Foundation of ML

Week 2

Assessment - 1

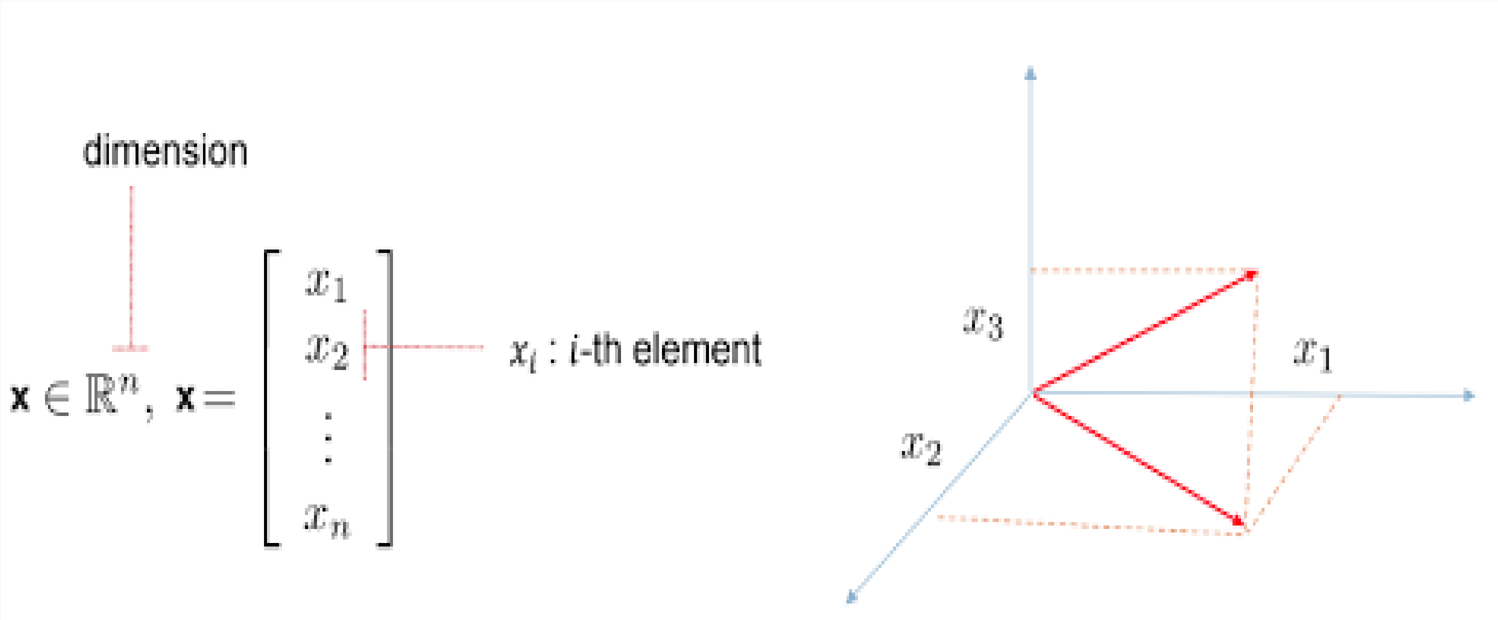
- You can access from:
  - “Content->Assessment->Assessment 1”
  - Menu – Assessment->Assignments->Assignment Task 1 Submission Folder”
- Open and use the “.ipynb” file to create your solution
- If you are confused – you can ask question
- You should do:
  - Think for the possible solution based on the given data and questions.
  - Propose the solution and the results.
  - Unleash your ideas rather than searching for specific solution.
  - Keep in mind there is not a single solution for a problem or there is not a single ML model that can solve all our problems.

Revising knowledge of Linear Algebra and Probability

- Linear Algebra
  - Vector & their operations
  - Matrix & their operations
  - Feature vectors and matrices
- Probability Concepts
  - Random experiment & Event
  - Joint probability
  - Conditional probability
  - Bayes Rules
- Random variable
  - Distribution of random variables

Vector

- In machine learning algorithms a data instance is represented by a vector, more precisely, by a feature vector



Vector operations

- Three main operations in vectors -
  - Transpose
  - Addition
  - Inner product
- Let's take two simple matrices x and y: $X = \begin{bmatrix} x_1 \\ x_2 \\ . \\ x_n \end{bmatrix}$ $Y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ y_n \end{bmatrix}$
- Transpose: $X^T = [x_1 \quad x_2 \quad .. \quad x_n]$

Vector operations...

- Addition:

$$X + Y = \begin{bmatrix} x_1 \\ x_2 \\ . \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ . \\ x_n + y_n \end{bmatrix}$$

- Inner product:

$$X^T Y = [x_1 y_1 + x_2 y_2 + .. + x_n y_n]$$

- Magnitude of length of a vector:

$$length(X) = \sqrt{x_1^2 + x_2^2 + .. + x_n^2}$$

- Above length known as 2-norm of vector:

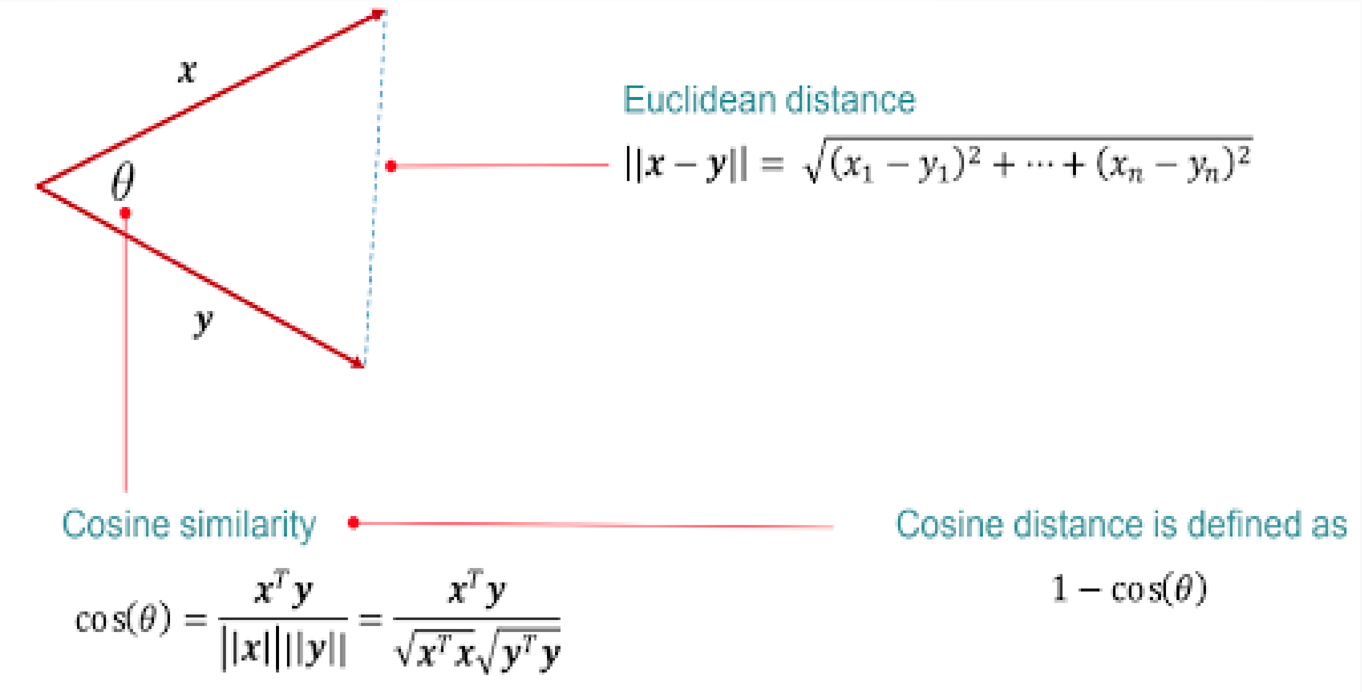
$$\|X\|_2 = \sqrt{x_1^2 + x_2^2 + .. + x_n^2}$$

$$\|X\|_2 = \left(x_1^2 + x_2^2 + .. + x_n^2\right)^{\frac{1}{2}}$$

- generalised to define a p-norm of a vector:

$$\|X\|_p = (|x_1|^p + |x_2|^p + .. + |x_n|^p)^{\frac{1}{p}}$$

Distances between vectors

- Cosine similarity - measures the cosine of the angle between two vectors
  - measure of similarity between two vectors of an inner product space
- 

Matrix

- Matrix has number of rows and columns

$$\begin{array}{c} \text{column vector} \\ A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2_n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \end{array}$$

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2_n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{array}{l} \\ \text{row vector} \\ \\ \\ \end{array}$$

$$\begin{array}{c} \text{number of columns} \\ A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2_n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \\ \text{number of rows} \end{array}$$

$$A \in \mathbb{R}^{\underset{\text{number of rows}}{m} \times \overset{\text{number of columns}}{n}}$$

Matrix types

- Rectangular and Square Matrices
  - If a matrix A has size m×n such that m=n, then it is called a square matrix; otherwise it is a rectangular matrix

$$\begin{bmatrix} 1 & 6 \\ 2 & 3 \end{bmatrix} \qquad \begin{bmatrix} 1 & 2 & 5 \\ 6 & 2 & 4 \end{bmatrix}$$

square matrix rectangular matrix

Matrix types

- Symmetric Matrices
  - a matrix is symmetric if it is equal to its transpose, that is $A = A^T$

$$\begin{bmatrix} 7 & 6 \\ 6 & 7 \end{bmatrix}$$

- Symmetric matrices are always square.

Matrix types

- Diagonal Matrix
  - A matrix A is called a diagonal matrix if A(i,j)=0 for all i≠j.
  - Diagonal matrix is always a square matrix.

$$A = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 5 \end{bmatrix}$$

- Identity Matrix
  - A matrix *I* is called an identity matrix if it is a diagonal matrix and *I*(i,i)=1
  - I_{nxn} denotes nxn identity matrix.

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

Matrix operations

- Transpose of a Matrix
  - putting all the matrix elements on rows on its columns. Lets say B is transpose of A, then B(i,j)=A(j,i)

$$\begin{bmatrix} 1 & 6 & 7 \\ 2 & 3 & 8 \end{bmatrix} \xleftrightarrow{\text{transpose}} \begin{bmatrix} 1 & 2 \\ 6 & 3 \\ 7 & 8 \end{bmatrix}$$

Matrix operations

- Matrix Addition/Subtraction
  - two matrices of same size

$$X + Y = \begin{bmatrix} 2 & 4 \\ 3 & 1 \\ 8 & 5 \end{bmatrix} + \begin{bmatrix} 6 & 7 \\ 4 & 4 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 8 & 11 \\ 7 & 5 \\ 9 & 8 \end{bmatrix}$$

- Scalar Multiplication/Division
  - to multiply a matrix A with scalar c, multiply each element of A with c

$$3x \begin{bmatrix} 6 & 7 \\ 4 & 4 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 18 & 21 \\ 12 & 12 \\ 3 & 9 \end{bmatrix}$$

- Element wise Matrix Multiplication
  - matrices have the same size

$$\begin{bmatrix} 2 & 4 \\ 3 & 1 \\ 8 & 5 \end{bmatrix} \odot \begin{bmatrix} 6 & 7 \\ 4 & 4 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 12 & 28 \\ 12 & 4 \\ 8 & 15 \end{bmatrix}$$

Matrix operations

- Matrix to Matrix Multiplication
  - number of columns in the first matrix is equal to the number of rows in the second matrix
  - Consider AB=C. Now C(i,j) is computed by dot product of A(i,:) and B(:,j)

$$\begin{bmatrix} 2 & 4 \\ 5 & 6 \\ 1 & 7 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 6 & 16 \\ 11 & 28 \\ 8 & 23 \end{bmatrix}$$

  - Matrix multiplication is NOT commutative. Multiplication order matters. In general AB≠BA.

Matrix operations

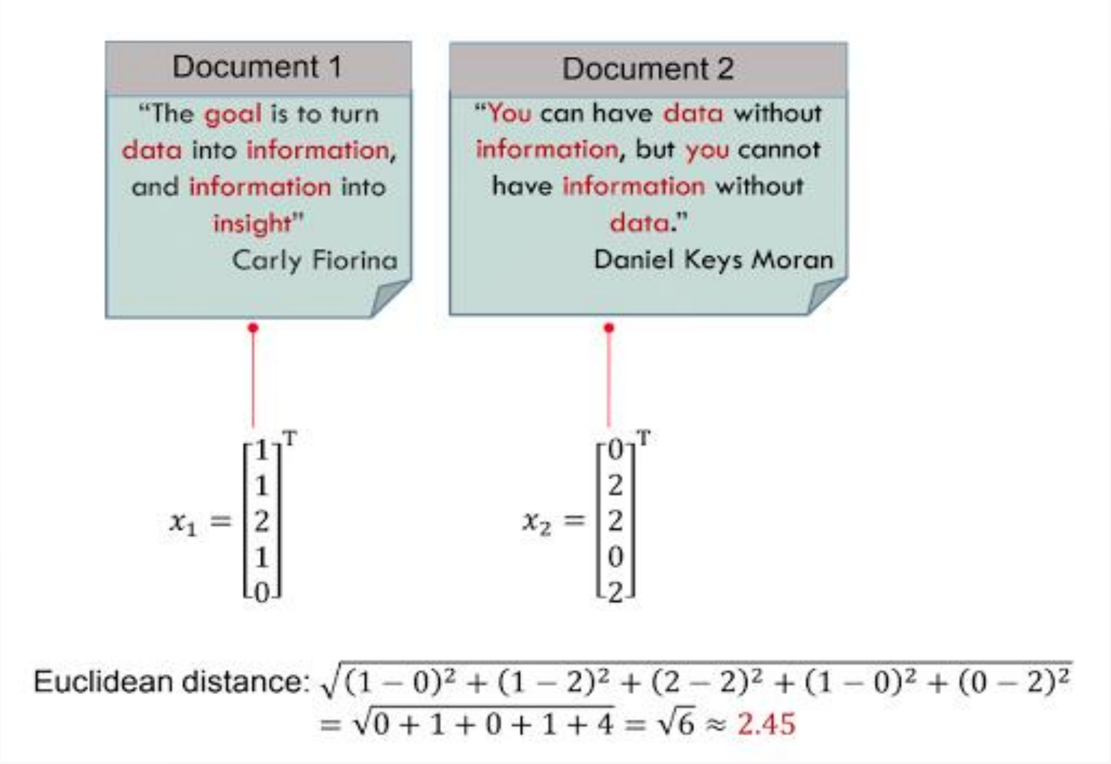
- Inverse Matrix
  - Matrix A is called as inverse of matrix B, if and only if *BA*=*AB*=*I*.
  - Since AB=BA, both A and B need to be a square matrix
  - If A is inverse of B, we denote it as $A = B^{-1}$
  - Inverse of a matrix A exists only if it’s determinant is nonzero
-

Revising knowledge of Linear Algebra and Probability

- Linear Algebra
  - Vector & their operations
  - Matrix & their operations
  - Feature vectors and matrices
- Probability Concepts
  - Random experiment & Event
  - Joint probability
  - Conditional probability
  - Bayes Rules
- Random variable
  - Distribution of random variables

Feature Vectors

- Vector space model is representation of set of documents as vectors.
- It is a fundamental step in information retrieval operations
- Text data representation as **Feature Vectors**

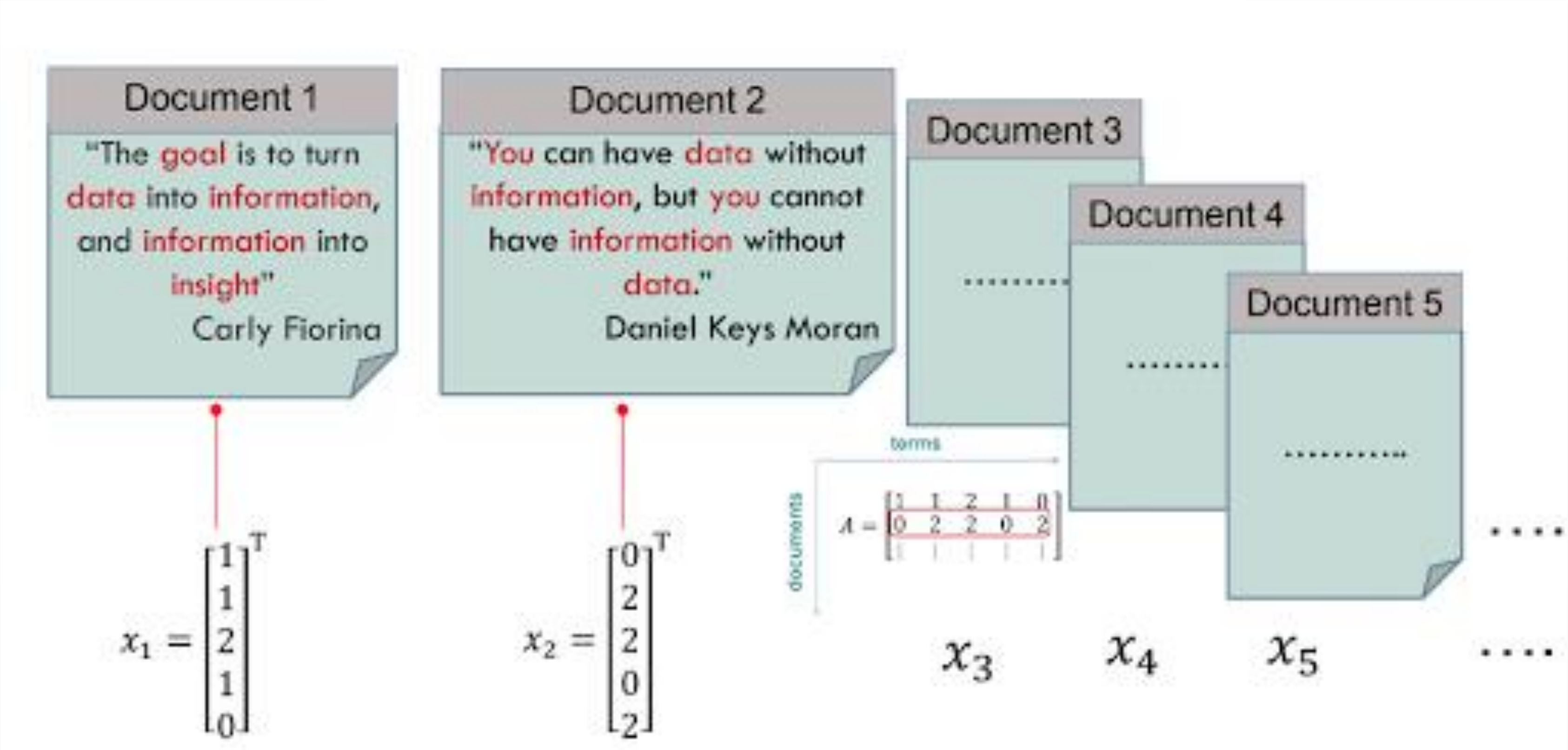


Feature matrix

- We can extend the concept of the feature vector towards a feature matrix by stacking feature vectors as a matrix X
  - We create a vocabulary of features for all the instances in the dataset
  - Represent each instance as a vector on features listed in the vocabulary
  - If our dataset has N instances, we create N vectors x1,x2,,...,xN
  - Each of these vectors is called a feature vector
  - We stack these vectors as a matrix X and call it a feature matrix

Feature matrix

- Example of steps mentioned earlier



Revising knowledge of Linear Algebra and Probability

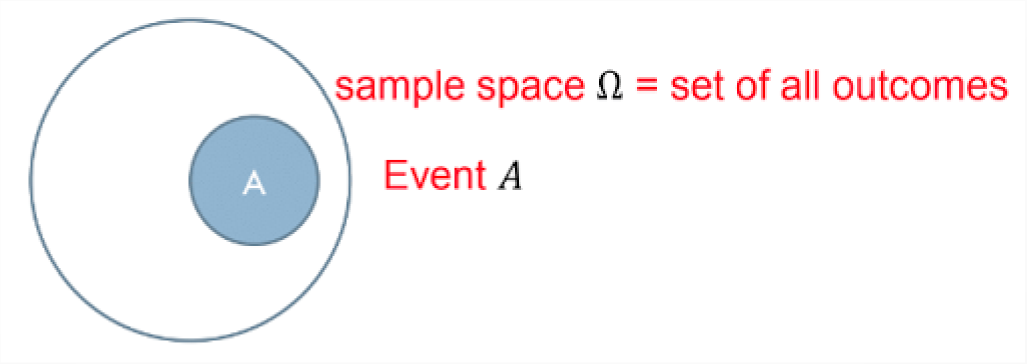
- Linear Algebra
  - Vector & their operations
  - Matrix & their operations
  - Feature vectors and matrices
- Probability Concepts
  - Random experiment & Event
  - Joint probability
  - Conditional probability
  - Bayes Rules
- Random variable
  - Distribution of random variables

Random experiment

- Probability plays a major role in many machine learning algorithms
- **Random experiment**: an experiment or a process for which the outcome cannot be predicted with certainty.
  - toss of a coin
  - roll of a dice
  - counting the number of phone calls received on a mobile phone in a given duration
  - daily temperature
  - how many times a specific word appears in each document of a corpus

Event

- **Event**: a set of outcomes of a random experiment
  - For a coin toss experiment, sample space Ω={head,tail}. And event A could be either {head} or {tail}
  - For a dice roll experiment, sample space Ω={1,2,3,4,5,6} and event A={{1},{2},{3},{4},{5},{6}}



Probability

- **Probability** is defined for an event and is the measure of the likelihood that an event will occur. It is quantified as a number between 0 and 1
- The probability of an event A occurring is denoted as P(A)
- The probability of an event A not occurring is denoted as

$$P(\bar{A}) = 1 - P(A)$$

Joint Probability

- Probability can be defined jointly for more than one event. Consider a random experiment where we toss two coins
- In this case the probability of seeing “head for coin-1” and “head for coin-2” is an example of two events. If two events, A and B are independent then the joint probability is
- Assuming fair coins with probability of head as 1/2. So the probability of head for the first coin and also head for the second coin is:

$$P(A \text{ and } B) = P(A)P(B)$$

$$P(\{head - first\}\text{and}\{head - second\}) = \frac{1}{2} * \frac{1}{2}$$

Conditional Probability

- is the probability of some event A, given the occurrence of another event B.
- Condition probability P(A|B), read as the probability of A given B is defined as
- Provided P(B) is not zero

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

Byes Rule

- essence of most of Bayesian approaches
- mathematical rule explaining how you should change your existing beliefs in the light of new occurrence
- Bayes rule describes the probability of an event A based on another event B that is related to A
- if cancer is related to age, using Bayes' rule information about a person's age can be used to more accurately assess the probability that the person has cancer.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \text{ in which} P(B) \neq 0$$

Revising knowledge of Linear Algebra and Probability

- Linear Algebra
  - Vector & their operations
  - Matrix & their operations
  - Feature vectors and matrices
- Probability Concepts
  - Random experiment & Event
  - Joint probability
  - Conditional probability
  - Bayes Rules
- Random variable
  - Distribution of random variables

Random Variable

- is a variable whose possible values are the generated outcomes of a random phenomenon
- is a function that can assign probabilities to events of interest in a random experiment
- if we toss a coin the possible outcomes are head or tail. Let us define a random variable X so that X=1 means head and X=0 means tail.
- *The function is nothing but the mapping X=1 to head or X=0 to tail*.
- let's say P(head)=0.6, P(tail)=0.4. Then we can say P(X=1)=0.6, P(X=0)=0.4
- This way a random variable can assign a probability to all possible outcomes of a random experiment

Random Variable...

- Two type of random variables
- Discrete Random Variable
  - countable number of values (i.e., faces of a dice, number of emails received in an hour)
- Continuous Random Variable
  - can take values on a infinite continuum (i.e., height of a person, time to failure)

Discrete Random Variable

- Discrete Random Variable
  - defined using a **Probability Mass Functions (PMF)**, denoted as π(x)
  - The PMF assigns a probability to each possible value of the random variable as π(x)=P(X=x) summing them to 1, i.e.

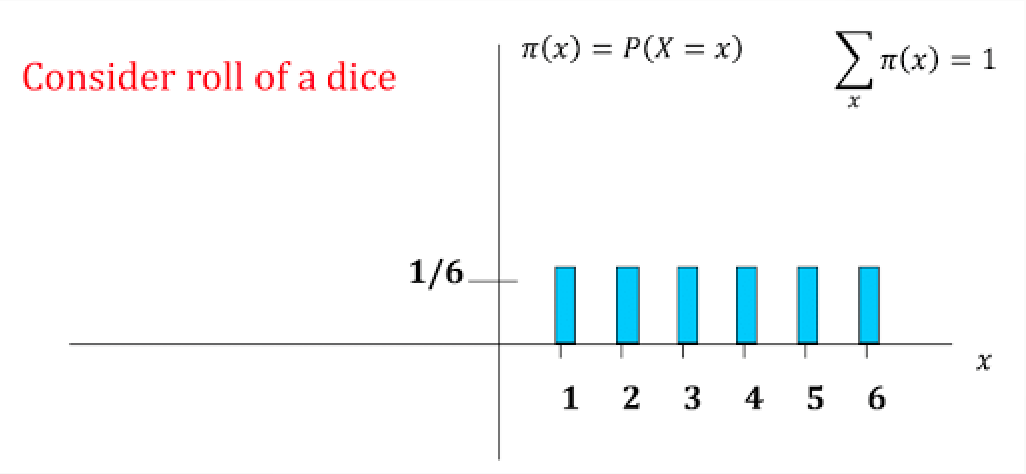
$$\sum_{x} \Pi\left(x\right) = 1$$

  - Rolling a dice is a perfect example of random variables. But what if someone asks about the probability of rolling a dice and getting a number less than 5?

Discrete Random Variable...

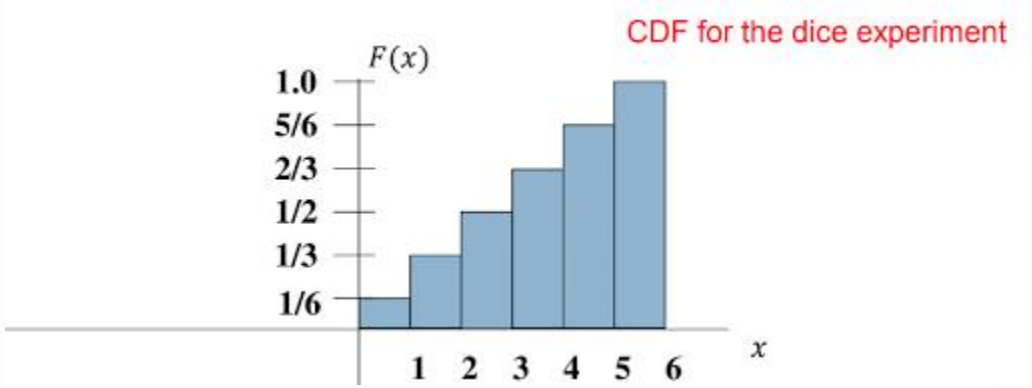
- In such cases we have to work with Cumulative Distribution Function (CDF).
- The cumulative distribution function gives us the cumulative probability associated with a function. it is defined as:

$$F(X) = P(X \leq x) = \sum_{x_i \leq x} P\,(X = x_i)$$



Discrete Random Variable...

- In the figure, it is discontinuous at points x_i's and constant in between



- The probability of seeing a number equal or less than five is
- probability of seeing a number greater than five is

$$P(X \leq 5) = \frac{5}{6}$$

$$P(X > 5) = 1 - \frac{5}{6} = \frac{1}{6}$$

Continuous Random Variable

- Continuous random variables are defined using Probability Density Functions (PDF, *statistical expression*) , denoted as f(x)
- PDF assigns a probability to a range of values of the random variable as $f(x)d(x) = P(x \leq X \leq x + dx)$integrating to 1
- So we can say: $\int_{-\infty}^{+\infty} f(x)dx = 1$
- Probability assigned at any exact value is zero (in the continues space). But we can talk about probabilities over a range such as

$$P(X \geq a), P(X < b) \text{or} P(c \leq X \leq d)$$

Distribution of Random Variable

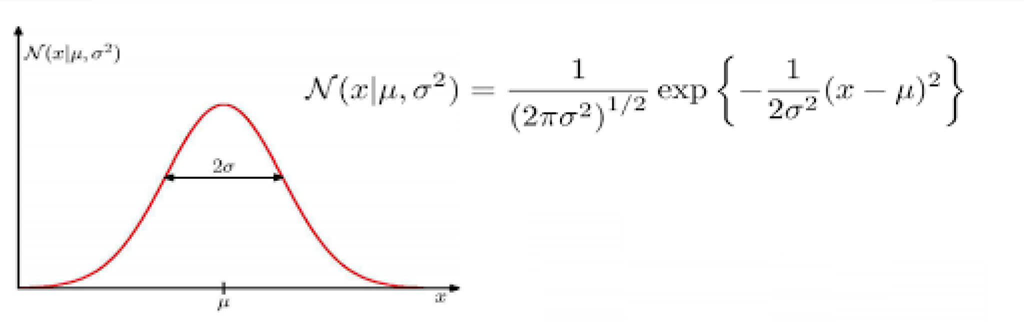
- Probability distribution is a function that links each outcome of a statistical experiment with its probability of occurrence
- **Bernoulli distribution**
  - discrete distribution and defined for a binary random variable with values X=0 and X=1
  - so $\pi(0) = P(X = 0) = p$ and $\pi(1) = P(X = 1) = 1 - p$
  - B means Bernoulli in notation $\pi(x) = B(x||p)$ or $x \sim B(x||p)$
  - For example
    - we can say a distribution over the outcome of an exam is Bernoulli. We may pass (x=1) or fail (x=0)

Distribution of Random Variable...

- **Uniform distribution**
  - can be defined for both discrete and continuous random variables. For a discrete random variable
  - For discrete $\pi(x_i) = P(X = x_i) = \frac{1}{N}, i = 1..N$
  - U means uniform in notation $\pi(x) = U(x||N) \vee x \sim U(x||N)$
  - For a continuous random variable $f(x) = \frac{1}{b-a}, a \leq x \leq b$
  - U means uniform in notation $f(x) = U(x||a, b) \vee x \sim U(x||a, b)$
  - Rolling a fair dice follows a uniform distribution (discrete space)

Distribution of Random Variable...

- **Normal distribution**
  - is defined for continuous random variables
  - most popular distribution
  - defined as $N(x||\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$
  - where N means normal
  - popular because natural phenomena are approximately following a normal distribution

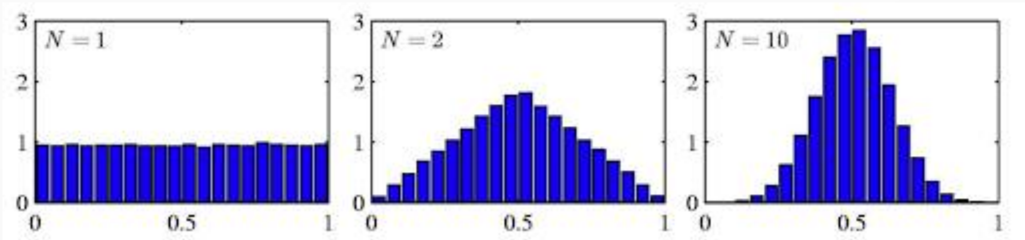


Distribution of Random Variable...

- **Central limit theorem**
  - if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the existing population, then the distribution of the sample means will be approximately normally distributed
  - So, the distribution of the sum of *N* i.i.d.(independent and identically distributed) random variables becomes increasingly normal (Gaussian) as *N* grows

$$Y = X_1 + X_2 + \ldots + X_N$$

  - N uniform [0,1] random variables, following central limit theorem
  -



Thank You