IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

MACHINE LEARNING TECHNIQUE
k-NEAREST NEIGHBORS (k-NN) PART 3

LECTURE 30

DR. GAUAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

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

k-NEAREST NEIGHBORS (k-NN)

- k-NN
  - Choosing appropriate value of k
  - k=1: powerful for large no. of records in training partition
  - k>1: smoothing effects (control overfitting issues)
  - Low value of k -> more likely to fit the noise
  - High value of k -> more likely to ignore the local patterns in the data
  - Trade-off between benefits from local pattern vs global effects
  - k=n: naïve rule

k-NEAREST NEIGHBORS (k-NN)

- k-NN
  - Value of k: depends on nature of the data as well
    - Low value of k for data with complex and irregular structures
  - Typical value of k: between ‘1-20’
  - Odd value of k is preferred to avoid ties in majority class decisions
- Best value of k
  - Classification performance on validation partition
- Open RStudio

k-NEAREST NEIGHBORS (k-NN)

- Majority decision rule vs. cutoff probability
  - Two class scenario: majority rule ≡ cutoff value of 0.5

- k-NN for multi-class scenario

- Class of interest
  - Instead of the majority rule, compare proportion of k neighbors belonging to class of interest to a user-specified cut off value

k-NEAREST NEIGHBORS (k-NN)

- k-NN for Prediction task
  - Main idea is to find k records in the training partition which are neighboring the new observation to be predicted
  - These k neighbors are used to predict the value of new observation
    - Average value of the outcome variable among the neighbors
    - Weighted average wherein weight for a neighbor decreases as its distance from new observation increases
  - Performance metric: RMSE or some other prediction error metric

k-NEAREST NEIGHBORS (k-NN)

- k-NN: Finding neighbors and Prediction
  - Compute the distance between the new observation and training partition records
  - Determine k nearest or closest records to the new observation
  - Compute the average or weighted average of outcome variable values among k neighbors and it would be the predicted value of new observation

k-NEAREST NEIGHBORS (k-NN)

- Further Comments on k-NN algorithm
  - Computation time to find nearest neighbors for large training partition
    - Dimension reduction techniques
    - Steps to find neighbors can be optimized using efficient data structures for search operations like trees
    - Identification and pruning of redundant records from training partition which will not be included in neighbor search steps
  - Curse of dimensionality
    - Sample size requirement depends on no. of predictors
    - Leads to more computations for neighbors

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…