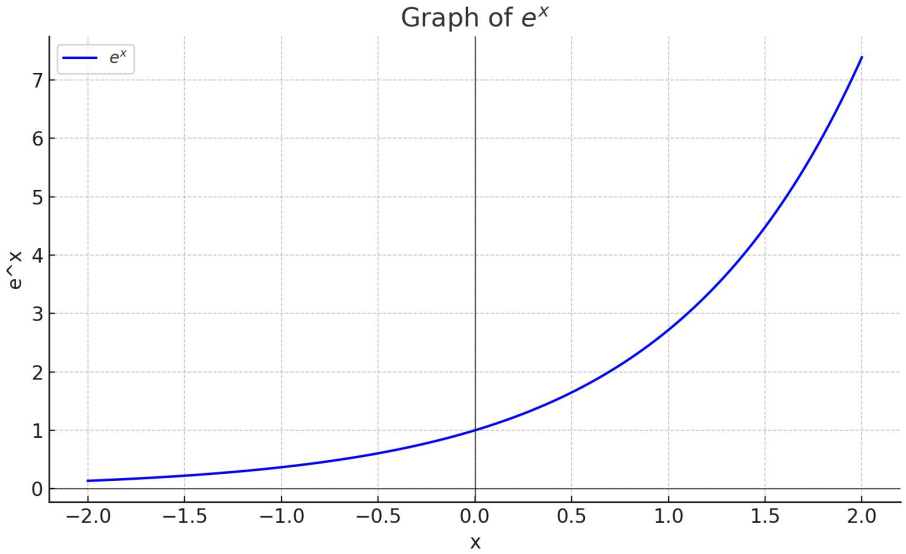
Softmax

In this submodule, you will learn

- The mathematical formula for the softmax
- The interpretation and purpose of softmax

The natural exponent

e=2.718….



- Strictly positive
- It never gets to zero
- This functions always be positive
- Natural exponent is used for softmax in order to generate probability

The Softmax formula: numerical example

$$\sigma_i = \frac{e^{z_i}}{\sum e^z}$$

$$z = \{1, 2, 3\}$$

$$e^z = \{2.72,\ 7, 39,\ 20.01\}$$

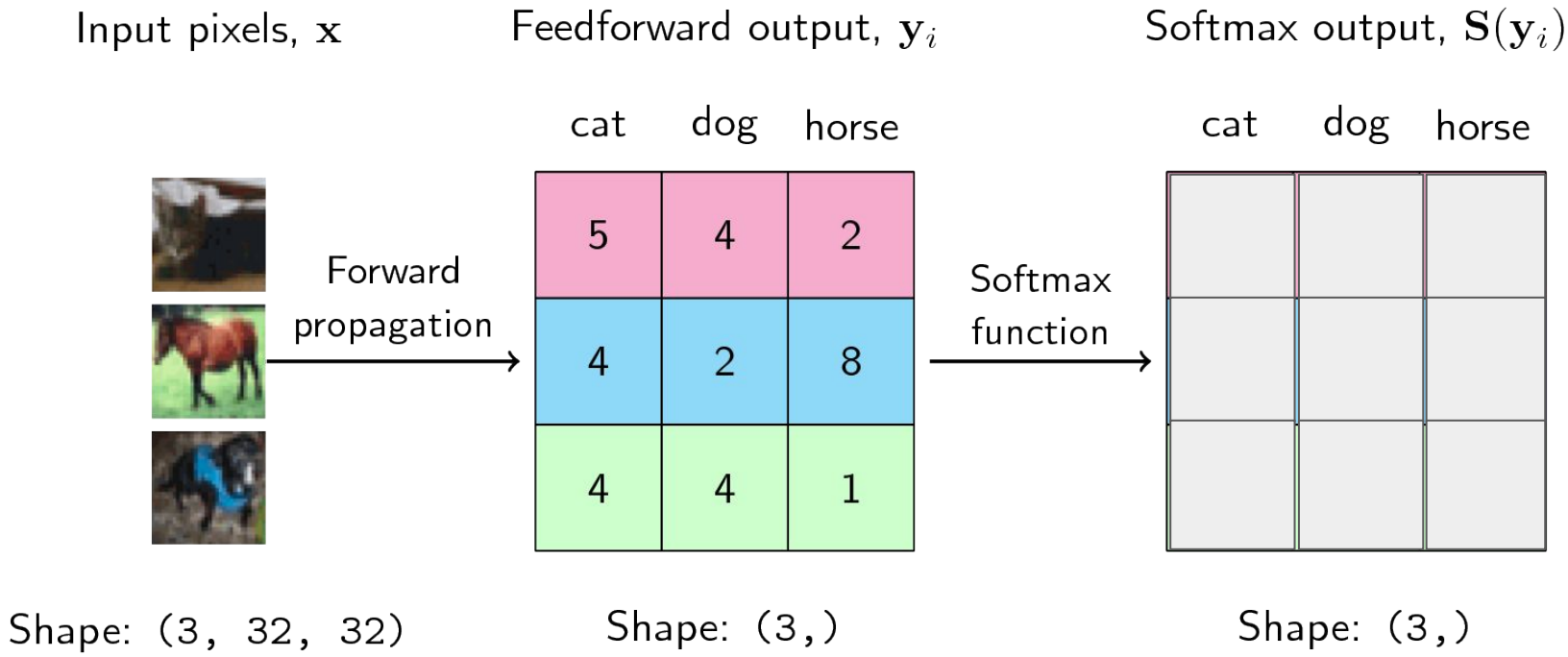
$$\Sigma\, e^z = 30.19$$

$$\sigma = \{0.09, .24, .67\}$$

The way to think about the softmax function

The way to think about the softmax function

Softmax input and output



Sum over inputs: any numerical value
Sum over outputs:Guaranteed to be 1.0

P-Code

A-Code

Softmax input and output

Input pixels, $\mathbf{x}$

Forward propagation →

Feedforward output, $\mathbf{y}_i$

cat	dog	horse
5	4	2
4	2	8
4	4	1

Softmax function →

Softmax output, $\mathbf{S}(\mathbf{y}_i)$

cat	dog	horse
0.71	0.26	0.04
0.02	0.00	0.98
0.49	0.49	0.02

Shape: (3, 32, 32)

Shape: (3,)

Shape: (3,)

Sum over inputs: any numerical value

Sum over outputs:Guaranteed to be 1.0

P-Code

A-Code

Softmax

In this submodule, you have learned

- The mathematical formula for the softmax
- The interpretation and purpose of softmax

Quiz-6

1. What is the primary role of the softmax function in machine learning?
   a. To normalize a vector into a probability distribution
   b. To compute the gradient of a loss function
   c. To reduce the dimensionality of input data
   d. To perform linear regression

2. The softmax function outputs values that are:
   a. Always greater than 1
   b. Any real number
   c. Between 0 and 1, summing to 1
   d. Negative and positive integers

3. Is the softmax function sigma(z)_i = e^(z_i) / sum_j e^(z_j) strictly monotonic with respect to each input z_i?
   a. Strictly increasing in z_i
   b. Strictly decreasing in z_i
   c. Non-monotonic in z_i
   d. Constant

4. If two inputs to the softmax function are equal (z_i = z_j), what can be said about their outputs?
   a. Their outputs are equal (sigma(z)_i = sigma(z)_j)
   b. Their outputs sum to 1
   c. Their outputs are zero
   d. Their outputs are undefined

5. Is the softmax function convex with respect to its input vector z?
   a. Always convex
   b. Always concave
   c. Neither convex nor concave
   d. Convex only for positive inputs

Quiz-6

1. What is the primary role of the softmax function in machine learning?
   a. To normalize a vector into a probability distribution
   b. To compute the gradient of a loss function
   c. To reduce the dimensionality of input data
   d. To perform linear regression

2. The softmax function outputs values that are:
   a. Always greater than 1
   b. Any real number
   c. Between 0 and 1, summing to 1
   d. Negative and positive integers

3. Is the softmax function sigma(z)_i = e^(z_i) / sum_j e^(z_j) strictly monotonic with respect to each input z_i?
   a. Strictly increasing in z_i
   b. Strictly decreasing in z_i
   c. Non-monotonic in z_i
   d. Constant

4. If two inputs to the softmax function are equal (z_i = z_j), what can be said about their outputs?
   a. Their outputs are equal (sigma(z)_i = sigma(z)_j)
   b. Their outputs sum to 1
   c. Their outputs are zero
   d. Their outputs are undefined

5. Is the softmax function convex with respect to its input vector z?
   a. Always convex
   b. Always concave
   c. Neither convex nor concave
   d. Convex only for positive inputs

7

Logarithms

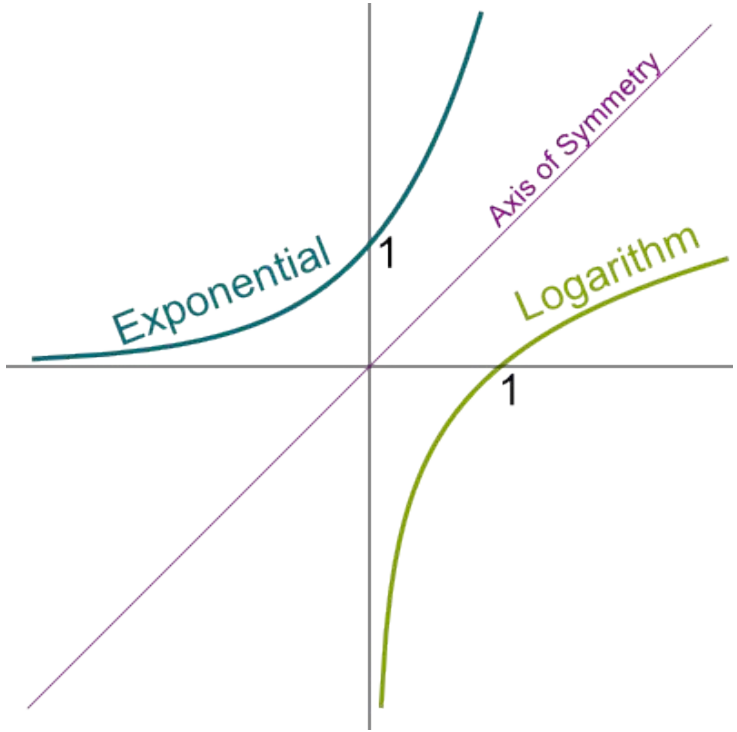
Logarithms

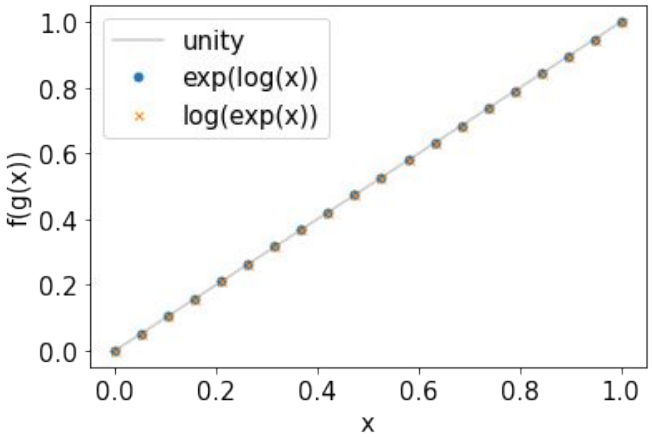
In this submodule, you will learn

- The logarithm(log) function
- Why logs are often used in ML and optimization

Logarithm: the inverse of the natural exponential function

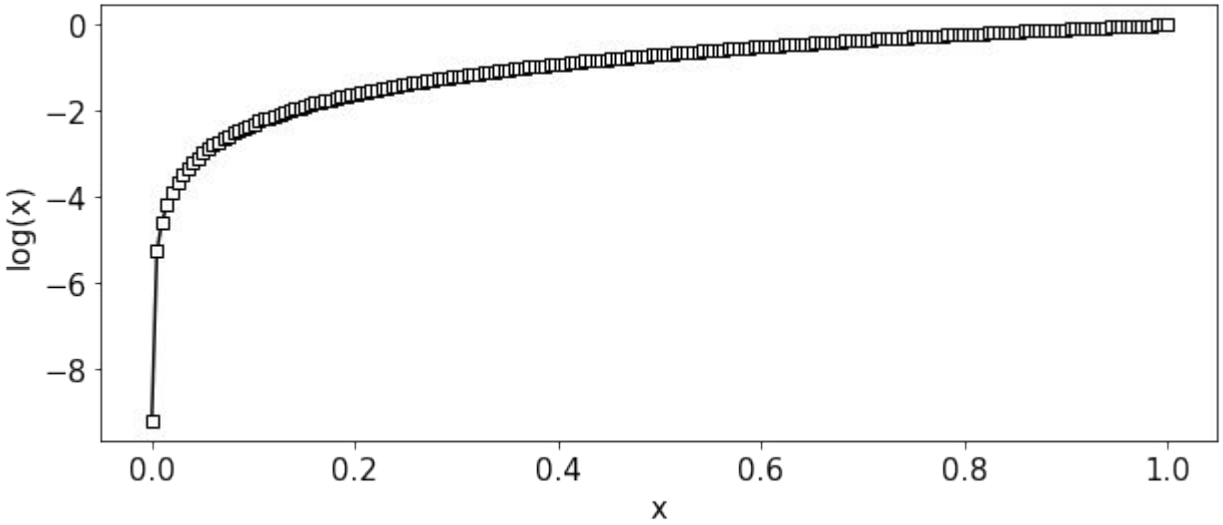
——-----exp(x)
——-----log(exp(x))

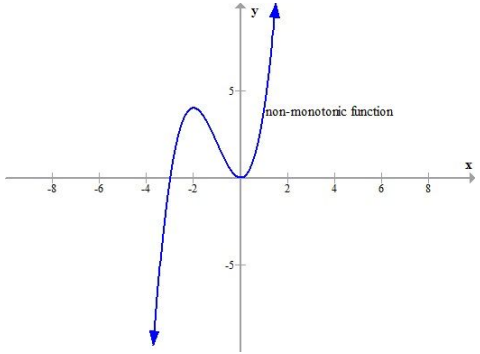




https://github.com/features/copilot/

Logarithm: The inverse of the natural exponential

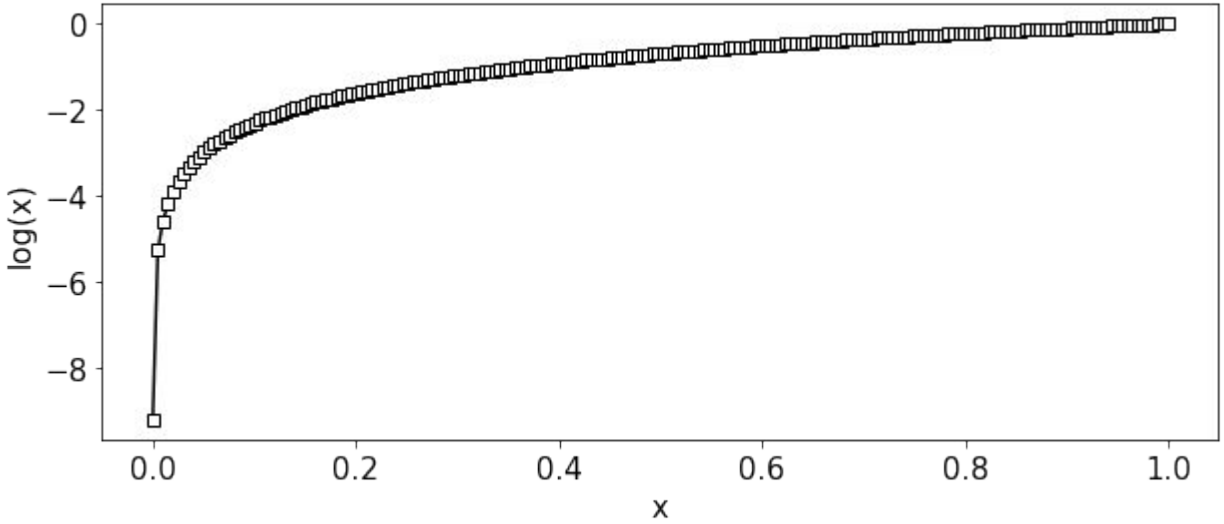


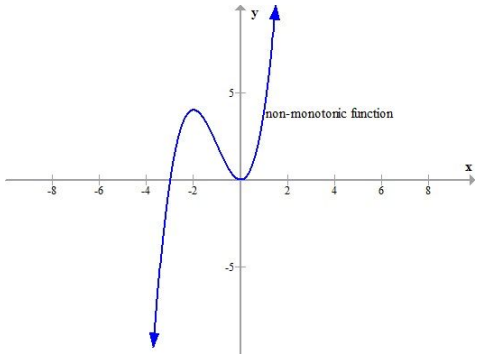


- Log is a monotonic function of x
- This is important because minimizing x is the same thing as minimizing log(x)! (n.b. Only for x>0)
- Log Stretches small values of x
- This is important because log better distinguishes small and closely spaced numbers

A-Code

Logarithm: The inverse of the natural exponential





- ML (and DL) often involves minimizing small quantities like probabilities
- Computers suffer from precision errors when working with very small numbers

P-Code

A-Code

Logarithms

In this submodule, you have learned

- The logarithm(log) function
- Why logs are often used in ML and optimization

Quiz-7

1. What is the inverse operation of a logarithm?
   a. Multiplication
   b. Division
   c. Exponentiation
   d. Subtraction

2. Which of the following is true about the logarithm of 1 (e.g., $\log_a(1)$)?
   a. 0
   b. 1
   c. It is always undefined
   d. It depends on the base

3. What is the value of $\log_{10}(100)$?
   a. 1
   b. 2
   c. 3
   d. 10

4. What is the primary purpose of a logarithm?
   a. To multiply numbers
   b. To find the exponent to which a base must be raised to produce a given number
   c. To divide numbers
   d. To add numbers

5. What does the logarithm base 10 represent?
   a. Natural logarithm
   b. Common logarithm
   c. Binary logarithm
   d. Exponential logarithm

Quiz-7

1. What is the inverse operation of a logarithm?
   a. Multiplication
   b. Division
   c. Exponentiation
   d. Subtraction

2. Which of the following is true about the logarithm of 1 (e.g., $\log_a(1)$)?
   a. 0
   b. 1
   c. It is always undefined
   d. It depends on the base

3. What is the value of $\log_{10}(100)$?
   a. 1
   b. 2
   c. 3
   d. 10

4. What is the primary purpose of a logarithm?
   a. To multiply numbers
   b. To find the exponent to which a base must be raised to produce a given number
   c. To divide numbers
   d. To add numbers

5. What does the logarithm base 10 represent?
   a. Natural logarithm
   b. Common logarithm
   c. Binary logarithm
   d. Exponential logarithm

8

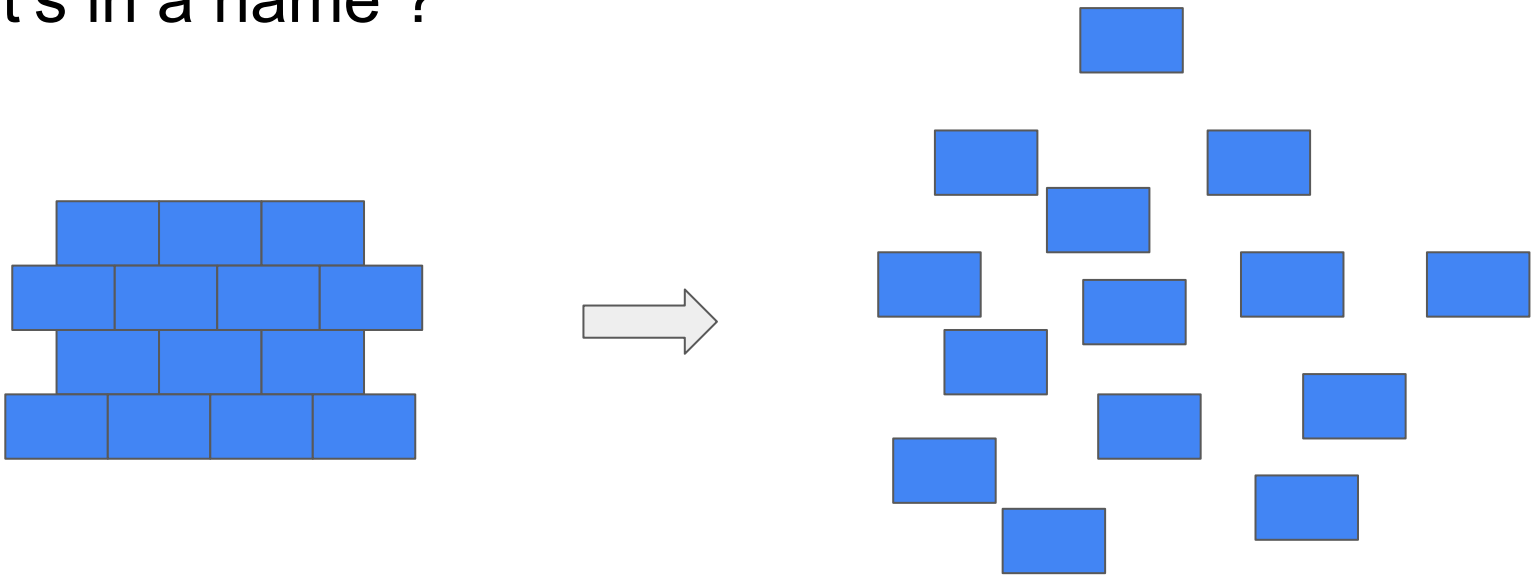
Entropy and cross-entropy

Entropy

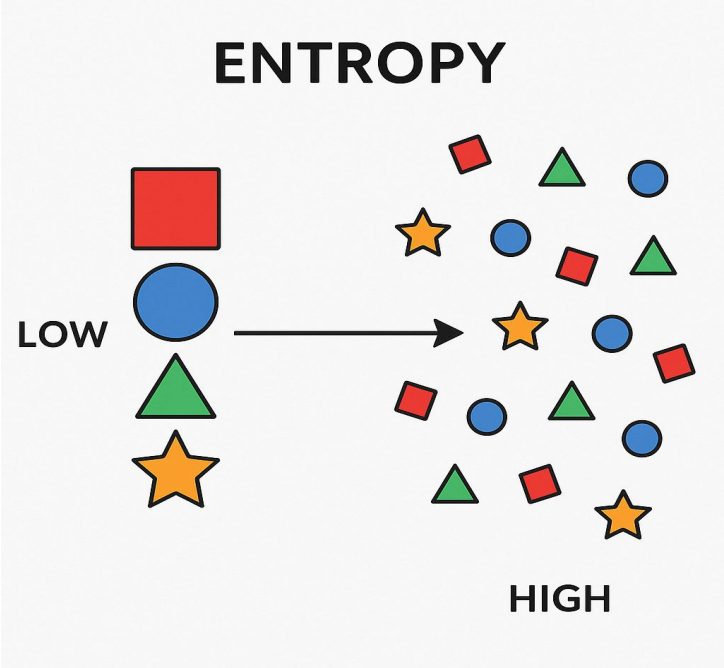
In this submodule, you will learn

- How to interpret entropy
- The formula for entropy
- The main application of entropy in deep learning

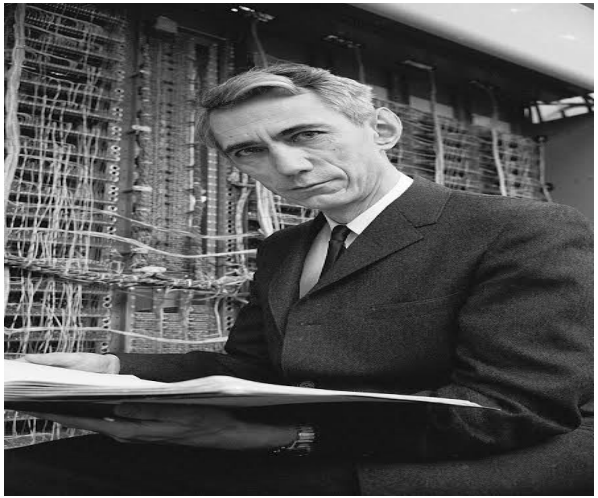
What’s in a name ?



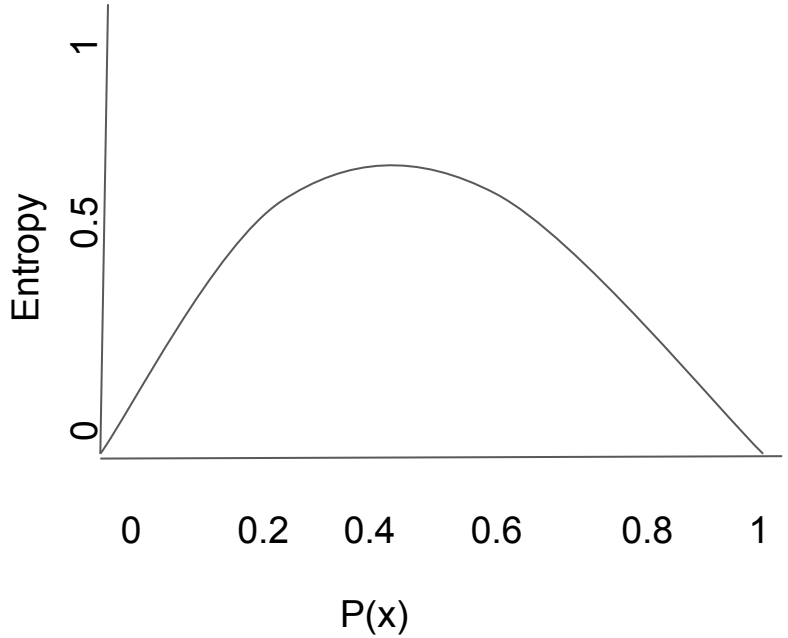
What’s in a name ?



What's in a name?



https://en.wikipedia.org/wiki/Claude_Shannon



Entropy in Information Theory:

- Entropy is a measure of the uncertainty or randomness in a set of outcomes.
- It is often used to quantify the amount of information contained in a message or a random variable.
- High entropy means high unpredictability or disorder, while low entropy means low unpredictability or more order.

Formula for entropy

$$H = -\sum_{i=1}^{n} p(x_i)\ \log_2\left(p(x_i)\right)$$

x=data values
p=probability

- High entropy means that the dataset has lot of variability.
- Low entropy means that most of the values of the dataset repeat (and therefore are redundant)

Important Note:

- How does entropy differ from variance?
- Entropy is nonlinear and makes no assumptions about the distribution.
- Variance depends on the validity of the mean and therefore is appropriate for roughly normal data.

Cross-Entropy

$$H(p) = -\sum p \log(p)$$

- Entropy describes one probability distribution

$$H(p, q) = -\sum p \log(q)$$

- Cross-entropy describes the relationship between two probability distributions.

(Note: in DL, events happen or don't happen->p=0 or p=1

P-Code

A-Code

Entropy and Cross entropy

In this submodule, you have learned

- How to interpret entropy
- The formula for entropy
- The main application of entropy in deep learning

Quiz-8

1. Which of the following distributions has the highest entropy for a given number of outcomes?
   a. Uniform distribution
   b. Gaussian distribution
   c. Binomial distribution
   d. Exponential distribution

2. What is the entropy of a system where the probability of all outcomes is 1?
   a. 0
   b. 1
   c. Infinite
   d. Undefined

3. Entropy is maximum when:
   a. All events have equal probability
   b. One event is more likely than others
   c. Only one event occurs
   d. No events occur

4. Which scientist introduced the concept of entropy in information theory?
   a. Isaac Newton
   b. Albert Einstein
   c. Claude Shannon
   d. Alan Turing

5. What does entropy measure in information theory?
   a. Speed of transmission
   b. Amount of noise in a signal
   c. Time taken to compress data
   d. Average amount of information in a message

Quiz-8

1. Which of the following distributions has the highest entropy for a given number of outcomes?
   a. Uniform distribution
   b. Gaussian distribution
   c. Binomial distribution
   d. Exponential distribution

2. What is the entropy of a system where the probability of all outcomes is 1?
   a. 0
   b. 1
   c. Infinite
   d. Undefined

3. Entropy is maximum when:
   a. All events have equal probability
   b. One event is more likely than others
   c. Only one event occurs
   d. No events occur

4. Which scientist introduced the concept of entropy in information theory?
   a. Isaac Newton
   b. Albert Einstein
   c. Claude Shannon
   d. Alan Turing

5. What does entropy measure in information theory?
   a. Speed of transmission
   b. Amount of noise in a signal
   c. Time taken to compress data
   d. Average amount of information in a message

9

Min/Max and argmin/argmax

Min/Max and argmin/argmax

In this submodule, you will learn

- The minimum and maximum
- The “argument” of the min/max functions.
- How to interpret the output of argmix/max functions.

Min/Max, argmin/argmax

$\min\{1,-1,3,0,4,3\}=-1$

$\max\{1,-1,3,0,4,3\}=4$

$\text{Arg min}\{1,-1,3,0,4,3\}=2$

$\text{Arg max}\{1,-1,3,0,4,3\}=5$

Min/Max, argmin/argmax

$$z = \arg_x \max f(x)$$

Application of argmax in deep learning

Model Input

Arg_y max M(y)=3

Model Output

Squirrel: p=0
Speed limit sign: p=0.05
Stop sign: p=0.8
Tomato:p=.1
Car:p=0.05

P-code

A- Code:

Min/Max and argmin/argmax

In this submodule, you have learned about

- The minimum and maximum
- The “argument” of the min/max functions.
- How to interpret the output of argmix/max functions.

Quiz-9

1. Which of the following is true about min and argmin?
   a. Both return values
   b. Both return indices
   c. min returns a value, argmin returns a location
   d. argmin returns a value, min returns a location

2. What does np.argmax([1, 3, 7, 4]) return in Python (NumPy)?
   a. 7
   b. 2
   c. 4
   d. [2,7]

3. If a function has multiple equal maximum values, argmax typically returns:
   a. All indices of maxima
   b. The average of maxima
   c. The first index of the maximum
   d. An error

4. In machine learning, which of the following is most commonly solved using argmin?
   a. Data visualization
   b. Loss function minimization
   c. Gradient ascent
   d. Feature selection

5. What is the purpose of argmin in optimization?
   a. To find the minimum value
   b. To find the index or input where the minimum occurs
   c. To normalize values
   d. To sort the array

Quiz-9

1. Which of the following is true about min and argmin?
   a. Both return values
   b. Both return indices
   c. min returns a value, argmin returns a location
   d. argmin returns a value, min returns a location

2. What does np.argmax([1, 3, 7, 4]) return in Python (NumPy)?
   a. 7
   b. 2
   c. 4
   d. [2,7]

3. If a function has multiple equal maximum values, argmax typically returns:
   a. All indices of maxima
   b. The average of maxima
   c. The first index of the maximum
   d. An error

4. In machine learning, which of the following is most commonly solved using argmin?
   a. Data visualization
   b. Loss function minimization
   c. Gradient ascent
   d. Feature selection

5. What is the purpose of argmin in optimization?
   a. To find the minimum value
   b. To find the index or input where the minimum occurs
   c. To normalize values
   d. To sort the array

10

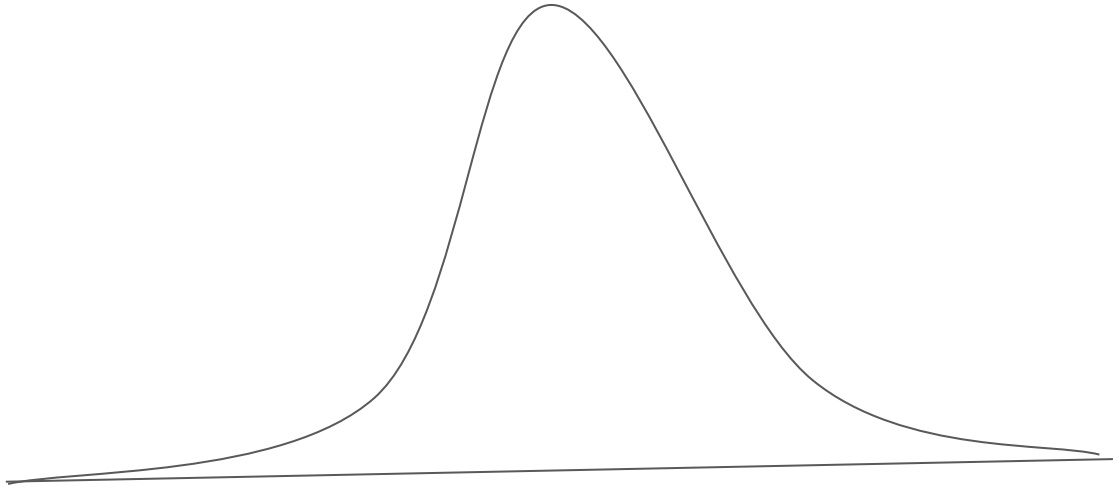
Mean and variance

Mean and Variance

In this submodule, you will learn

- The definition of “average”
- The definition and interpretation of “variance”

Central tendency



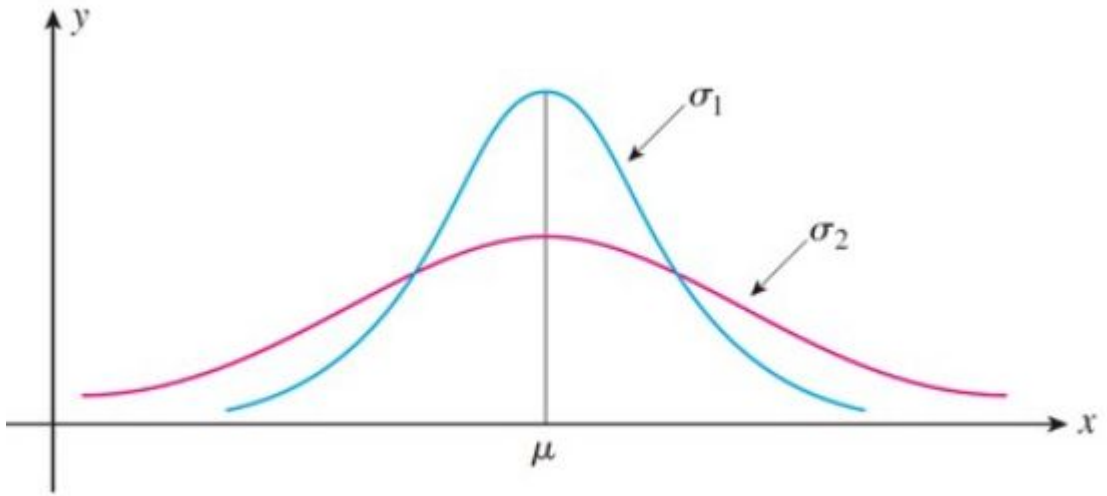
Mean (arithmetic mean, average)

Formula: $$\hat{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Suitable for:
Roughly normally distributed data

Suitable data types:
Interval, ration

Concept of dispersion



Two normal curves that have the same mean but different standard deviations

Variance

Formula: $$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Example:
x=[8,0,4,1,-2,7]

var(x)=sum([5,-3,1,-2,-5,4]^2)/5
var(x)=16
The numbers are wider spreaded

Suitable for:
Any distribution.

Suitable data types:
Numerical
Ordinal(but requires mean…)

Variance

Formula: $$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Example:
x=[2,3,4,3,4,4]

var(x)=0.667

Implication:the numbers are closer

Suitable for:
Any distribution.

Suitable data types:
Numerical
Ordinal(but requires mean…)

Some questions about the formula

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Why mean-center?

- Variance indicates the dispersion around the average:
- The following two datasets should have the same variance:

d1=[1,2,3,3,2,1]
d2=[101,102,103,103,102,101]

Some questions about the formula

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Why are the differences squared?

- We want the distances to the average;
- Without squaring the variance would be 0.

d1=[1,2,3,3,2,1]
Mean-centered d1=[-1,0,1,1,0,-1]=> sums to 0!

Some questions about the formula

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}|x_i - \bar{x}|$$

Why not take the absolute value(“mean absolute difference”)?

Squaring: emphasizes large values; is better for optimization (continuous and differentiable); is closer to Euclidean distance; is the second “moment” of the distributions; better link to least-squares regression; other nice properties

Also good; robust to outliers; less commonly used.

Standard deviation

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

P-code

A-code

Mean and Variance

In this submodule, you have learned

- The definition of “average”
- The definition and interpretation of “variance”

Quiz-10

1. What does the 'mean' of a dataset represent?
   a. The middle value
   b. The most frequent value
   c. The average value
   d. The highest value

2. Variance measures:
   a. Central tendency
   b. Skewness
   c. Dispersion or spread of data
   d. Frequency of data

3. If all values in a dataset are equal, then the variance is:
   a. 0
   b. 1
   c. Undefined
   d. Maximum

4. Which measure is sensitive to outliers?
   a. Median
   b. Mode
   c. Mean
   d. None of the above

5. What does a higher variance in a dataset imply?
   a. Data values are closer to the mean
   b. Data values are more spread out
   c. All data values are equal
   d. There are fewer data points

Quiz-10

1. What does the 'mean' of a dataset represent?
   a. The middle value
   b. The most frequent value
   c. The average value
   d. The highest value

2. Variance measures:
   a. Central tendency
   b. Skewness
   c. Dispersion or spread of data
   d. Frequency of data

3. If all values in a dataset are equal, then the variance is:
   a. 0
   b. 1
   c. Undefined
   d. Maximum

4. Which measure is sensitive to outliers?
   a. Median
   b. Mode
   c. Mean
   d. None of the above

5. What does a higher variance in a dataset imply?
   a. Data values are closer to the mean
   b. Data values are more spread out
   c. All data values are equal
   d. There are fewer data points

11

Random sampling and sampling variability

Sampling variability

In this submodule, you will learn

- What random sampling means and why we need it.
- The variability in sampling and the problems it can cause

Sampling variability:example

Scientific Questions:

How tall is the average Indian?

Don’t worry about the actual answer.

The question is :How do we know the answer?

Sampling variability:example

Random person-1:
160 cm
Random person-2:
192 cm

Etc.

But the internet says:
183.2cm

Sampling variability:definition

Sampling Variability:

Different samples from the same population can have different values of the same measurements.

Implication of sampling variability:

A single measurement may be an unreliable estimate of a population parameter

Sources of sampling variability

Natural variation: Often seen in biology(e.g. Height, weight) and physics(e.g. Earthquake magnitude, number of stars per galaxy).

Measurement noise: The sensors are imperfect (e.g. electrical line noise, measuring microgram with a gram-precision scale)

Complex systems: Measuring some factors while ignoring others (3.g. Measuring height while ignoring age)

Stochasticity(randomness):The universe is a wild and unpredictable place (3.g., photons hitting a camera lens)

What to do about sampling variability?

Take many samples! Averaging together many samples will approximate the true population mean (Law of Large Numbers).

Why sampling variability is important in DL

DL models learn by examples.

Non-random sampling can introduce systematic biases in DL models.

Non-representative sampling causes overfitting and limits generalizability

P-Code

A-Code

Sampling variability

In this submodule, you have learned about

- What random sampling means and and why we need it.
- The variability in sampling and the problems it can cause

Quiz-11

1. Which of the following reduces sampling variability?
   a. Smaller sample size
   b. Larger sample size
   c. Sampling with replacement
   d. Ignoring outliers

2. Why do different random samples give different results?
   a. Due to incorrect algorithms
   b. Because samples are chosen from different populations
   c. Due to natural variability in sample selection
   d. Because data gets corrupted during sampling

3. Which of these statements is true?
   a. Random samples always have zero variability
   b. Sampling variability increases with larger samples
   c. Every random sample from the same population may give different estimates
   d. Random sampling is only used in theoretical research

Quiz-11

1. Which of the following reduces sampling variability?
   a. Smaller sample size
   b. Larger sample size
   c. Sampling with replacement
   d. Ignoring outliers

2. Why do different random samples give different results?
   a. Due to incorrect algorithms
   b. Because samples are chosen from different populations
   c. Due to natural variability in sample selection
   d. Because data gets corrupted during sampling

3. Which of these statements is true?
   a. Random samples always have zero variability
   b. Sampling variability increases with larger samples
   c. Every random sample from the same population may give different estimates
   d. Random sampling is only used in theoretical research

12

Reproducible randomness via seeding

Reproducible randomness via seeding

In this submodule, you will learn

- How to use numpy's and PyTorch's seed functions.
- That there are multiple seed's in python, and you need to be mindful of which are set, and their scope.

P-Code

A-Code

Reproducible randomness via seeding

In this submodule, you have learned

- How to use numpy's and PyTorch's seed functions.
- That there are multiple seed's in python, and you need to be mindful of which are set, and their scope.

Quiz-12 

1. What does setting a seed in a random number generator help achieve?
   a. More randomness
   b. Less memory usage
   c. Reproducible results
   d. Faster computation

2. Which function is used to set the seed for reproducibility in NumPy?
   a. numpy.seed(42)
   b. numpy.random.set_seed(42)
   c. numpy.random.seed(42)
   d. set_seed(42)

3. In PyTorch, which function is used to set the seed for reproducibility?
   a. torch.manual_seed(seed)
   b. torch.set_seed(seed)
   c. random.seed(seed)
   d. numpy.random.seed(seed)

4. What will happen if you don't set a seed while using randomness in training a model?
   a. Your model won't train
   b. You'll always get the same results
   c. Results may vary on each run
   d. It'll use the system date as seed

5. Why is reproducibility important in scientific experiments involving randomness?
   a. To hide randomness
   b. To improve memory allocation
   c. To validate and verify results
   d. To reduce accuracy

Quiz-12 

1. What does setting a seed in a random number generator help achieve?
   a. More randomness
   b. Less memory usage
   c. Reproducible results
   d. Faster computation

2. Which function is used to set the seed for reproducibility in NumPy?
   a. numpy.seed(42)
   b. numpy.random.set_seed(42)
   c. numpy.random.seed(42)
   d. set_seed(42)

3. In PyTorch, which function is used to set the seed for reproducibility?
   a. torch.manual_seed(seed)
   b. torch.set_seed(seed)
   c. random.seed(seed)
   d. numpy.random.seed(seed)

4. What will happen if you don't set a seed while using randomness in training a model?
   a. Your model won't train
   b. You'll always get the same results
   c. Results may vary on each run
   d. It'll use the system date as seed

5. Why is reproducibility important in scientific experiments involving randomness?
   a. To hide randomness
   b. To improve memory allocation
   c. To validate and verify results
   d. To reduce accuracy

Acknowledgement

Thanks to the following courses and corresponding researchers for making their teaching/research material online

- Mike X Cohen's Master Deep Learning course (Udemy)
- AI Tools such as Chat GPT, Grok, etc
- Introduction to Deep Learning, University of Illinois at Urbana-Champaign
- Introduction to Deep Learning, Carnegie Mellon University
- Convolutional Neural Networks for Visual Recognition, Stanford University
- Natural Language Processing with Deep Learning, Stanford University
- And Many More .....