**In-Class Learning Activity 1(LA1)**

**Part I :**

**a)**

Data mining also known as knowledge discovery in data (KDD) is the process of uncovering patterns and other valuable information from large data sets.

Some areas of overlap include: supervised and unsupervised(dimension reduction & clustering)

**b)**

**Bernoulli Distribution:**

Probability Mass Function (PMF):

*P*(*X*=*k*)= 🡪 p if k=1

* q=1−p if k=0

Cumulative Distribution Function (CDF):

F(x)= 🡪

0 for x<0

1−q for 0≤x<1

1 for x≥1

Example:

A Bernoulli distribution could model the outcome of a single coin flip, where p is the probability of getting heads.

**Binomial Distribution:**

Probability Mass Function (PMF):

*P*(*X*=*k*)=(*kn*​)*pk*(1−*p*)*n*−*k*

Cumulative Distribution Function (CDF):

*F*(*x*)=∑*k*=0⌊*x*⌋​(*kn*​)*pk*(1−*p*)*n*−*k*

Example:

If you flip a coin (with p=0.5)5 times the binomial distribution can model the probability of getting exactly k heads.

**Poisson Distribution:**

Probability Mass Function (PMF):

*P*(*X*=*k*)=*k*!*e*−*λλk*  
Cumulative Distribution Function (CDF):

*F*(*x*)=∑*i*=0⌊*x*⌋​*i*!*e*−*λλi*​  
Example:

The Poisson distribution is useful for modeling the number of COVID patients arriving in the ICU between 9 am - 5 pm if the average arrival rate is known (e.g., an average of 2 patients per hour). The parameter λ is the average rate of events per unit of time.

**Uniform Distribution:**

Probability Density Function (PDF):

*f*(*x*)=*b*−*a*1​ for *a*≤*x*≤*b*

Cumulative Distribution Function (CDF):

*F*(*x*)=*b*−*ax*−*a*​ for *a*≤*x*≤*b*

Example:

A uniform distribution could model the probability of a randomly selected point in the interval [0, 1].

**Gaussian/Normal Distribution:**

Probability Density Function (PDF):

*f*(*x*)=2*πσ*2​1​*e*−2*σ*2(*x*−*μ*)2​  
Cumulative Distribution Function (CDF):

*F*(*x*)=21​[1+erf(*σ*2​*x*−*μ*​)]

Example:

The normal distribution is often used to model the distribution of heights in a population.

**Student t Distribution:**

Probability Density Function (PDF):

*f*(*x*)=*πν*​Γ(2*ν*​)Γ(2*ν*+1​)​(1+*νx*2​)−2*ν*+1​

Cumulative Distribution Function (CDF):

*F*(*x*)=21​+21​⋅sign(*x*)⋅*I*2*ν*​,21​​(1+*x*2*ν*​)

**Example:** The Student t distribution is used in hypothesis testing, particularly in t-tests.

**Chi-squared Distribution**:

Probability Density Function (PDF):

*f*(*x*)=22*k*​Γ(2*k*​)1​*x*2*k*​−1*e*−2*x*​

Cumulative Distribution Function (CDF):

*F*(*x*)=*γ*(2*k*​,2*x*​)

**Example:** The chi-squared distribution is commonly used in hypothesis testing and confidence interval construction.

**Gamma Distribution:**

Probability Density Function (PDF):

*f*(*x*)=*θk*Γ(*k*)*xk*−1*e*−*θx*​​

Cumulative Distribution Function (CDF):

*F*(*x*)=Γ(*k*)*γ*(*k*,*x*/*θ*)​  
**Example:** The gamma distribution is often used to model wait times until a Poisson process reaches a certain number of events.

**Beta Distribution:**

Probability Density Function (PDF):

*f*(*x*)=*B*(*α*,*β*)*xα*−1(1−*x*)*β*−1​

Cumulative Distribution Function (CDF):

*F*(*x*)=*Ix*+(1−*x*)*x*​​(*α*,*β*)

**Example:** The beta distribution is commonly used as a conjugate prior in Bayesian statistics.

**Pareto Distribution:**

Probability Density Function (PDF):

*f*(*x*)=*xα*+1*αxmα*​​  
Cumulative Distribution Function (CDF):

*F*(*x*)=1−(*xxm*​​)*α*

**Example:** The Pareto distribution is often used to model the distribution of wealth in a population.

**RV** stands for Random Variable, it is a variable that can take on different values based on the outcome of a random event.

There are two main types of random variables: discrete and continuous

**Part - 2**

1. **Importing data:**

* **Directly importing:**

> data(iris)

* **(using read.csv)**

> my.data<- read.csv("iris.csv", sep = ',', header = TRUE)

> my.data

1. **Exporting Data:**

* Saving to Rdata file format

> save(iris,file='iris.rda') -----------> save the data into rda format

> load(file=’iris.rda’)

* **To save the change to the csv file:**

> write.csv(iris, file='iris.csv')