UNIT II

Statistical Modelling

- It is a Process of applying statistical analysis to a dataset.

- A statistical model is a mathematical representation (or mathematical model) of observed data and is a collection of probability distributions on a set of all possible outcomes of an experiment

- When data analysts apply various statistical models to the data they are investigating, they are able to understand and interpret the information more strategically.

- This practice allows data analysts to identify relationships between variables, make predictions about future sets of data, and visualize that data so that non-analysts can understand and use it.

- Common data sets for statistical analysis include Internet of Things (IoT) sensors, census data, public health data, social media data, imagery data, and other public sector data that benefit from real-world predictions.

Statistical Modelling Techniques

- Supervised

  Regression model

  Classification model:

- Un Supervised

  Clustering

  Reinforcement learning

3 main types of statistical models:

- parametric
- nonparametric
- Semiparametric

How to make a statistical model?

- Start with univariate descriptives and graphs. Visualizing the data helps with identifying errors, understanding the variables you're working with.
- Build predictors in theoretically distinct sets first in order to observe how related variables work together.
- Next, run bivariate descriptives with graphs in order to visualize and understand how each potential predictor relates individually to every other predictor and to the outcome.
- Record, compare and interpret results from models
- Eliminate non-significant interactions first

Machine Learning vs Statistical Modeling

- Machine learning models seek out patterns hidden in data independent of all assumptions and hence predictive power is typically very strong.
- Requires little human input and does well with large numbers of attributes and observations.

- Statistical modelling seeks out relationships between variables in order to predict an outcome.
- They are based on coefficient estimation, and are typically applied to smaller sets of data with fewer attributes.
- Require the human designer to understand the relationships between variables before inputting.

Reasons to Learn Statistical Modeling:

- Will be better equipped to choose the right model for your needs.
- Will be better able to prepare your data for analysis.

Random variable

- A random variable is a numerical description of the outcome of a statistical experiment.

- If a random variable assumes only a finite number or an infinite sequence of values , it is discrete

  and one that may assume any value in some interval on the real number line is said to be continuous.

Ex: No. of automobiles sold at a particular showroom on one day– discrete

Weight of a person in kilograms -- continuous.

- The probability distribution for a random variable describes how the probabilities are distributed over the values of the random variable.
- For a discrete random variable, x, the probability distribution is defined by a probability mass function, denoted by f(x).
- The probability function for a discrete random variable must satisfy two conditions :

  f(x) must be nonnegative for each value of the random variable

  The sum of the probabilities for each value of the random variable must equal one.