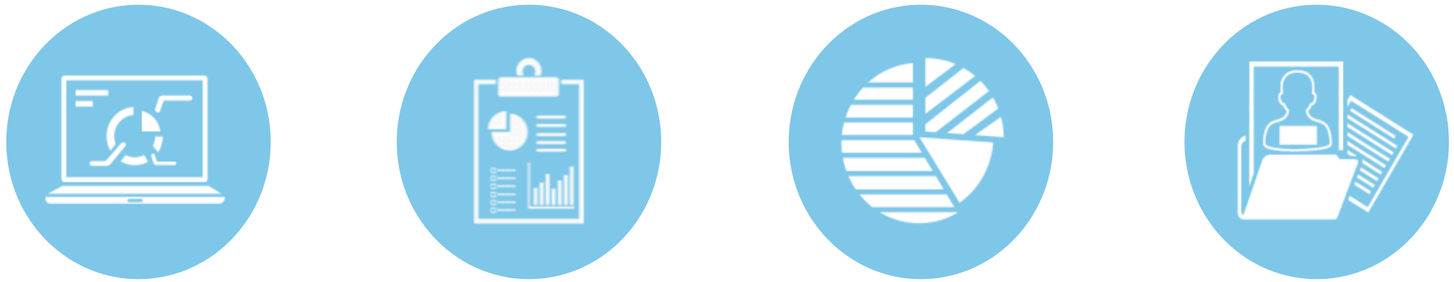
Machine Learning

Lesson 4—Math Refresher



Learning Objectives

- Explain the concepts of Linear Algebra
- Describe eigenvalues, eigenvectors, and eigendecomposition
- Define differential and integral calculus
- Explain the concepts of probability and statistics

Math Refresher

Topic 1—Concepts of Linear Algebra

Introduction to Linear Algebra

Linear algebra is a branch of mathematics that deals with the study of vectors and linear functions and equations.

Linear Equations

The main purpose of linear algebra is to find systematic methods for solving systems of linear equations.

A linear equation of n variables is of the form:

$$a_1x_1 + a_2x_2 + \ldots + a_nx_n = b$$

where x_1, x_2, ..., x_n are the unknown quantities to be found, a_1, $\cdots$, a_n are the coefficients (given numbers), and b is the constant term.

A linear equation does not involve any products, inverses, or roots of variables. All variables occur only to the first power and not as arguments for trigonometric, logarithmic, or exponential functions.

System of Linear Equations

A system of linear equations is a finite collection of linear equations.

A linear system of m equations in n variables has the form:

$$a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1$$
$$a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n = b_m$$

In the case of a single linear equation, a linear system can have infinitely many solutions, one solution, or no solutions at all.

A linear system that has a solution is called consistent, and the one with no solution is termed inconsistent.

Solving Linear Systems of Equations

Solving a linear system of equations is a long and tedious process.

The concept of **matrix** was introduced to simplify the computations involved in this process.

A matrix contains the essential information of a linear system in a rectangular array.

Matrix

A matrix of size m × n is a rectangular array of the form:

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

where the a_{ij}'s are the entries of the matrix, n represents the number of columns, and m represents the number of rows.

Forms of Matrix

- If n = m, that is, the number of columns and rows are equal, the matrix is called a square matrix.
- An entry of the form a_{ii} is said to be on the main diagonal.
- A is called a diagonal matrix if $a_{ij} = 0$, where $i \neq j$.

Matrix Operations

ADDITION

Consider the following two matrices:

$$A = \begin{pmatrix} 22 & 32 \\ 11 & 16 \end{pmatrix} \qquad B = \begin{pmatrix} 13 & 8 \\ 13 & 16 \end{pmatrix}$$

$$A + B = \begin{pmatrix} 22 + 13 & 32 + 8 \\ 11 + 13 & 16 + 16 \end{pmatrix}$$

The corresponding elements in the rows are added.

$$A + B = \begin{pmatrix} 35 & 40 \\ 24 & 32 \end{pmatrix}$$

Two matrices can be added only if they have the same number of rows and columns. Also, during addition, A + B = B + A

Matrix Operations

SUBTRACTION

Now consider the same matrices again:

$$A = \begin{pmatrix} 22 & 32 \\ 11 & 16 \end{pmatrix} \qquad B = \begin{pmatrix} 13 & 8 \\ 13 & 16 \end{pmatrix}$$

$$A - B = \begin{pmatrix} 22 - 13 & 32 - 8 \\ 11 - 13 & 16 - 16 \end{pmatrix}$$

The corresponding elements in the rows are subtracted.

$$A - B = \begin{pmatrix} 9 & 24 \\ -2 & 0 \end{pmatrix}$$

Two matrices can be subtracted only if they have the same number of rows and columns. Also, during subtraction, A - B ≠ B - A

Matrix Operations

MULTIPLICATION

Consider the same matrices again:

$$A = \begin{pmatrix} 22 & 32 \\ 11 & 16 \end{pmatrix} \qquad B = \begin{pmatrix} 13 & 8 \\ 13 & 16 \end{pmatrix}$$

Let AB = C. To compute the value of every element in the 2 x 2 matrix C, use the formula $C_{ik} = \Sigma_j A_{ij} B_{jk}$,

Dot product

$$A.B = \begin{pmatrix} (22 \times 13) + (32 \times 13) & (22 \times 8) + (32 \times 16) \\ (11 \times 13) + (16 \times 13) & (11 \times 8) + (16 \times 16) \end{pmatrix}$$

The 1st and 2nd rows of A are multiplied with the 1st and 2nd columns of B and added.

$$C = \begin{pmatrix} 702 & 688 \\ 351 & 344 \end{pmatrix}$$

The matrix product AB is defined only when the number of columns in A is equal to the number of rows in B. BA is defined only when the number of columns in B is equal to the number of rows in A. AB is not always equal to BA.

Matrix Operations

TRANSPOSE

A transpose is a matrix formed by turning all the rows of a given matrix into columns and vice versa. The transpose of matrix A is denoted as A^T

From the previous examples:

$$A = \begin{pmatrix} 22 & 32 \\ 11 & 16 \end{pmatrix}$$

$$A^T = \begin{pmatrix} 22 & 11 \\ 32 & 16 \end{pmatrix}$$

The rows become columns and vice versa.

Matrix Operations

ANALOGY TO DIVISION

Now that you know about matrix addition, subtraction, and multiplication, is division of matrices possible?

There is no matrix division, but there is a similar analogy called inverse.

Matrix Operations

INVERSE

An n-by-n square matrix A is called invertible (also nonsingular or nondegenerate) if there exists an n-by-n square matrix B such that

$$\mathbf{AB = BA = I_n}$$

where I_n denotes the n-by-n **identity matrix** and the multiplication used is ordinary matrix multiplication.

When the matrix B is uniquely determined by A, it is called the inverse of A, denoted by A^{-1}

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

Special Matrix Types

- Diagonal Matrix: a matrix $\boldsymbol{D}$ is diagonal only if $D_{i,j} = 0$ for all $i \neq j$
- Symmetric Matrix: a matrix A for which $\boldsymbol{A} = \boldsymbol{A}^T$
- Identity matrix: denoted as $\boldsymbol{I}_n$ such that $\mathbf{I}_n\mathbf{A} = \mathbf{A}$

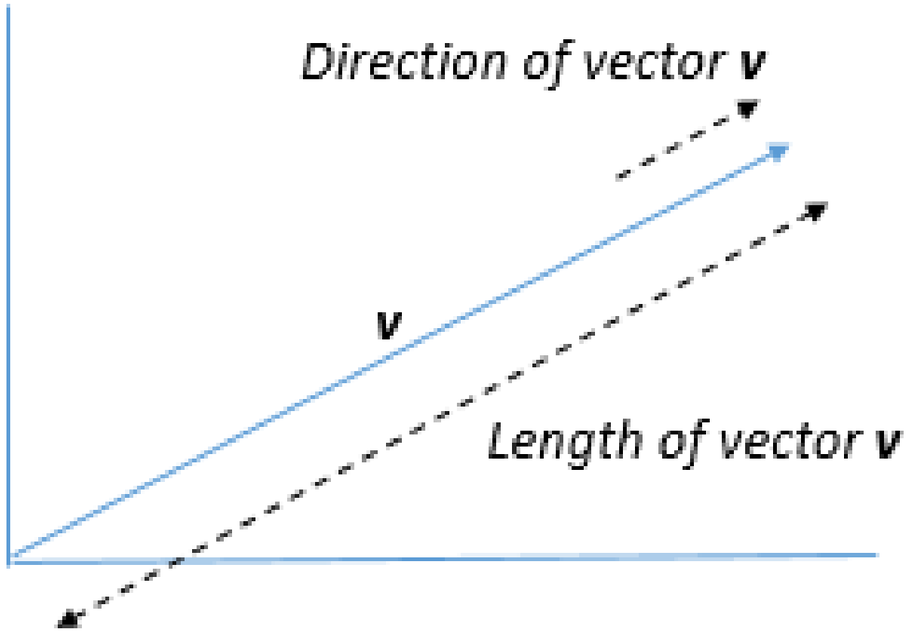
$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

An array with more than two axes is called a tensor. Example: A tensor might have 3 dimensions, so the value at coordinates (i, j, k) is $A_{i,j,k}$

What Is a Vector?

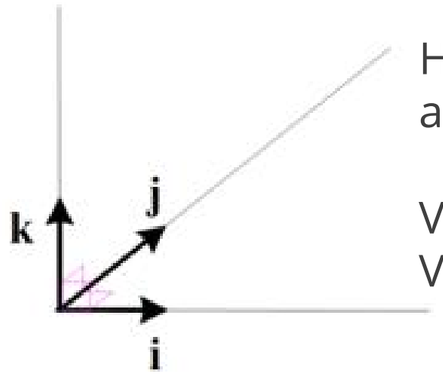
A vector (*v)* is an object with both magnitude (length) and direction.

It starts from origin (0,0), and its length is denoted by ||*v*||



Properties of Vectors

- A unit vector is a vector with unit norm (unit length) $||x||_2 = 1$
- A vector x and a vector y are orthogonal to each other if $x^Ty = 0$. This also means that both vectors are at a 90 degree angle to each other.
- An orthogonal vector that has a unit norm is called an orthonormal vector.



Here, vector i and k are orthogonal (orthonormal if they are assumed to be of unit norm).

Vectors k and j are not orthogonal.
Vectors j and i and not orthogonal.

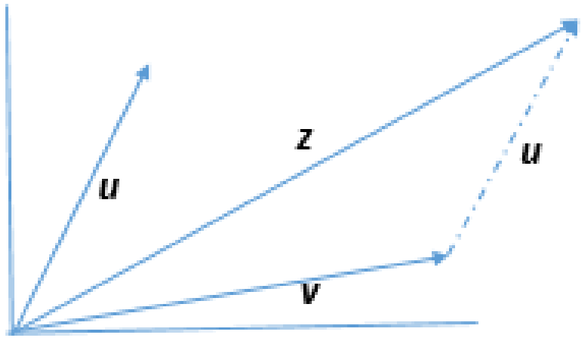
An orthogonal matrix is a square matrix whose rows are mutually orthonormal and whose columns are mutually orthonormal. In this case $A^TA = AA^T = I$. Also, for an orthogonal matrix, $A^{-1} = A^T$

Vector Operations

ADDITION AND SUBTRACTION

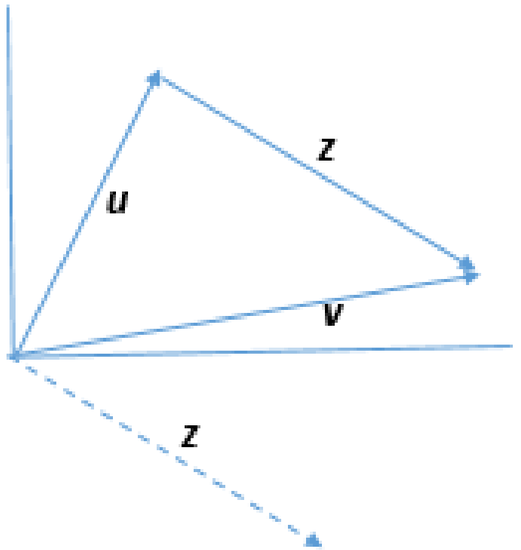
The operation of adding two or more vectors together into a vector sum is referred to as vector addition. For two vectors u and v, the vector sum is:

u + v = z



Vector subtraction is the process of subtracting two or more vectors to get a vector difference. For two vectors u and v, the vector difference is:

u - v = z



Vector Operations

MULTIPLICATION

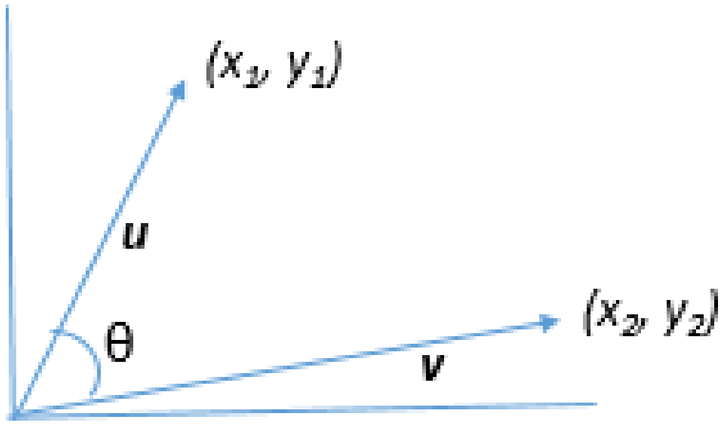
Vector multiplication refers to a technique for the multiplication of two (or more) vectors with themselves

$u*v = z$

$u*v = (x_1x_2 + y_1y_2) = \sum(x_iy_i)$

This can be shown to equal :

$u*v = \|x\| \|y\| \cos \theta$



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

Vector Operations

NORM

In machine learning, the size of a vector is called a norm. On an intuitive level, the norm of a vector x measures the distance from the origin to the point x.

Example: a vector with L^p norm where $p \geq 1$.

L^p norm

$$||\boldsymbol{x}||_p = \left(\sum_i |x_i|^p \right)^{\frac{1}{p}}$$

Vector Operations

NORM: FEATURES

- The most popular norm is the L^2 norm with $p = 2$, also called the **Euclidean norm**.
- It is simply the Euclidean distance from the origin to the point identified by $\boldsymbol{x}$.
- The 2 in the L^2 norm is frequently omitted, that is, $||x||_2$ is just written as $||x||$
- Size of vector is often measured as squared L^2 norm and equals $\boldsymbol{x}^T\boldsymbol{x}$. This is better as its differential only depends on x.
- L^1 norm is commonly used when the difference between zero and non-zero elements is very important. This is due to the fact that L^1 norm increases slowly in all directions from the origin.

$$\text{L1 norm, } p=1: \quad ||\boldsymbol{x}||_1 = \sum_i |x_i|$$

- Every time an element of x moves away from 0 by ϵ, the L^1 norm increases by ϵ.
- Max norm, infinite p: $||\boldsymbol{x}||_\infty = \max_i |x_i|$.

Math Refresher

Topic 2—Eigenvalues, Eigenvectors, and Eigendecomposition

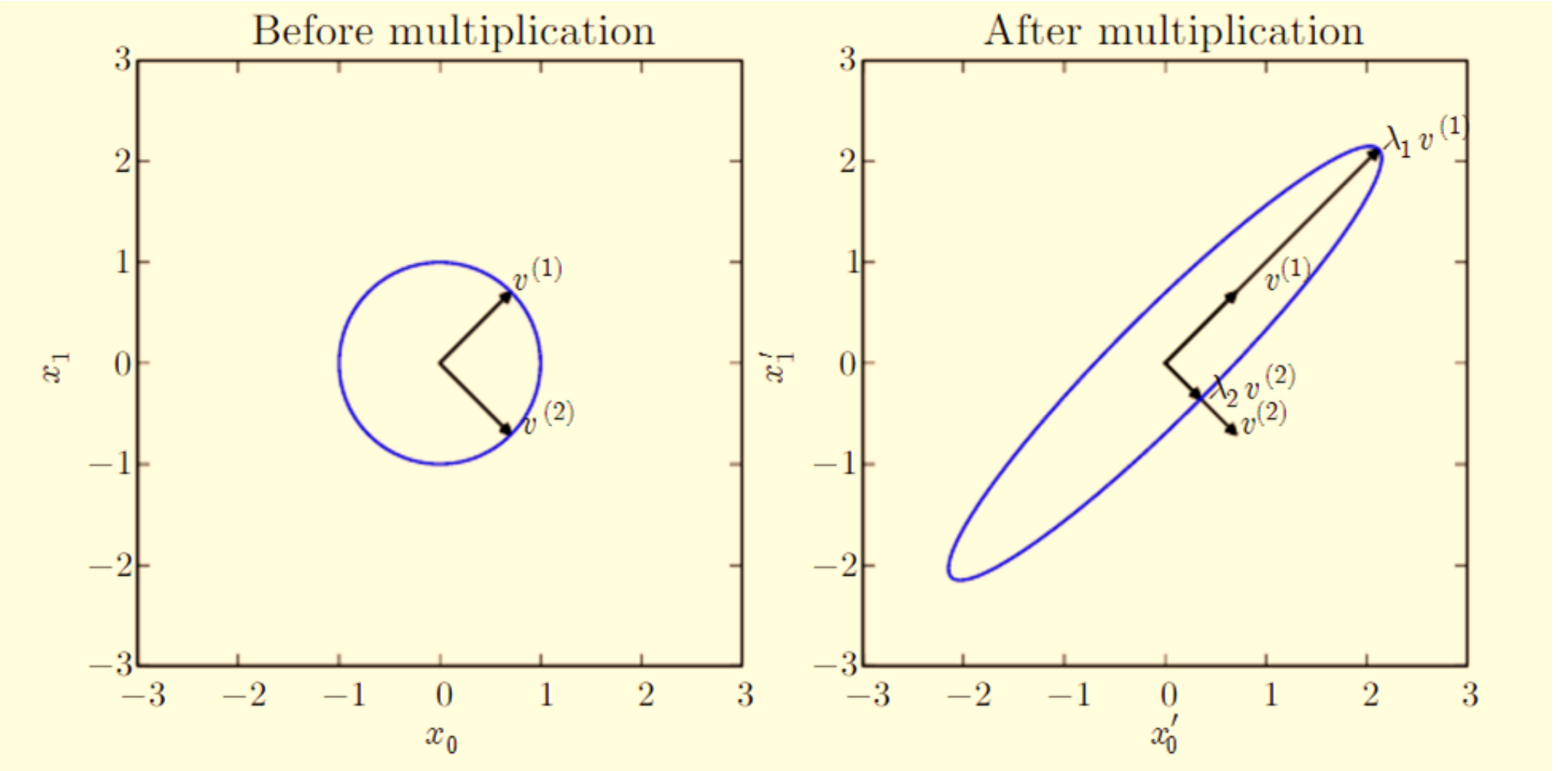
Eigenvector and Eigenvalue

An eigenvector of a square matrix A is a non-zero vector such that multiplication by A alters only the scale of v.

$$Av = \lambda v$$

where λ is eigenvalue corresponding to this eigenvector.

Effect of Eigenvectors and Eigenvalues



Here, matrix A has two orthonormal eigenvectors, $v^{(1)}$ with eigenvalue λ_1 and $v^{(2)}$ with eigenvalue λ_2. *(left).* Plot the set of all unit vectors u € R^2 as a unit circle. *(Right)* Plot the set of all points *Au*. By observing the way A distorts the unit circle, you can see that it scales space in the direction $v^{(i)}$ by λ_i

Eigendecomposition

Integers can be broken into their prime factors to understand them, example: 12 = 2 x 2 x 3. From this, useful properties can be derived, for example, the number is not divisible by 5 and is divisible by 2 and 3.

Similarly, matrices can be decomposed. This will help you discover information about the matrix.

If A has a set of eigenvectors v_1, v_2,... represented by matrix V, and the corresponding eigenvalues λ_1, λ_2,... represented by vector λ, then eigendecomposition of A is given by:

$$A = V \, diag(\lambda) \, V^{-1}$$

Math Refresher

Topic 3—Introduction to Calculus

What Is Calculus?

> Calculus is the study of change. It provides a framework for modelling systems in which there is change and ways to make predictions of such models.

http://www-math.mit.edu/~djk/calculus_beginners/chapter01/section02.html

Differential Calculus

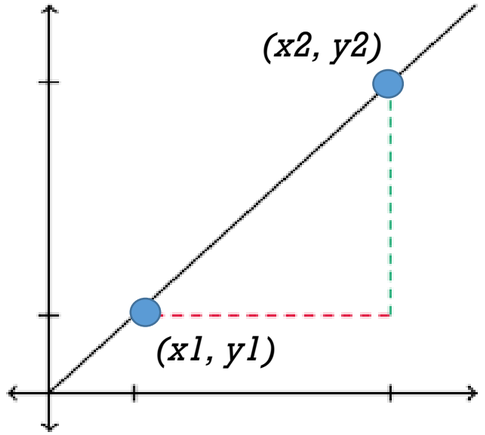
Differential calculus is a part of calculus that deals with the study of the rates at which quantities change.

Let x and y be two real numbers such that y is a function of x, that is, y = f(x).

If f(x) is the equation of a straight line (linear equation), then the equation is represented as y = mx + b.

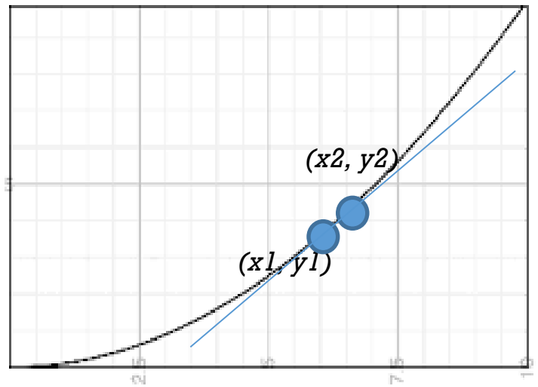
Where m is the slope determined by the following equation:

$$m = \frac{\text{change in } y}{\text{change in } x} = \frac{\Delta y}{\Delta x}$$



Δy/Δx or dy/dx is the derivative of y with respect to x and is also the rate of change of y per unit change in x.

Slope of a curvature changes at various points of the graph. It represents slope of an imaginary straight line drawn through that small graph segment.

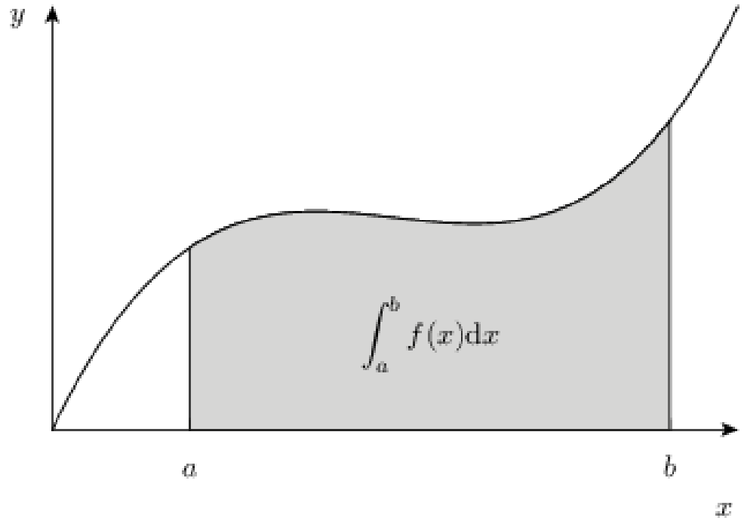


Integral Calculus

Integral Calculus assigns numbers to functions to describe displacement, area, volume, and other concepts that arise by combining infinitesimal data.

Given a function f of a real variable x and an interval [a, b] of the real line, the definite integral is defined informally as the signed area of the region in the xy-plane that is bounded by the graph of f, the x-axis, and the vertical lines x = a and x = b.

$$\int_a^b f(x)\,dx$$



An integral is the inverse of a differential and vice versa.

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

Math Refresher

Topic 4—Probability and Statistics

Probability Theory

Probability is the measure of the likelihood of an event's occurrence.

Example: The chances of getting heads on a coin toss is ½ or 50%

Probability of any specific event is between 0 and 1 (inclusive). The sum of total probabilities of an event cannot exceed 1, that is, 0 <= p(x) <= 1. This implies that ∫p(x)dx =1 (integral of p for a distribution over x)

Conditional Probability

Conditional probability of y=y given x=x is:

$$P(\mathrm{y} = y \mid \mathrm{x} = x) = \frac{P(\mathrm{y} = y, \mathrm{x} = x)}{P(\mathrm{x} = x)}$$

This is also called **Bayesian probability**.

Conditional Probability

BAYES MODEL EXAMPLE

Bayes model defines the probability of event A occurring, given event B has occurred.

P(A) = probability of event A
P(B) = probability of event B
P(A ∏ B) = probability of both events happening

Consider the coin example:

P(Coin1-H) = 2/4
P(Coin2-H) = 2/4
P(Coin1-H ∏ Coin2-H) = ¼
P(Coin1-H | Coin2-H) = (1/4)/(2/4) = ½ = 50% (probability of Coin1-H, given Coin2-H)

Two Coin Flip	
Coin 1	Coin 2
H	T
T	H
H	H
T	T

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \Rightarrow \quad P(A \cap B) = P(A|B)P(B)$$

Conditional Probability

SIMPLIFYING THE BAYES EQUATION

Events A and B are statistically independent if:

$$P(A \cap B) = P(A|B)P(B)$$

=> $$P(A \cap B) = P(A)P(B)$$

$P(A|B) = P(A)$ assumes P(B) is not zero

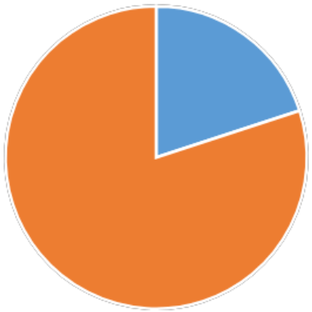
$P(B|A) = P(B)$ assumes P(A) is not zero

AI with Bayes Model: Example

Calculating the chance of developing diabetes given the incidence of fast food.

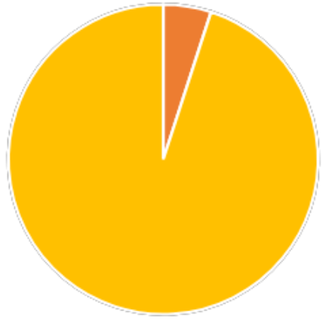
Observed Data

Fast Food audience 20%



Diabetes prevalence 10%

Fast Food and Diabetes 5%



Chances of Diabetes, given fast food: (conditional probability)

⇒ (D and F)/F = 5%/20% = ¼ = 25%

Analysis : If you eat fast food, you have 25% chance of developing Diabetes.

Chain Rule of Probability

Joint probability distribution over many random variables may be decomposed into conditional distributions over only one variable. It can be represented as:

$$P(\mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(n)}) = P(\mathrm{x}^{(1)}) \Pi_{i=2}^{n} P(\mathrm{x}^{(i)} \mid \mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(i-1)})$$

For example: **P(a, b, c) = P(a | b, c) * P(b | c) * P(c)**

Standard Deviation

Standard deviation is a quantity that expresses the value by which the members of a group differ from the mean value for the group.

- Its symbol is σ (the Greek letter sigma). If the data points are further from the mean, there is higher deviation within the data set.
- For example, a volatile stock has a high standard deviation, while the deviation of a stable blue-chip stock is usually rather low.

FORMULA

If X is a random variable with mean value (or expected value) μ such that: $\mathrm{E}[X] = \mu$

Then standard deviation : $\sigma = \sqrt{\mathrm{E}[(X - \mu)^2]}$

Standard deviation is used more often than variance because the unit in which it is measured is the same as that of mean, a measure of central tendency.

Variance

Variance (σ^2) refers to the spread of the data set, for example, how far the numbers are in relation to the mean.

- Variance is particularly useful when calculating the probability of future events or performance.
- Example: India has high variation in environmental temperature, from very hot to very cold weather. That is, India has high variance of weather conditions.

FORMULA

If X is a random variable with mean value (or expected value) μ such that: $\mathbf{E}[X] = \mu$

Then variance: $\mathbf{Var}(X) = \mathbf{E}\left[(X - \mu)^2\right]$

Notice that variance is just the square of standard deviation.

Covariance

Covariance is the measure of how two random variables change together. It is used to calculate the correlation between variables.

- A positive covariance indicates that both variables from the prior line tend to move upward and downward in value at the same time. An inverse or negative covariance means that variables move counter to each other: when one rises, the other falls

- Example: Purchasing stock with a negative covariance is a great way to minimize risk in a portfolio (diverse portfolio)

FORMULA

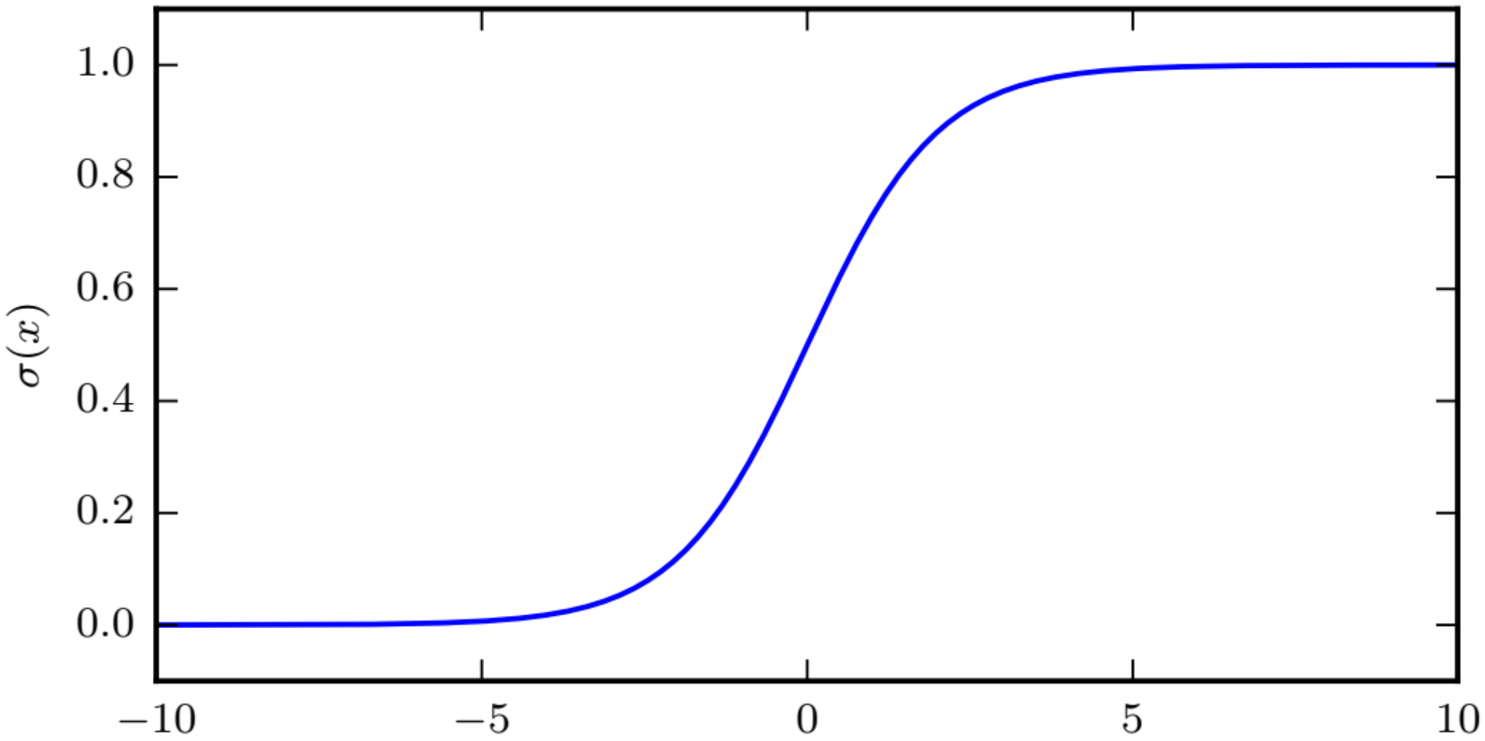
If X is a random variable with mean value (or expected value) μ such that: $\mathrm{E}[X] = \mu$

Then covariance: $\operatorname{cov}(X, Y) = \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right]$

Logistic Sigmoid

The Logistic Sigmoid is a useful function that follows the S curve. It saturates when input is very large or very small.

$$\sigma(x) = \frac{1}{1 + \exp(-x)}$$

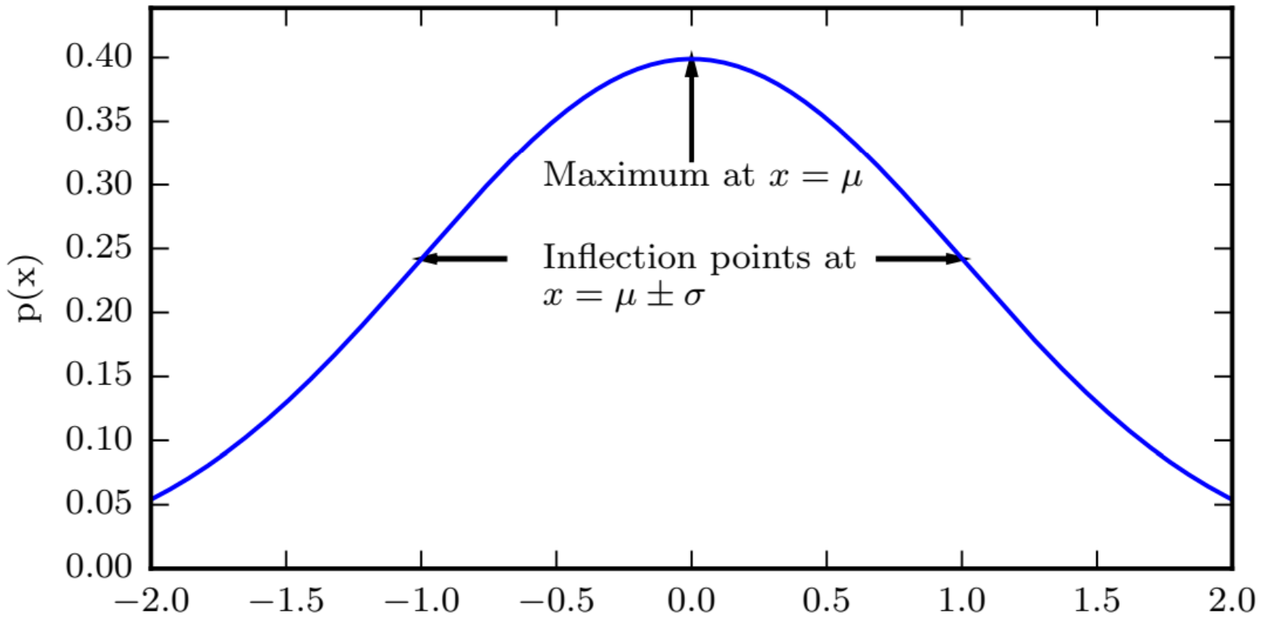


Gaussian Distribution

DEFINITION

Data can be distributed in various ways.

The distribution where the data tends to be around a central value with lack of bias or minimal bias toward the left or right is called Gaussian distribution, also known as normal distribution.



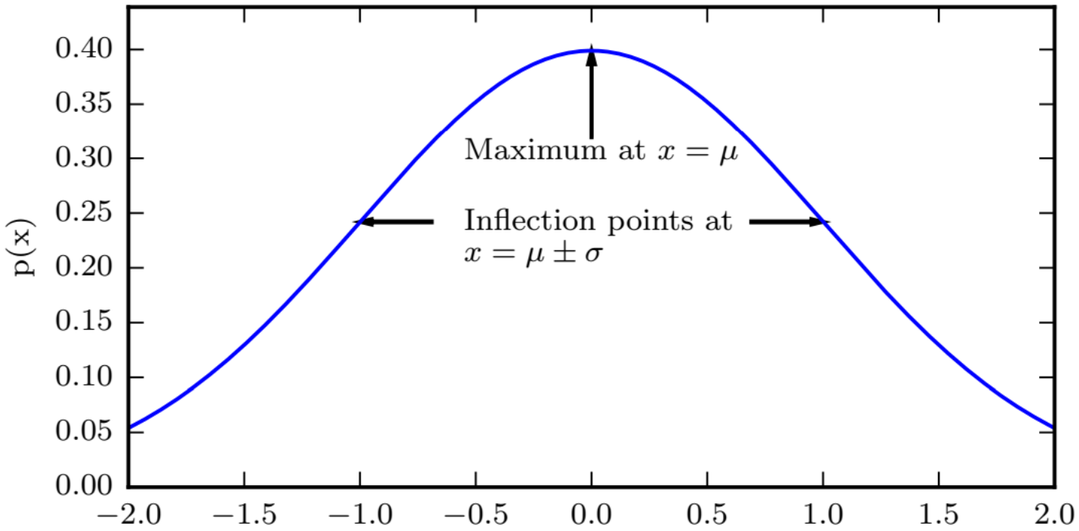
In absence of prior knowledge, normal distribution is often a good assumption in machine learning.

Gaussian Distribution

EQUATION

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

$$\mathcal{N}(x;\mu,\beta^{-1}) = \sqrt{\frac{\beta}{2\pi}}\exp\left(-\frac{1}{2}\beta(x-\mu)^2\right)$$

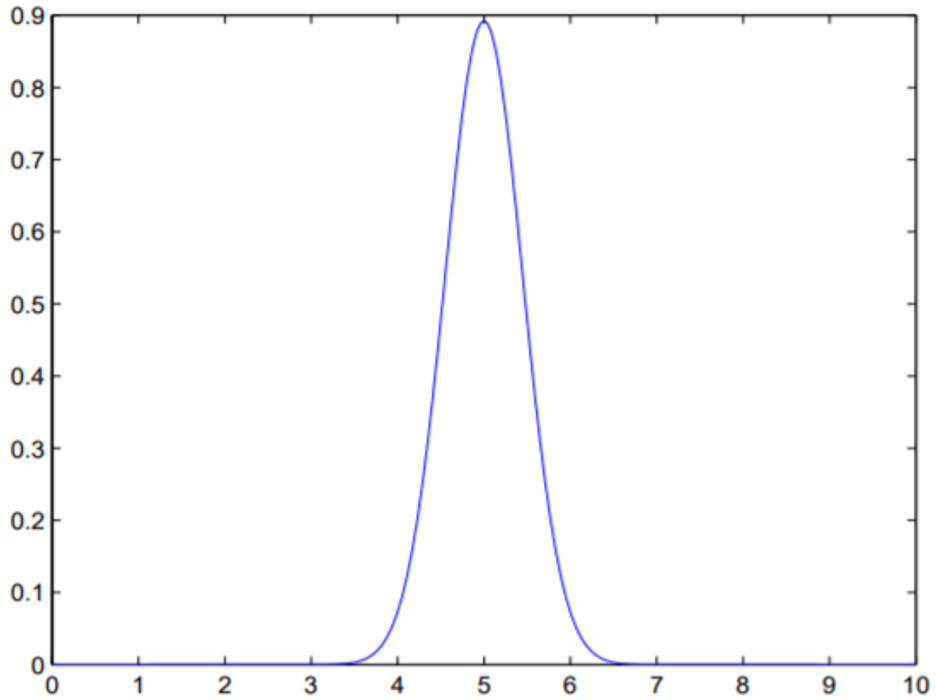


- μ = mean or peak value, which also means E[x] = μ
- σ = standard deviation and σ^2 = variance
- A “standard normal distribution” has μ = 0 and σ = 1
- For efficient handling, invert σ and use precision β (inverse variance) instead

Types of Gaussian Distribution

UNIVARIATE

Univariate Gaussian distribution refers to the distribution over a single variable.

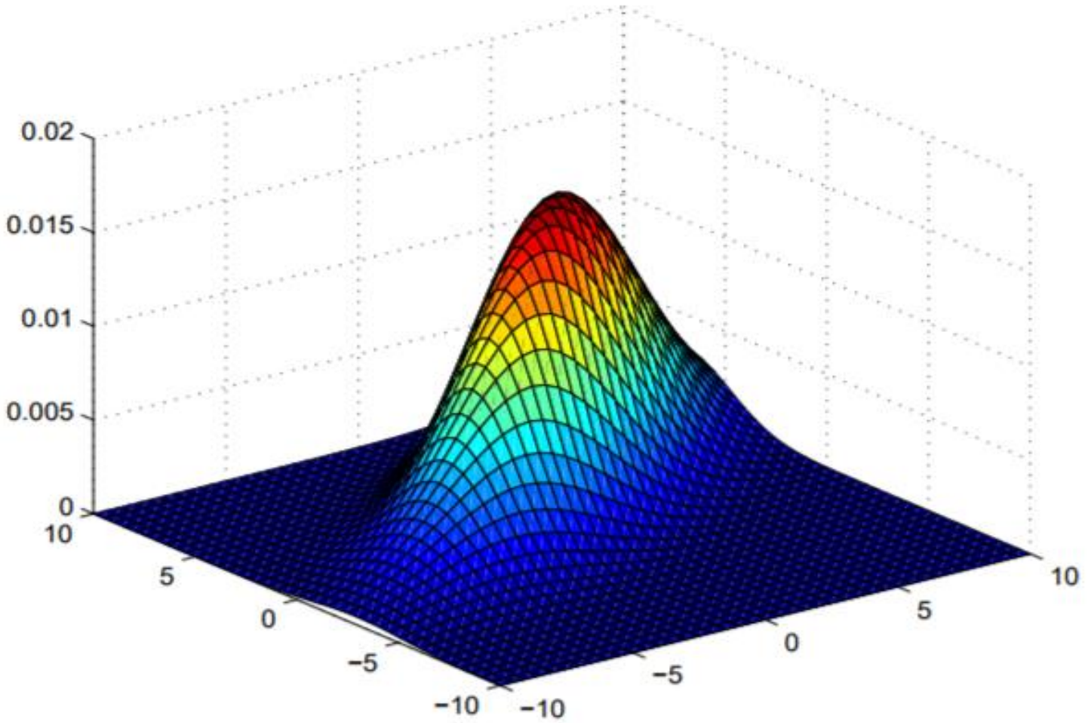


Univariate Gaussian distribution over single variable x

Types of Gaussian Distribution

MULTIVARIATE

Multivariate normal distribution is the generalization of the univariate normal distribution to multiple variables.



Multivariate Gaussian distribution over two variables x_1 and x_2

Key Takeaways

- Linear algebra is a branch of mathematics that deals with the study of vectors and linear functions and equations.
- A matrix of size m × n is a rectangular array.
- A vector (*v*) is an object with both magnitude (length) and direction. It is represented by an arrow on a graph.
- Eigenvector of a matrix A is a vector v that only changes the scale of the vector v when multiplied by A.
- Differential Calculus is the incremental rate of change of dependent variable y with respect to x. Integral Calculus is summation of a function f(x) over x.
- Probability is the chance of something happening.
- Standard deviation and variance indicate the spread of a data distribution around its mean.

Quiz

QUIZ 1

Matrix Multiplication of two matrices A and B involves:

a. multiplying rows of both matrices

b. multiplying columns of both matrices

c. multiplying both matrices diagonally

d. None of the above

QUIZ 1

Matrix Multiplication of two matrices A and B involves:

a. multiplying rows of both matrices

b. multiplying columns of both matrices

c. multiplying both matrices diagonally

d. None of the above

The correct answer is **d. None of the above**

Matrix Multiplication of two matrices A and B involves the 1st and 2nd rows of A multiplied with the 1st and 2nd columns of B and added.

QUIZ 2

When are two vectors v_1 and v_2 considered orthogonal?

a. When both are at 90 degrees to each other

b. When both are at 180 degrees to each other

c. When both are pointing in the same direction

d. When both are pointing in the opposite direction

QUIZ 2

When are two vectors v_1 and v_2 considered orthogonal?

a. When both are at 90 degrees to each other

b. When both are at 180 degrees to each other

c. When both are pointing in the same direction

d. When both are pointing in the opposite direction

The correct answer is **a. When both are at 90 degrees to each other**

Two vectors v_1 and v_2 are considered orthogonal when both are at 90 degrees to each other.

This concludes “Math Refresher.”

The next lesson is “Regression.”