IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

MACHINE LEARNING TECHNIQUE
k-NEAREST NEIGHBORS (k-NN)

LECTURE 28

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

k-NEAREST NEIGHBORS (k-NN)

- k-NN
  - No assumptions about the form of relationship between outcome variable and the set of predictors
  - Non-parametric method
    - No parameters from the assumed functional form to estimate
  - Useful information for modeling is extracted using the similarities between the records based on predictors' values
    - Typically, distance based similarity measures are used

k-NEAREST NEIGHBORS (k-NN)

- k-NN: distance metrics
  - Most popular metric is Euclidean distance

  For two records having values of the predictors denoted by $(x_1, x_2, \ldots, x_p)$ and $(w_1, w_2, \ldots, w_p)$

$$D_{Eu} = \sqrt{(x_1 - w_1)^2 + (x_2 - w_2)^2 + \ldots + (x_p - w_p)^2}$$

    - Low computation costs
  - Other distance metrics: statistical distance or Mahalanobis distance and Manhattan distance
  - Euclidean distance is preferred in k-NN due to many distance computations

k-NEAREST NEIGHBORS (k-NN)

- k-NN
  - Scaling of predictors: standardized values of predictors
- k-NN for Classification task
  - Main idea is to find k records in the training partition which are neighboring the new observation to be classified
  - These k neighbors are used to classify the new observation into a class
    - Predominant class among the neighbors

k-NEAREST NEIGHBORS (k-NN)

- k-NN: Finding neighbors and Classification
  - Compute the distance between the new observation and training partition records
  - Determine k nearest or closest records to the new observation
  - Find most prevalent class among k neighbors and it would be the predicted class of new observation

- Open RStudio

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…