BITS Pilani
Pilani Campus

M.Tech DSE
Machine Learning

Dr. Monali Mavani

Source: "Probabilistic Machine Learning, An Introduction", Kevin P. Murphy, Slides of Prof. Chetana from BITS Pilani , CS109 stanford lecture notes and many others who made their course materials freely available online.

- Lecture No. – 2 | Math and Stat Preliminaries

Session Content

- Linear Algebra Review
- Calculus Review
- Probability Theory
- Decision Theory

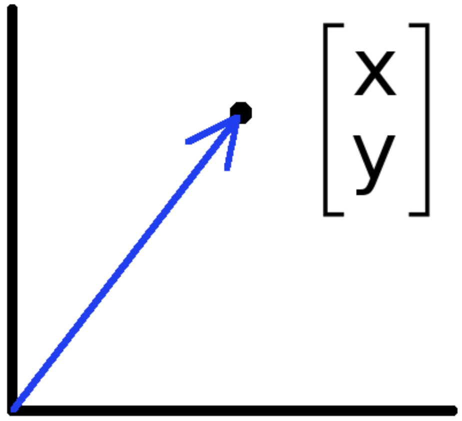
Linear algebra Review

Vectors and Matrices

- Collections of ordered numbers that represent movements in space, word counts, movie ratings, pixel brightness, etc.
- Vector is a mathematical quantity that has magnitude and direction

Vectors

- Vectors can represent an offset in 2D or 3D space
- Points are just vectors from the origin



- Data can also be treated as a vector

Vector

- A column vector $\mathbf{v} \in \mathbb{R}^{n \times 1}$ where

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

- A row vector $\mathbf{v}^T \in \mathbb{R}^{1 \times n}$ where

$$\mathbf{v}^T = \begin{bmatrix} v_1 & v_2 & \dots & v_n \end{bmatrix}$$

T denotes the transpose operation

Product of 2 Vectors

Three ways to multiply

- Element-by-element
- Inner product
- Outer product
- Cross product

Element-by-element product (Hadamard product)

$$\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} .* \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} a_1 b_1 \\ a_2 b_2 \end{pmatrix}$$

Multiplication: Dot product (inner product)

The dot product represents the similarity between vectors as a single number:

If two vectors are in the same direction the dot product is positive and if they are in the opposite direction the dot product is negative.

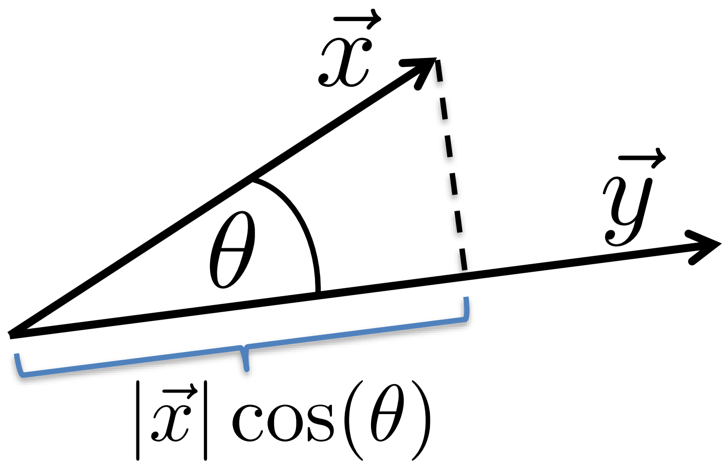
$$\vec{x} \cdot \vec{y} =$$

$$\begin{pmatrix} x_1 & x_2 & \cdots & x_N \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = x_1 y_1 + x_2 y_2 + \cdots + x_N y_N$$

1 X N N X 1 1 X 1

Outer dimensions give size of resulting matrix

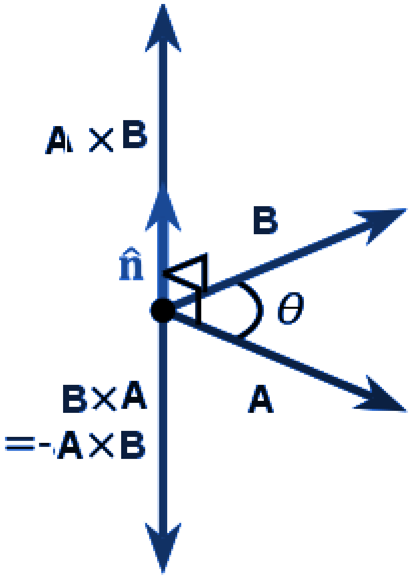
Dot product geometric intuition: "Overlap" of 2 vectors



$$\vec{x} \cdot \vec{y} = |\vec{x}||\vec{y}| \cos(\theta)$$

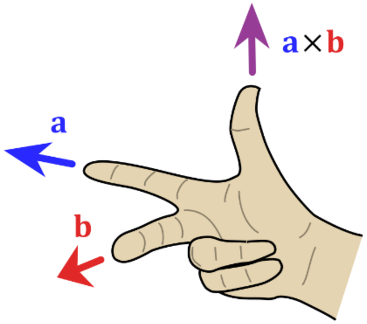
Cross product

The cross product of two vectors **a** and **b** is defined only in three-dimensional space and is denoted by **a** × **b**.



$$\mathbf{a} \times \mathbf{b} = \|\mathbf{a}\| \, \|\mathbf{b}\| \sin(\theta) \, \mathbf{n}$$

n is a unit vector perpendicular to the plane containing **a** and **b**, in the direction given by the right-hand rule



Multiplication: Outer product

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} \begin{pmatrix} y_1 & y_2 & \cdots & y_M \end{pmatrix} = \begin{pmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_M \\ x_2y_1 & x_2y_2 & \cdots & x_2y_M \\ \vdots & \vdots & \ddots & \dots \\ x_Ny_1 & x_Ny_2 & \cdots & x_Ny_M \end{pmatrix}$$

N X 1 1 X M N X M

Slide credit: Mark Goldman and Emily Mackevicius

Norm

- **Norm** is a function that assigns a strictly positive *length* or *size* to each vector in a vector space—except for the zero vector

- **L^1 norm -** *Manhattan/Taxicab Distance*, the *Mean Absolute Error (MAE)*, or the *Least Absolute Shrinkage and Selection Operator* (*LASSO)*

$$\|\boldsymbol{x}\|_1 := \sum_{i=1}^{n} |x_i|$$

- **L^2 norm -** *Euclidean Distance*, the *Mean Squared Error (MSE) / Least Squares Error*, or the *Ridge Operator*

$$\|\boldsymbol{x}\| := \sqrt{x_1^2 + \cdots + x_n^2}$$

Norm

L^p norm - Let p ≥ 1 be a real number. The p norm of vector x=(x_1, x_2,…x_n)

$$\|\mathbf{x}\|_p := \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$$

Matrix norm

$$\|A\| = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2}.$$

Orthogonal and Orthonormal Vectors

- A.B=0 then A and B are orthogonal
- A collection of vectors $a_1, . . . , a_k$ is orthogonal if $a_i \perp a_j$ for with $i \neq j$
- A collection of vectors $a_1, . . . , a_k$ is orthonormal if the set of vectors are mutually orthogonal and if every vector has magnitude 1 ($||a_i|| = 1$)

- normalized or unit vector: A vector of norm one
- normalizing : dividing a vector by its norm

$$\hat{v} = \frac{1}{||v||} v = \frac{v}{||v||}$$

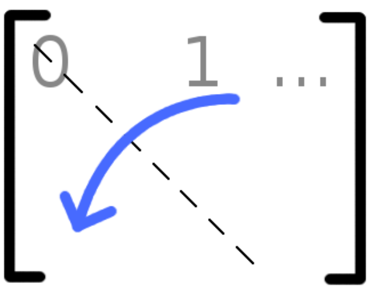
Euclidean distance

u and v are n dim vectors

$$d(u, v) = \|u - v\| = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_n - b_n)^2}$$

Matrix Operations

- Transpose – flip matrix, so row 1 becomes column 1



$$\begin{bmatrix} 0 & 1 \\ 2 & 3 \\ 4 & 5 \end{bmatrix}^T = \begin{bmatrix} 0 & 2 & 4 \\ 1 & 3 & 5 \end{bmatrix}$$

- A useful identity:

$$(ABC)^T = C^T B^T A^T$$

Matrix times a vector

$$\overrightarrow{y} = \overleftrightarrow{W}\,\overrightarrow{x}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{pmatrix} = \begin{pmatrix} W_{11} & W_{12} & \cdots & W_{1N} \\ W_{21} & W_{22} & \cdots & W_{2N} \\ \vdots & \vdots & & \vdots \\ W_{M1} & W_{M2} & \cdots & W_{MN} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix}$$

M X 1 M X N N X 1

Matrix times a vector: inner product interpretation

$$
\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_M \end{pmatrix} = \begin{pmatrix} W_{11} & W_{12} & \cdots & W_{1N} \\ W_{21} & W_{22} & \cdots & W_{2N} \\ \vdots & \vdots & & \vdots \\ W_{i1} & W_{i2} & \cdots & W_{iN} \\ \vdots & \vdots & \ddots & \vdots \\ W_{M1} & W_{M2} & \cdots & W_{MN} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix}
$$

$$
y_i = \sum_{j=1}^{N} W_{ij} x_j
$$

- Rule: the i^{th} element of **y** is the dot product of the i^{th} row of W with **x**
- matrix-vector product is really a dot product in disguise.

Product of 2 Matrices

- **Note:** Matrix multiplication doesn't (generally) commute, **AB ≠ BA**

$$
\underbrace{\begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1P} \\ A_{21} & A_{22} & \cdots & A_{2P} \\ \vdots & \vdots & & \vdots \\ A_{N1} & A_{N2} & \cdots & A_{NP} \end{pmatrix}}_{N \times P}
\underbrace{\begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1M} \\ B_{21} & B_{22} & \cdots & B_{2M} \\ \vdots & \vdots & & \vdots \\ B_{P1} & B_{P2} & \cdots & B_{PM} \end{pmatrix}}_{P \times M}
=
\underbrace{\begin{pmatrix} C_{11} & C_{12} & \cdots & C_{1M} \\ C_{21} & C_{22} & \cdots & C_{2M} \\ \vdots & \vdots & & \vdots \\ C_{N1} & C_{N2} & \cdots & C_{NM} \end{pmatrix}}_{N \times M}
$$

Matrix times Matrix: Matrix Multiplication by inner products

$$\begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1P} \\ A_{21} & A_{22} & \cdots & A_{2P} \\ \vdots & \vdots & & \vdots \\ A_{i1} & A_{i2} & \cdots & A_{iP} \\ \vdots & \vdots & & \vdots \\ A_{N1} & A_{N2} & \cdots & A_{NP} \end{pmatrix} \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1j} & \cdots & B_{1M} \\ B_{21} & B_{22} & \cdots & B_{2j} & \cdots & B_{2M} \\ \vdots & \vdots & & \vdots & & \vdots \\ B_{P1} & B_{P2} & \cdots & B_{Pj} & \cdots & B_{PM} \end{pmatrix} = \begin{pmatrix} C_{11} & C_{12} & \cdots & C_{1M} \\ C_{21} & C_{22} & \cdots & C_{2M} \\ \vdots & \vdots & C_{ij} & \vdots \\ C_{N1} & C_{N2} & \cdots & C_{NM} \end{pmatrix}$$

$$C_{ij} = \sum_{k=1}^{P} A_{ik} B_{kj}$$

- C_{ij} is the inner product of the i^{th} row of **A** with the j^{th} column of **B**
- *matrix product* **C** = **AB** (denoted without multiplication signs or dots)

Special matrices: diagonal matrix

This acts like scalar multiplication

$$\overleftrightarrow{D} = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{pmatrix}$$

$$\overleftrightarrow{D}\,\overrightarrow{x} = \begin{pmatrix} d_1 x_1 \\ d_2 x_2 \\ \vdots \\ d_n x_n \end{pmatrix}$$

Special matrices: identity matrix

$$\overleftrightarrow{1} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

$$\text{for all } \overleftrightarrow{A},\ \overleftrightarrow{1}\overleftrightarrow{A} = \overleftrightarrow{A}\overleftrightarrow{1} = \overleftrightarrow{A}$$

A square matrix A is symmetric if $A = A^T$, i.e., $A_{ij} = A_{ji}$ for all i; j.

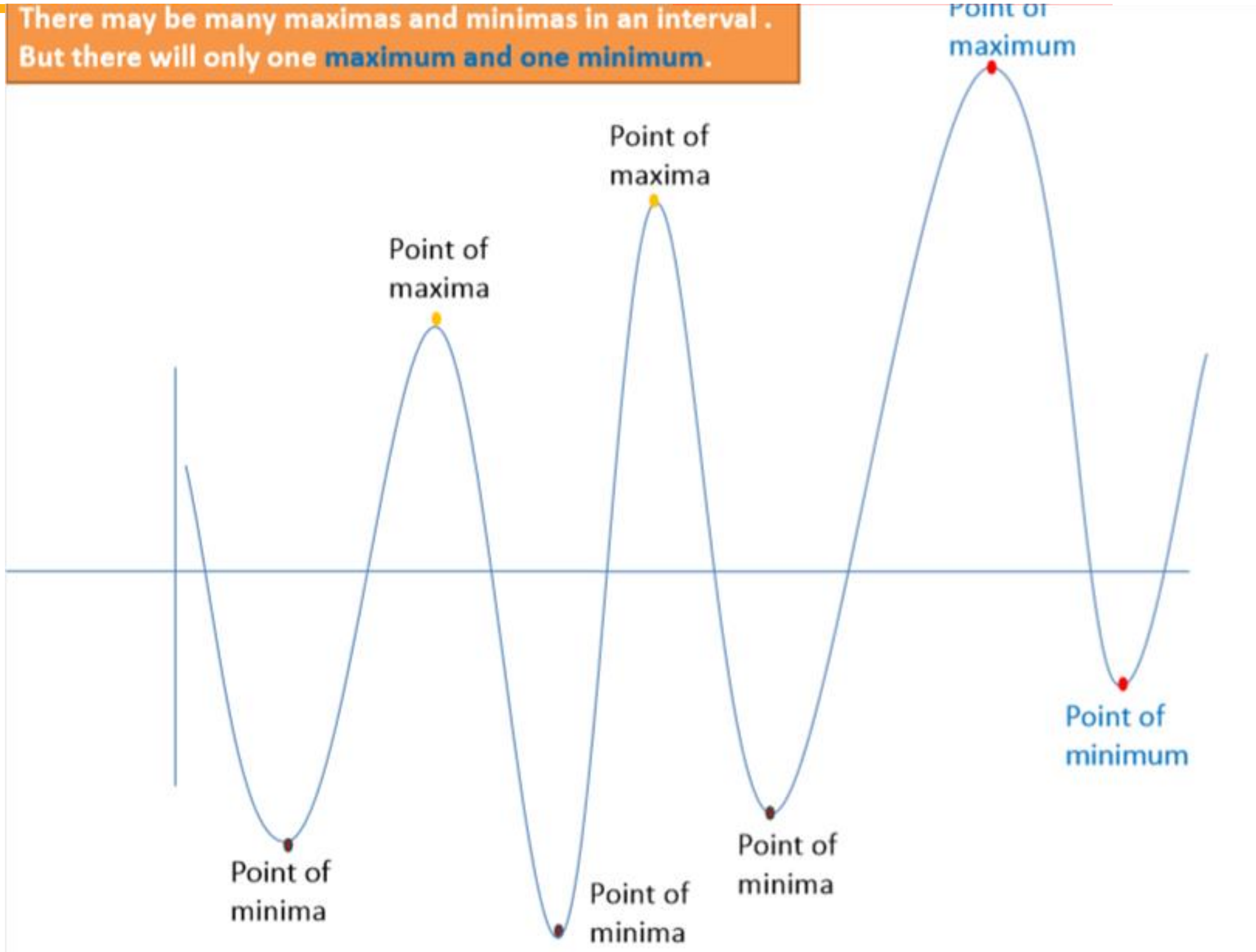
Different types of product

- $\mathbf{x}, \mathbf{y}$ = column vectors (nx1)
- $\mathbf{X}, \mathbf{Y}$ = matrices (mxn)
- x, y = scalars (1x1)

- $\mathbf{x}^T \mathbf{y} = \mathbf{x} \cdot \mathbf{y}$ = inner product (1xn x nx1 = scalar)
- $\mathbf{x} \otimes \mathbf{y} = \mathbf{x}\, \mathbf{y}^T$ = outer product (nx1 x 1xn = matrix)

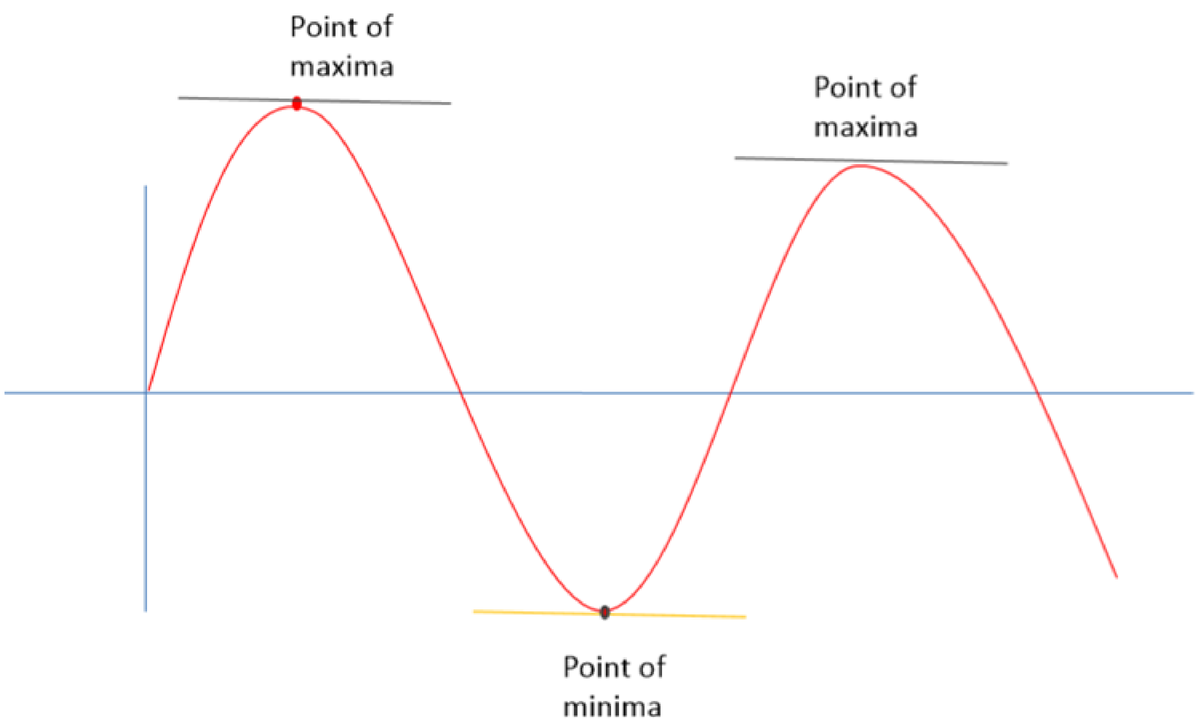
- $\mathbf{X} * \mathbf{Y}$ = matrix product
- $\mathbf{X} .* \mathbf{Y}$ = element-wise product

Maxima and Minima of a function

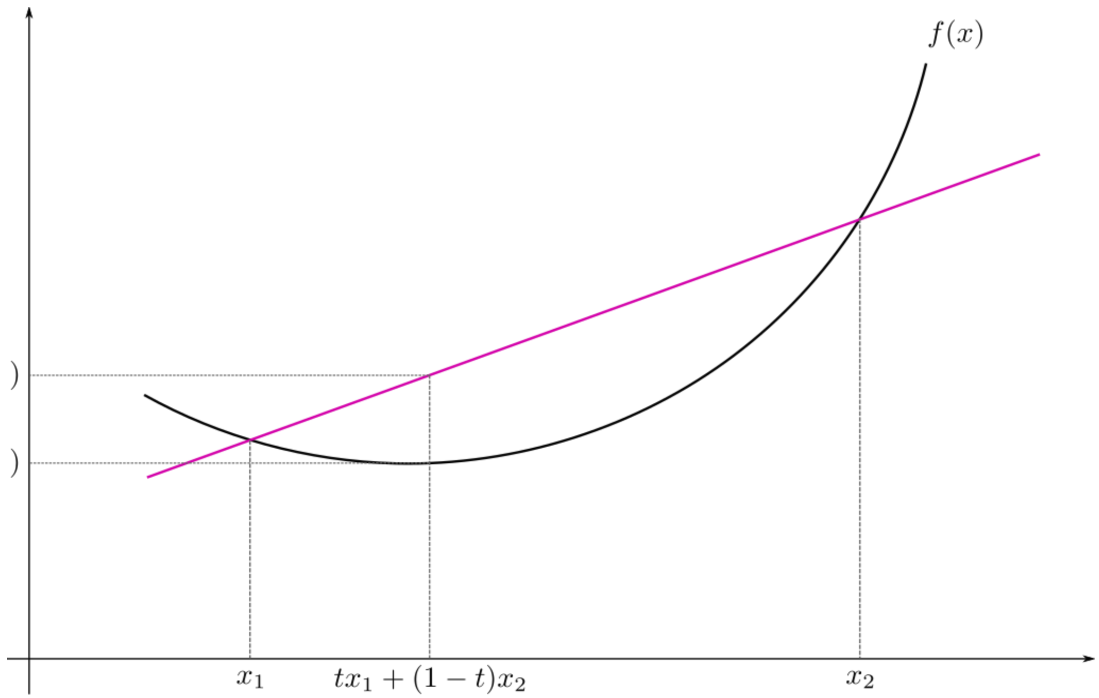


Maxima and Minima

- For maxima and minima $m=dy/dx=\tan 0^0=0$
- $dy/dx = 0$ means tangent is parralel to X –axis.

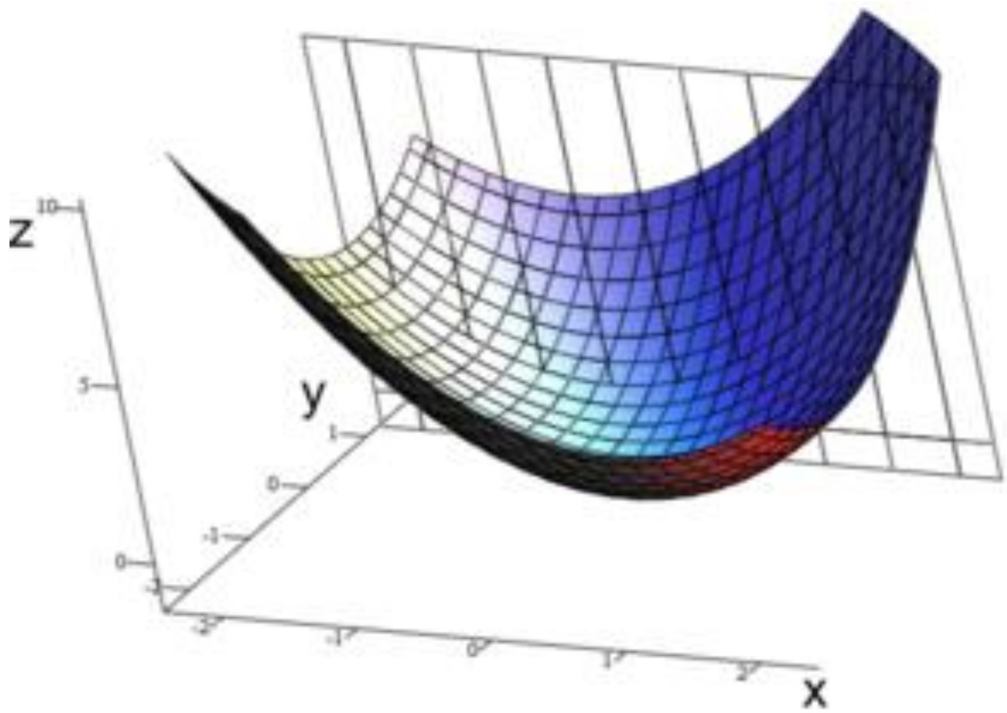


Convex Function



Real-valued function defined on an *n*-dimensional interval is called **convex** if the line segment between any two points on the graph of the function lies above or on the graph

Convex Function : Multivariate



Calculus review

Differentiation

- The derivative provides us information about the rate of change of a function.

- The derivative of a function is also a function.

  Example:

  - The derivative of the acceleration function is the velocity function.

Derivative = rate of change

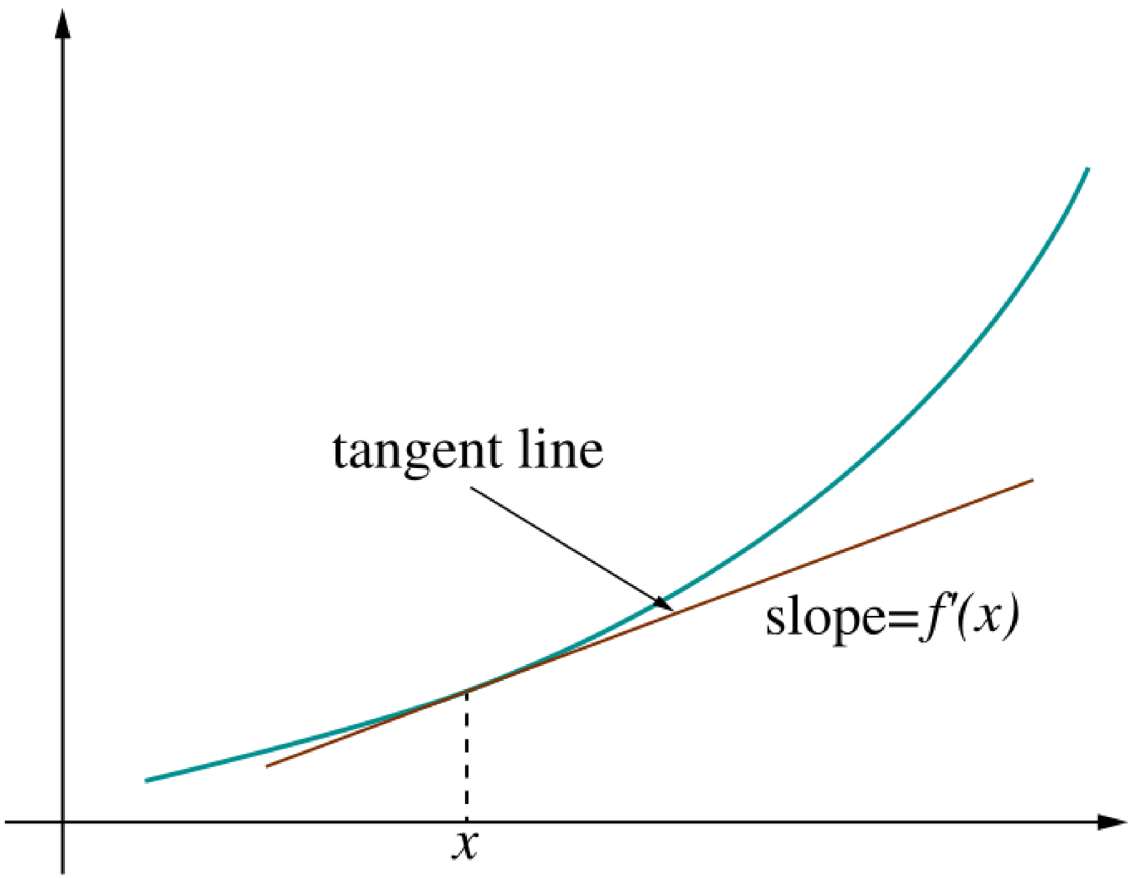


Image: Wikipedia

Derivative = rate of change

- Linear function y = mx + b
- Slope

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

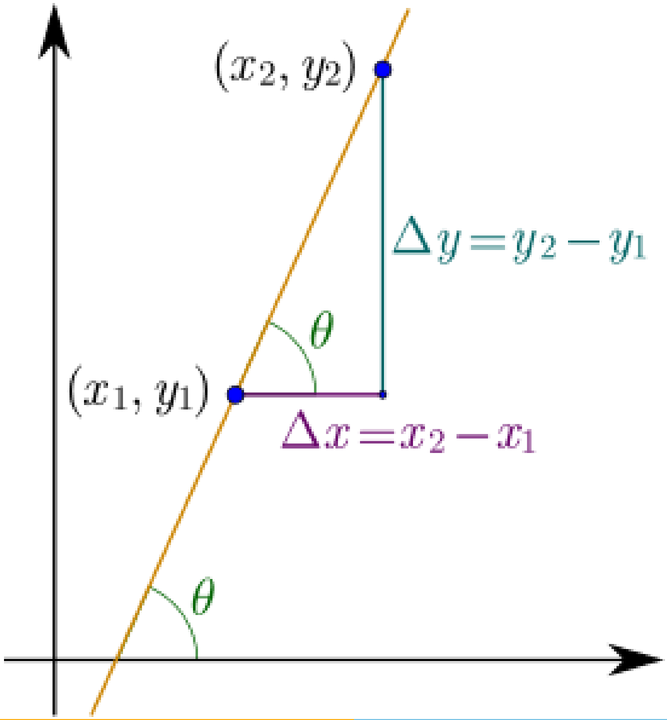


Image: Wikipedia

Ways to Write the Derivative

Given the function f(x), we can write its derivative in the following ways:

- $f'(x)$
- $\frac{d}{dx}f(x)$

The derivative of x is commonly written dx.

Differentiation Formulas

The following are common differentiation formulas:

- The derivative of a constant is 0.

$$\frac{d}{du}c = 0$$

- The derivative of a sum is the sum of the derivatives.

$$\frac{d}{du}(f(u) + g(u)) = f'(u) + g'(u)$$

More Formulas

- The derivative of *u* to a constant power:

$$\frac{d}{du} u^n = n * u^{n-1}$$

- The derivative of *e*:

$$\frac{d}{du} e^u = e^u$$

- The derivative of *log*:

$$\frac{d}{du} \log(u) = \frac{1}{u}$$

Product and Quotient

The product rule and quotient rules are commonly used in differentiation.

- Product rule:

$$\frac{d}{du}(f(u)*g(u)) = f(u)g'(u) + g(u)f'(u)$$

- Quotient rule:

$$\frac{d}{du}\left(\frac{f(u)}{g(u)}\right) = \frac{g(u)f'(u) - f(u)g'(u)}{g^{2}(u)}$$

Chain Rule

The chain rule allows you to combine any of the differentiation rules we have already covered.

- First, do the derivative of the outside and then do the derivative of the inside.

$$\frac{d}{du} f(g(u)) = f'(g(u)) * g'(u) * du$$

Try These

$$f(z) = z + 11$$

$$s(y) = 4ye^{2y}$$

$$g(y) = 4y^3 + 2y$$

$$p(x) = \frac{\log(x^2)}{x}$$

$$h(x) = e^{3x}$$

Solutions

$$f'(z) = 1$$

$$s'(y) = 8ye^{2y} + 4e^{2y}$$

$$g'(y) = 12y^2 + 2$$

$$p'(x) = \frac{2 - \log(x^2)}{x^2}$$

$$h'(x) = 3e^{3x}$$

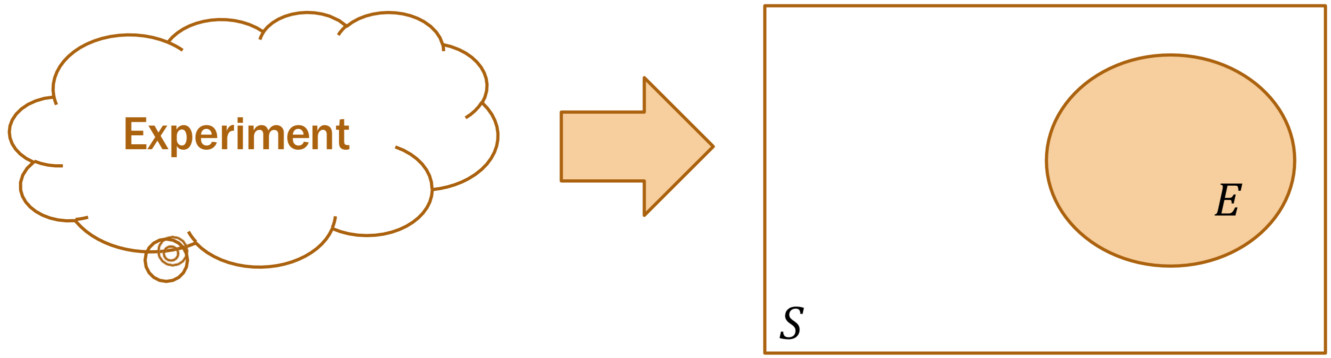
Probability Review

Two main "schools of thought"

- **Bayesian probability= degree of belief**
  - Probabilities are assigned to events based on evidence and personal belief and are centered around Bayes' theorem
  - p(heads=1)=0.5 means you think the event that a particular coin will land heads is 50% likely.
- **Frequentist probability= long run frequencies**
  - Events are observed and counted, and their frequencies provide the basis for directly calculating a probability
  - p(heads=1)=0.5 means that the empirical fraction of times this event will occur across infinitely repeated trials

Probability – Meaning & Concepts

An experiment in probability:



Sample Space, S: The set of all possible outcomes of an experiment

Event, E: Some subset of S ($E \subseteq S$).

Probability – Meaning & Concepts

Sample Space, S

- **Coin flip**
  $S = \{\text{Heads}, \text{Tails}\}$
- **Roll of 6-sided die**
  $S = \{1, 2, 3, 4, 5, 6\}$
- **# emails in a day**
  $S = \{x \mid x \in \mathbb{Z}, \; x \geq 0\}$

Event, E

- **Flip lands heads**
  $E = \{\text{Heads}\}$
- **Roll is 3 or less:**
  $E = \{1, 2, 3\}$
- **Low email day (≤ 20 emails)**
  $E = \{x \mid x \in \mathbb{Z}, \; 0 \leq x \leq 20\}$

$$P(E) = \lim_{n \to \infty} \frac{n(E)}{n}$$

n = # of total trials

$n(E)$ = # trials where E occurs

Slide Credit : Lisa Yan

3 Axioms of Probability

- Axiom 1: $0 \leq P(E) \leq 1$

- Axiom 2: $P(S) = 1$

- Axiom 3: If E and F mutually exclusive ($E \cap F = \emptyset$),

then $P(E) + P(F) = P(E \cup F)$

Conditional Probability

- The probability of *E* given that (aka conditioned on) event *F* already happened:

$$P(E \mid F) = \frac{P(EF)}{P(F)} = \frac{P(E \cap F)}{P(F)}$$

- **It allows us to update our beliefs in the face of new evidence – critical in machine learning**

- **The Chain Rule:**

$$P(EF) = P(E \mid F)\, P(F)$$

general form:

$$P(E_1 E_2 \ldots E_n) = P(E_1)P(E_2 \mid E_1) \ldots P(E_n \mid E_1 E_2 \ldots E_{n-1})$$

The Law of Total Probability

For events E and F

$$P(F) = P(F \mid E)P(E) + P(F \mid E^C)P(E^C)$$

Let E1,E2,……., En are mutually exclusive and exhaustive i.e every outcome in sample space falls into exactly one of those events then giving a general form:

$$P(F) = \sum_{i=1}^{n} P(F \mid E_i)P(E_i)$$

Example

P(Cancer) =1/100

P(+/Cancer)=90/100

P(+/Not Cancer)=8/100

P(Cancer/+)=?

Bayes Rule

$P\,(E = \textbf{Evidence} \mid F = \textbf{Fact})$

(collected from data)

Bayes

$P\,(F = \textbf{Fact} \mid E = \textbf{Evidence})$

(categorize a new datapoint)

Bayes Rule
Common form

$$P(F|E) = \frac{P(E|F)P(F)}{P(E)}$$

Bayes Rule Expanded form

$$P(E \mid F) = \frac{P(F \mid E)P(E)}{P(F \mid E)P(E) + P(F \mid E^C)P(E^C)} = \frac{P(F \mid E)P(E)}{\sum_i P(F \mid E_i)P(E_i)}$$

Bayes' Theorem terminology

posterior likelihood prior

$$\boxed{P(F|E) = \frac{P(E|F)P(F)}{P(E)}}$$

normalization constant

Given new evidence E, update belief of fact F

Prior belief → Posterior belief

$P(F)$ → $P(F|E)$

Example

- 60% of all email received is spam.
- 20% of spam has the word “Dear”
- 1% of non-spam has the word “Dear”

You get an email with the word “Dear” in it. What is the probability that the email is spam?

Independence

Two events E and F are called **independent** if:

$P(EF) = P(E)\,P(F)$ equivalently: $P(E \mid F) = P(E)$

Otherwise, they are called **dependent** events

Three events E, F, and G independent if:

$P(EFG) = P(E)\,P(F)\,P(G)$, and

$P(EF) = P(E)\,P(F)$, and

$P(EG) = P(E)\,P(G)$, and

$P(FG) = P(F)\,P(G)$

Probability of events

<u>P(AB)</u>

Generally: P(A)P(B|A)

Independent: P(A)P(B)

<u>P(A∪B)</u>

Generally: P(A) + P(B) – P(AB)

Mutually Exclusive: P(A) + P(B)

Independent Events-Example

Example:

The probability that you will get an A grade in Quantitative Methods is 0.7. The probability that you will get an A grade in Marketing is 0.5. Assuming these two courses are independent, compute the probability that you will get an A grade in both these subjects.

Solution:

Let A = getting A grade in Quantitative Methods

Let B =getting A grade in Marketing

It is given that A and B are independent.

$$P(A \cap B) = P(A).P(B) = 0.7.0.5 = 0.35.$$

Random Variable

CS variables

type name value

```
int a = 5;
```

Random variables

A is the number of Pokemon we bring to our *future* battle.

$$A \in \ 1, 2, \dots, 6$$

Random variables are like typed variables (with uncertainty)

A random variable is a real-valued function defined on a sample space.

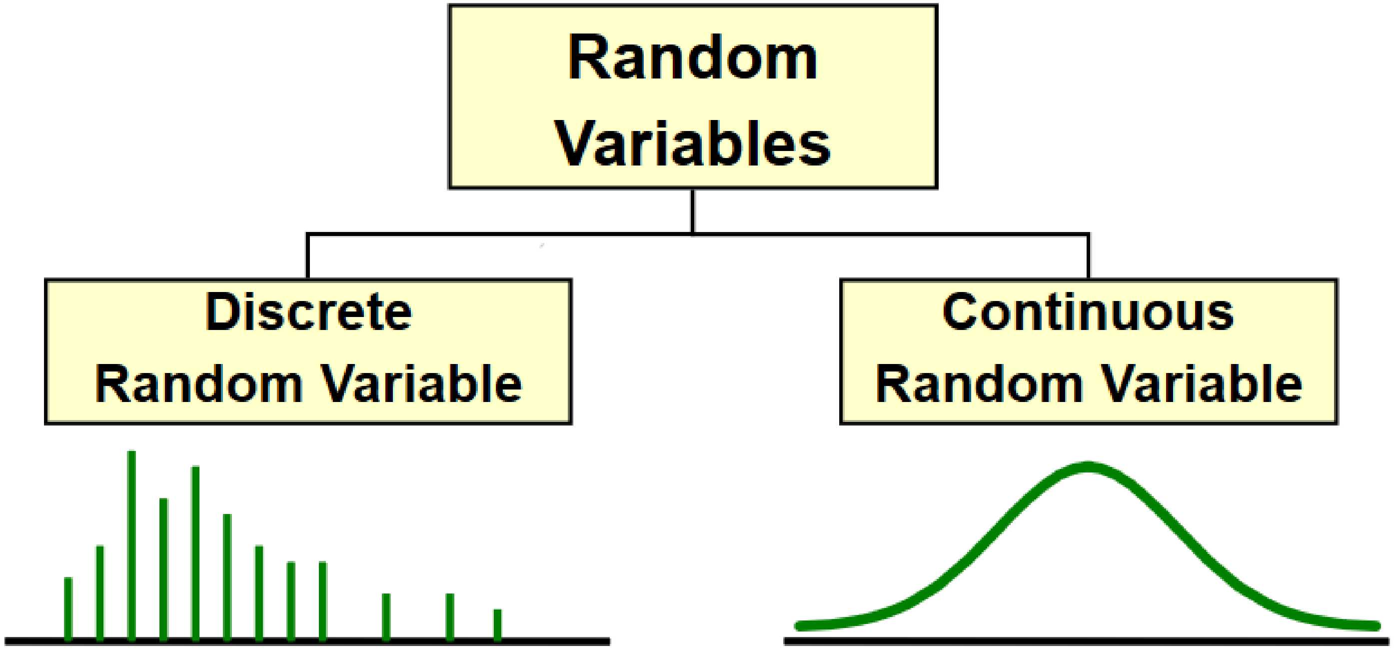
Experiment

Outcome

$0 = k$

Random Variable

Represents a possible numerical value from a random event



Example

- we flip three fair coins. A random variable Y is the total number of "heads" on the three coins
  - $P(Y = 0) = 1/8$ (T, T, T)
  - $P(Y = 1) = 3/8$ (H, T, T), (T, H, T), (T, T, H)
  - $P(Y = 2) = 3/8$ (H, H, T), (H, T, H), (T, H, H)
  - $P(Y = 3) = 1/8$ (H, H, H)
  - $P(Y = 4) = 0$
- **It is confusing that random variables and events use the same notation.**
  - **Random variables ≠ events.**
  - **We can define an event to be a particular assignment of a random variable.**

Probability distributions

It is the mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment.

PMF and CMF

- The **probability mass function** (PMF) of a random variable is a function that maps possible outcomes of a random variable to the corresponding probabilities.
- **It can be written as :** $P(\mathrm{X} = x) = p(x) = p_{\mathrm{X}}(x)$

- **the cumulative distribution function (CDF) is defined as**

  $F(a) = F_{\mathrm{x}}(a) = P(\mathrm{X} \leq a)$ **where** $-\infty < a < \infty$

- **For a discrete RV** X **the CDF is**

$$F(a) = P(X \leq a) = \sum_{\text{all } x \leq a} p(x)$$

Expectation

- Average value of the random variable over many repetitions of the experiment it represents

- The **expectation** of a discrete random variable *X* is defined as

$$E[X] = \sum_{x:P(x)>0} xP(x)$$

- It goes by many other names: *mean*, *expected value*, *weighted average*, *center of mass*, *1st moment*.

Important properties of expectation

1. **Linearity**:

$$E[aX + b] = aE[X] + b$$

2. **Expectation of a sum = sum of expectation:**

$$E[X + Y] = E[X] + E[Y]$$

3. **expected value of a function**

$$E[g(X)] = \sum_{x} g(x) \cdot px(x)$$

Variance and Standard deviation

- The variance of a discrete random variable X with expected value is: $$\mathrm{Var}(X) = E[(X-\mu)^2]$$
- different form of the same equation:

$$\mathrm{Var}(X) = E[X^2] - E[X]^2$$

- some useful identities for variance

$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$$

- Standard deviation:

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}.$$

Bernoulli Random Variable

- Consider an experiment with two outcomes: “success” and “failure.”
- A Bernoulli random variable X maps “success” to 1 and “failure” to 0.
- Other names: indicator random variable, boolean random variable

$$X \sim \text{Ber}(p)$$

PMF	$P(X=1) = p(1) = p$ $P(X=0) = p(0) = 1-p$
Expectation	$E[X] = p$
Variance	$\text{Var}\ (X) = p(1-p)$

Examples:

- Coin flip
- Random binary digit
- Whether a disk drive crashed

Binomial Random Variable

- Consider an experiment: n independent trials of Ber(p) random variables.
- A Binomial random variable X is the number of successes in n trials.

$X \sim \text{Bin}(n, p)$

PMF	$k = 0, 1, \ldots, n$: $P(X = k) = p(k) = \binom{n}{k} p^k (1-p)^{n-k}$
Expectation	$E[X] = np$
Variance	$\text{Var}(X) = np(1-p)$

Examples:

- # heads in n coin flips
- # of 1's in randomly generated length n bit string
- # of disk drives crashed in 1000 computer cluster (assuming disks crash independently)

Binomial distribution

two possible outcomes in n independent trials, then the probability of exactly X "successes"=

n = number of trials

$$\boxed{\binom{n}{X} p^X (1-p)^{n-X}}$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

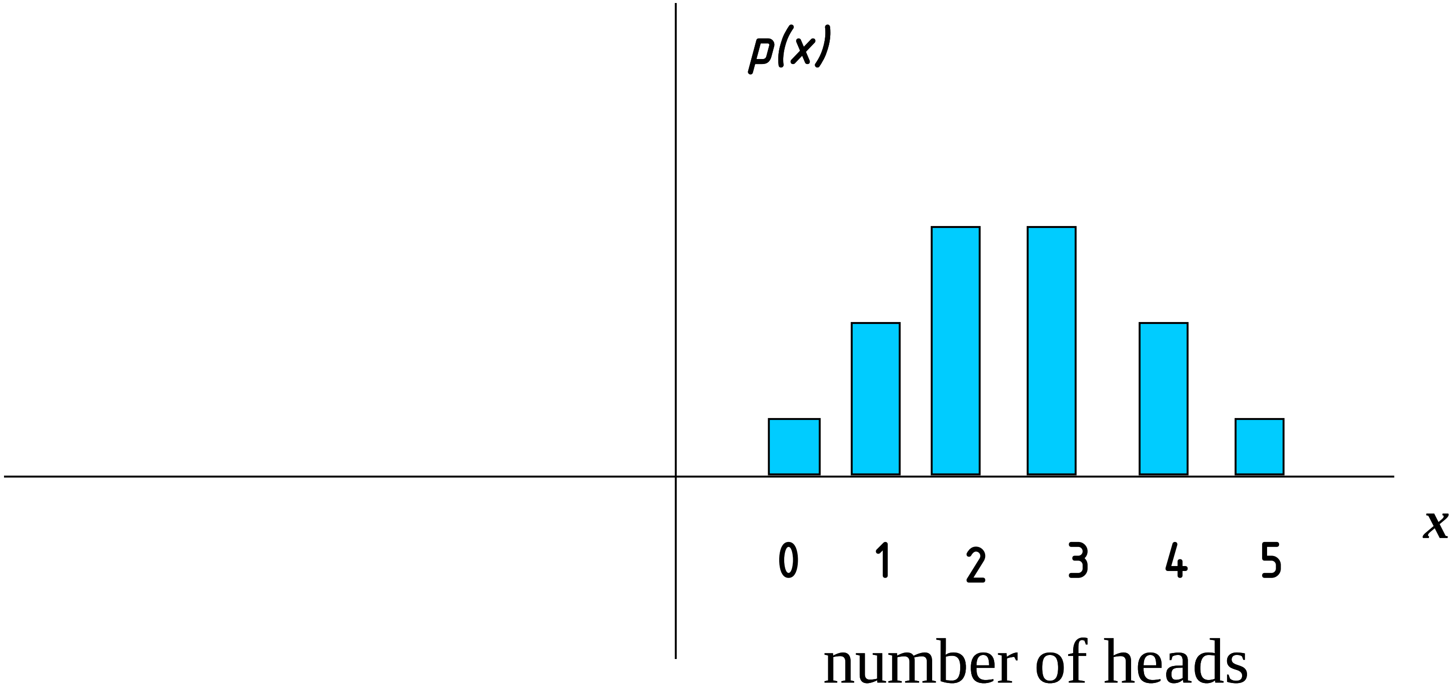
1-p = probability of failure

X = # successes out of n trials

p = probability of success

Binomial distribution

X= the number of heads tossed in 5 coin tosses



Bernoulli Vs Binomial

1. The random variable

$$X \sim \text{Ber}(p)$$

2. is distributed as a

3. Bernoulli

4. with parameter

Example: Heads in one coin flip, P(heads) = 0.8 = p

$$Y \sim \text{Bin}(n, p)$$

Example: # heads in 40 coin flips, P(heads) = 0.8 = p

Poisson Random Variable

- Consider an experiment that lasts a fixed interval of time.
- A Poisson random variable X is the number of successes over the experiment duration.
- A Poisson random variable approximates Binomial where n is large, p is small, and λ = np is “moderate”.
- Examples:

\# earthquakes per year

\# server hits per second

\# of emails per day

$$P(Y = i) = \frac{\lambda^i}{i!} e^{-\lambda}$$

$$E[Y] = \lambda$$

$$Var(Y) = \lambda$$

Continuous probability distribution

Continuous random variables

- *X* is a **Continuous Random Variable** if there is function
  f(*x*) ≥ 0 for -∞ ≤ *x* ≤ ∞, such that:

$$P(a \le X \le b) = \int_a^b f(x)dx$$

- *f* is a Probability Density Function (PDF) if:

$$P(-\infty < X < \infty) = \int_{-\infty}^{\infty} f(x)dx = 1$$

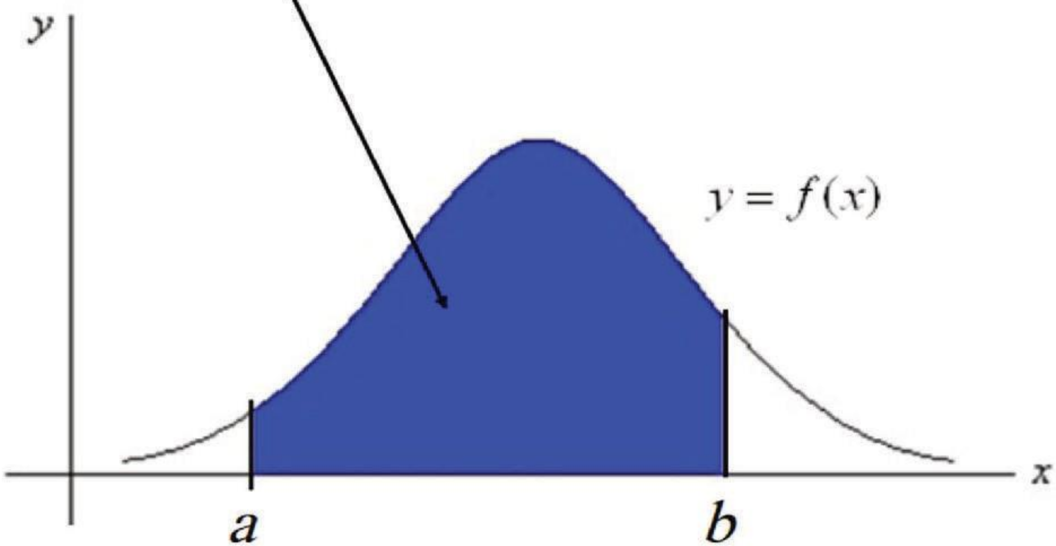
- *f*(*x*) is **not** a probability, it is probability/units of X, not meaningful without some subinterval over X

$$P(X = a) = \int_a^a f(x)dx = 0$$

Probability Densities

Probability Density Function

$P(a \leq X \leq b)$ is given by the area of the shaded region.



BITS Pilani

CDF for Continuous random variables

For a continuous random variable *X*, the **cumulative distribution function F(a)** is:

$$F_X(a) = P(X \le a) = \int_{-\infty}^{a} f(x)dx$$

Probability Query	**Solution**	**Explanation**
$P(X \le a)$	$F(a)$	This is the definition of the CDF
$P(X < a)$	$F(a)$	Note that $P(X = a) = 0$
$P(X > a)$	$1 - F(a)$	$P(X \le a) + P(X > a) = 1$
$P(a < X < b)$	$F(b) - F(a)$	$F(a) + P(a < X < b) = F(b)$

Expectation and Variance

For <u>discrete</u> RV X:

$$E[X] = \sum_x x\, p(x)$$

$$E[g(X)] = \sum_x g(x)\, p(x)$$

$$E[X^n] = \sum_x x^n\, p(x)$$

For <u>continuous</u> RV X:

$$E[X] = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)\, f(x)\, dx$$

$$E[X^n] = \int_{-\infty}^{\infty} x^n\, f(x)\, dx$$

For both continuous and discrete RVs:

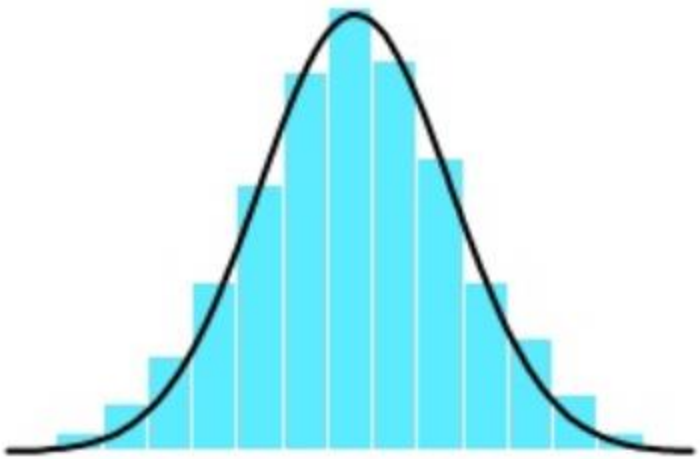
$$E[aX + b] = aE[X] + b$$

$$\mathrm{Var}(X) = E[(X-\mu)^2] = E[X^2] - (E[X])^2 \qquad (\text{with } \mu = E[X])$$

$$\mathrm{Var}(aX + b) = a^2\, \mathrm{Var}(X)$$

Gaussian Distribution

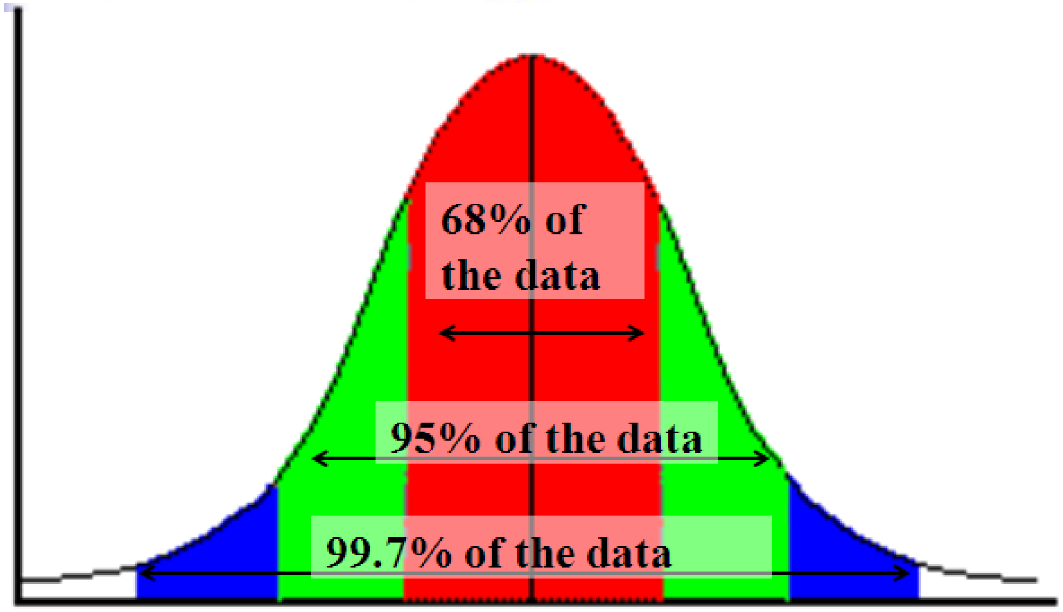
- The commonest and the most useful continuous distribution.
- A symmetrical probability distribution where most results are located in the middle and few are spread on both sides.
- It has the shape of a bell.
- Can entirely be described by its mean and standard deviation.



Gaussian Distribution

Empirical Rule:

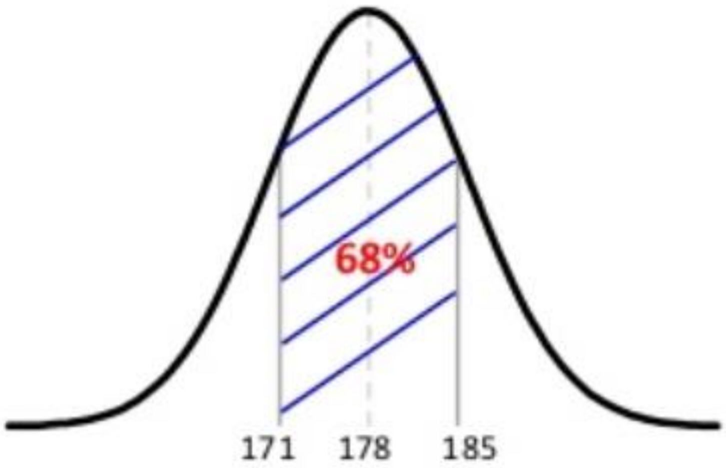
- For any normally distributed data:
  - **68%** of the data fall within **1** standard deviation of the mean.
  - **95%** of the data fall within **2** standard deviations of the mean.
  - **99.7%** of the data fall within **3** standard deviations of the mean.



Gaussian Distribution

- Suppose that the heights of a sample men are normally distributed.
- The mean height is **178** cm and a standard deviation is **7** cm.

- **We can generalize that:**
  - 68% of population are between 171 cm and 185 cm.
  - This might be a generalization, but it's true if the data is normally distributed.



Gaussian Distribution

In one dimension

Causes pdf to decrease as distance from center increases

$$N(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

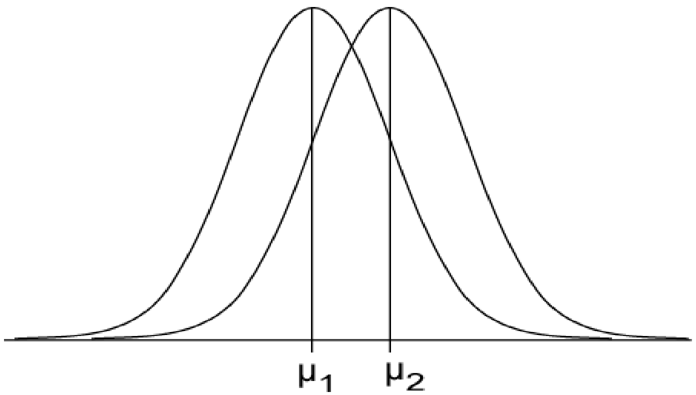
Controls width of curve

Normalizing constant: insures that distribution integrates to 1

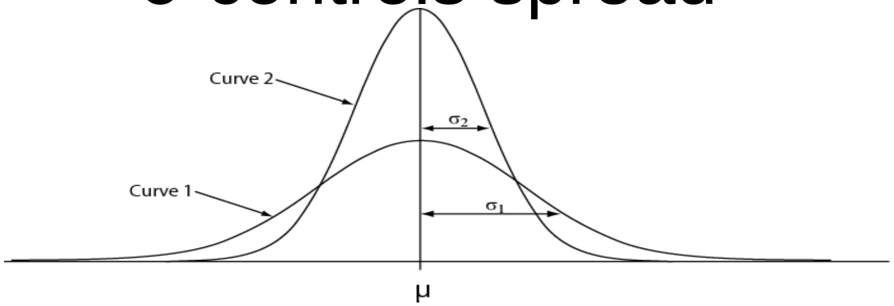
BITS Pilani

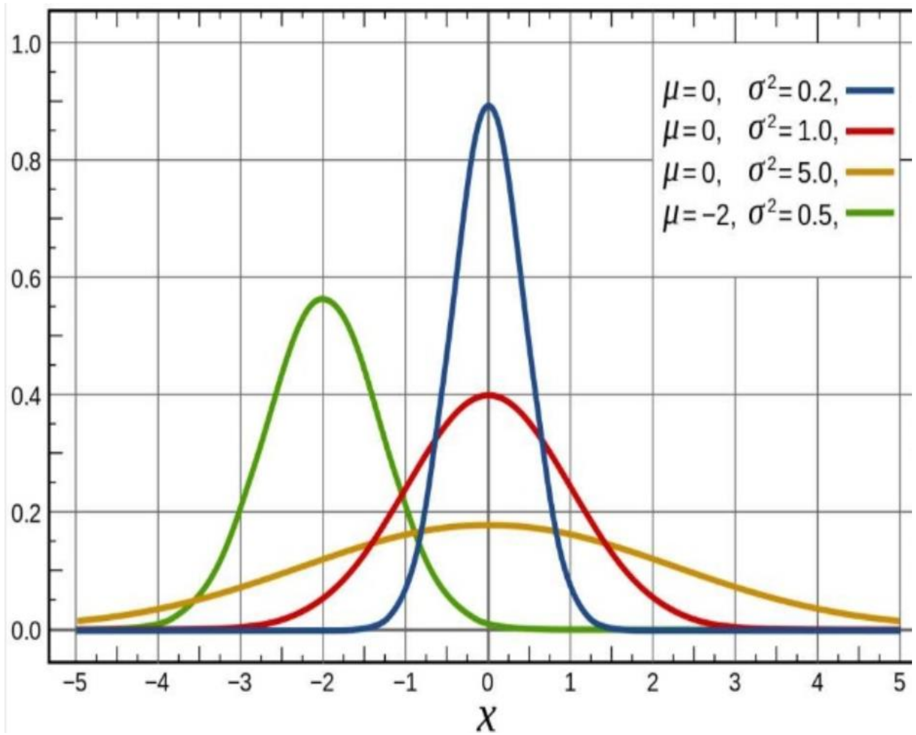
Gaussian Distribution

μ controls location



σ controls spread





Standard - Gaussian Distribution

$$P(a \leq X \leq b) = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx$$

- No closed form for the integral
- No closed form for F(x)

$\mathcal{N}(\mu, \sigma^2)$

A function that has been solved for numerically

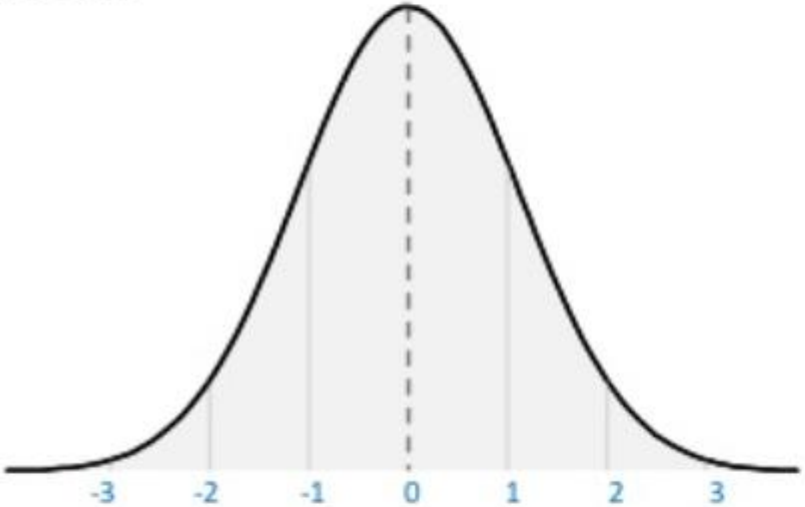
$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

The cumulative density function of any normal

Standard - Gaussian Distribution

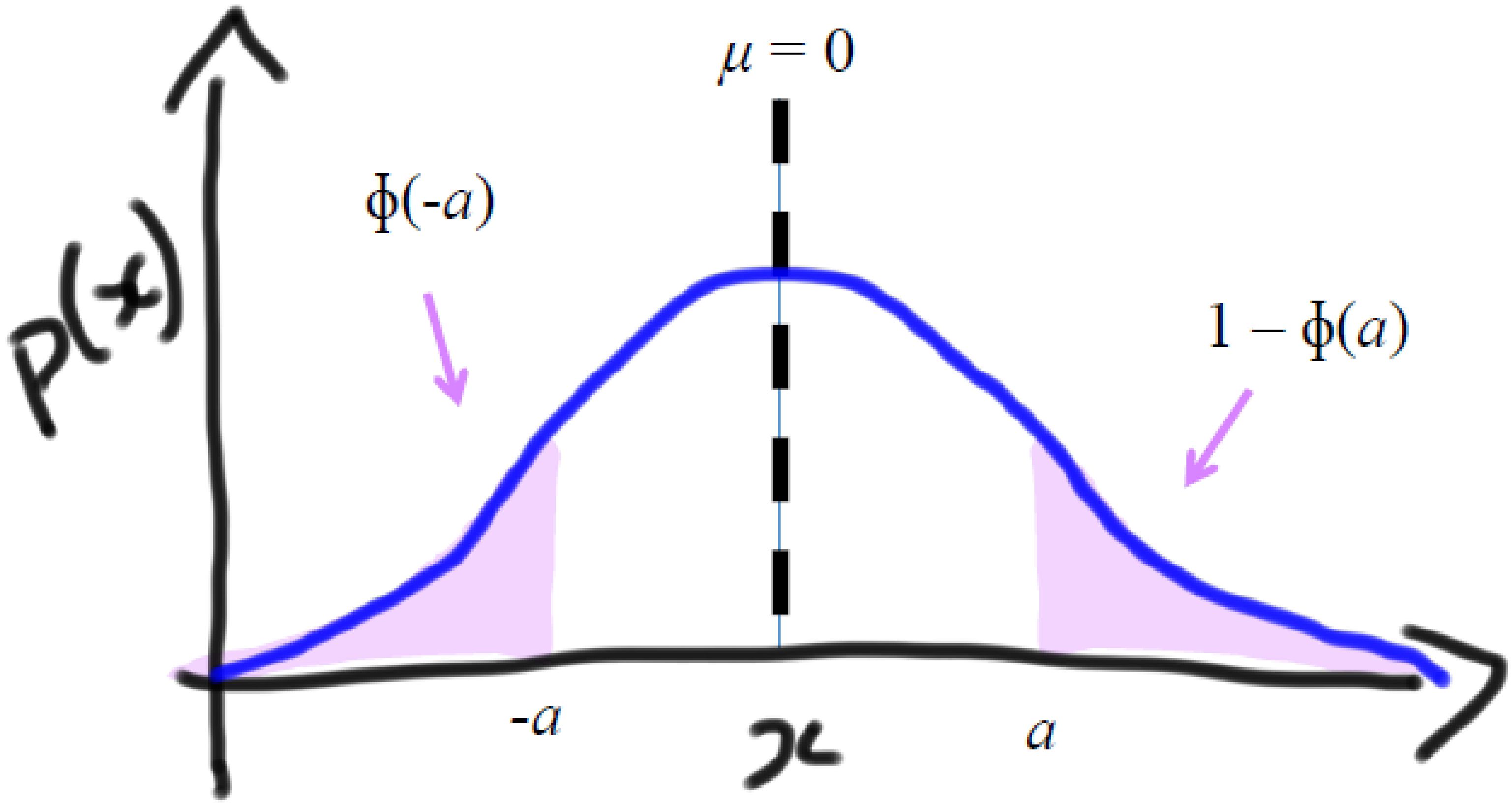
Standard Normal Distribution:

- A common practice to convert any normal distribution to the standardized form and then use the standard normal table to find probabilities.
- The **Standard Normal Distribution** (Z distribution) is a way of standardizing the normal distribution.
- It always has a mean of **0** and a standard deviation of **1**.



Standard - Gaussian Distribution

$$\phi(-a) = 1 - \phi(a)$$



Computing probabilities with Normal RVs

Let $X \sim \mathrm{N}\ (\mu, \sigma^2)$ What is $P(X \leq x) = F(x)$?

1. Rewrite in terms of standard normal CDF Φ by computing $z = \frac{x - \mu}{}$.
   Linear transforms of Normals are Normal:

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right) \qquad Z = \frac{(x - \mu)}{\sigma}, \text{ where } Z \sim \mathrm{N}(0,1)$$

2. Then, look up in a Standard Normal Table, where $z \geq 0$.
   Normal PDFs are symmetric about their mean:

$$\Phi\ (-z) = 1 - \Phi(z)$$

Standard Normal Table

Note: An entry in the table is the area under the curve to the left of

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.

Joint probability mass functions

Roll two 6-sided dice, yielding values X and Y.

X random variable	$P(X = 1)$ probability of an event	$P(X = k)$ probability mass function
X, Y random variables	$P(X = 1 \cap Y = 6)$ $P(X = 1, Y = 6)$ new notation: the comma	$P(X = a, Y = b)$

The marginal distributions of the joint PMF are defined as:

$$p_X(a) = P(X = a) = \sum_y p_{X,Y}(a, y)$$

Multinomial distribution

- The multinomial is a generalization of the binomial.
- It is used when there are more than 2 possible outcomes
- partitioning *n* trials into 3 or more outcomes (with probabilities: $p_{1,}\ p_{2,}\ p_{3,}..)$
  - General formula for 3 outcomes:

$$P(D = x, R = y, G = z) = \frac{n!}{x!\,y!\,z!}\, p_D^x\, p_R^y\, (1 - p_D - p_R)^z$$

Multinomial example

randomly choosing 8 people from an audience that contains 50% democrats, 30% republicans, and 20% green party, what's the probability of choosing exactly 4 democrats, 3 republicans, and 1 green party member?

$$\boxed{P(D=4, R=3, G=1) = \frac{8!}{4!\,3!\,1!}(.5)^4(.3)^3(.2)^1}$$

Decision Theory

Decision Theory

- Suppose **x** is an input vector together with a corresponding vector **t** of target variables
- Goal: predict **t** given a new value for **x.**
- The joint probability distribution ***p(x, t)*** provides a complete summary of the uncertainty associated with these variables.
- Determination of ***p(x, t)*** from a set of training data is called ***inference*** and is a difficult problem.

Decision Theory

Inference step

Determine either $p(t|\mathbf{x})$ or $p(\mathbf{x}, t)$.

Decision step

For given $\mathbf{x}$, determine optimal t.

Example : Medical diagnosis problem

Input: X-ray image of a patient
Input vector x is the set of pixel intensities in the image

Output: Presence of cancer = Class C1,
Absence of cancer, = Class C2.
Choose t to be a binary variable such that
t =0 corresponds to C1 and t =1 corresponds to C2.

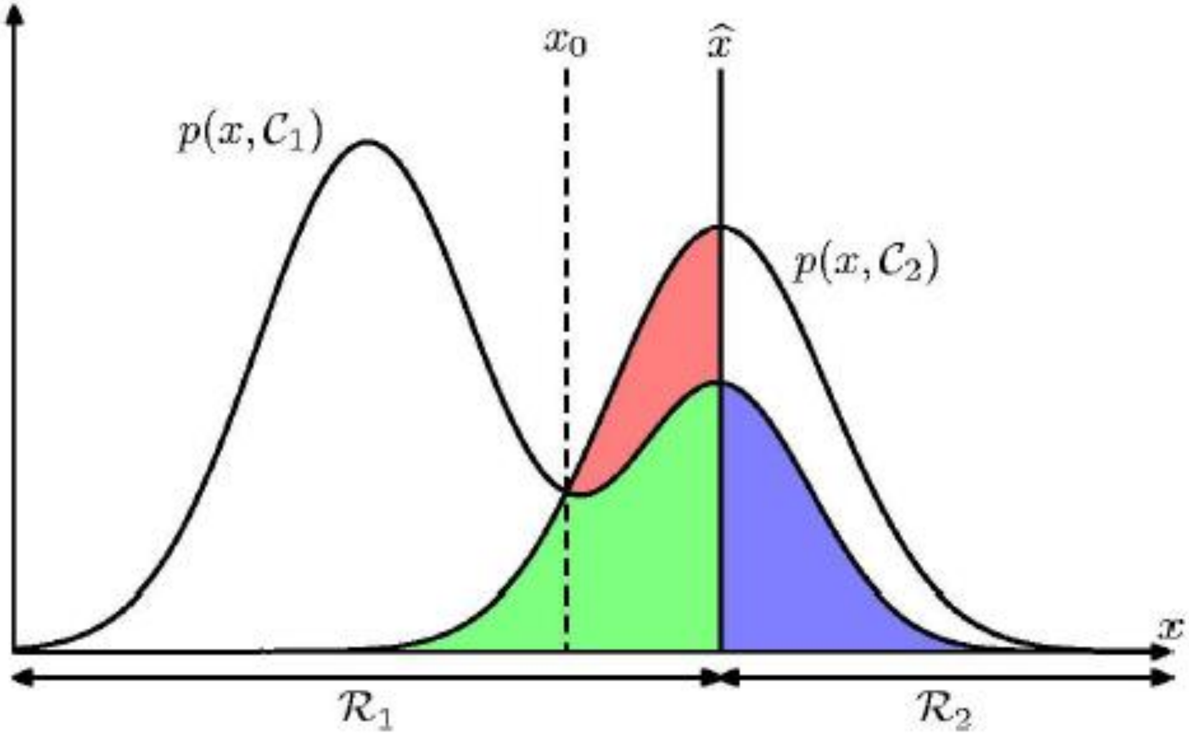
We are interested in the probabilities of the two classes given the image, which are given by $p(C_k|x)$.
Using Bayes' theorem,

$$p(C_k|\mathbf{x}) = \frac{p(\mathbf{x}|C_k)p(C_k)}{p(\mathbf{x})}$$

Minimum Misclassification Rate

- Divide the input space into regions Rk called decision regions, one for each class, such that all points in Rk are assigned to class Ck
- Boundaries between decision regions are called decision boundaries or decision surfaces
- A mistake occurs when an input vector belonging to class C1 is assigned to class C2 or vice versa.

Minimum Misclassification Rate



$$\begin{aligned} p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\,\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\,\mathrm{d}\mathbf{x}. \end{aligned}$$

Minimum Misclassification Rate

$$
\begin{aligned}
p(\text{correct}) &= \sum_{k=1}^{K} p(\mathbf{x} \in \mathcal{R}_k, \mathcal{C}_k) \\
&= \sum_{k=1}^{K} \int_{\mathcal{R}_k} p(\mathbf{x}, \mathcal{C}_k)\, \mathrm{d}\mathbf{x}
\end{aligned}
$$

References

https://web.stanford.edu/~boyd/vmls/vmls.pdf

https://ocw.mit.edu/resources/res-6-012-introduction-to-probability-spring-2018/index.htm

https://web.stanford.edu/class/cs109/

http://cs229.stanford.edu/section/cs229-linalg.pdf

https://www.alextsun.com/files/distributions.pdf

Thank you !