IIT ROORKEE

NPTEL

NPTEL ONLINE CERTIFICATION COURSE

DATA MINING PROCESS

LECTURE 02

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

DATA MINING PROCESS

- Phases in a typical Data Mining effort:

  1. Discovery

     Frame business problem

     Identify analytics component

     Formulate initial hypotheses

  2. Data Preparation

     Obtain dataset form internal and external sources

     Data consistency checks in terms of definitions of fields, units of measurement, time periods etc.,

     Sample

DATA MINING PROCESS

- Phases in a typical Data Mining effort:

  3. Data Exploration and Conditioning

     Missing data handling, Range reasonability, Outliers,

     Graphical or Visual Analysis

     Transformation, Creation of new variables, and Normalization

     Partitioning into Training, Validation, and Test datasets

DATA MINING PROCESS

- Phases in a typical Data Mining effort:

4. Model Planning

   Determine data mining task such as prediction, classification etc.

   Select appropriate data mining methods and techniques such as regression, neural networks, clustering etc.

5. Model Building

   Building different candidate models using selected techniques and their variants using training data

   Refine and select the final model using validation data

   Evaluate the final model on test data

DATA MINING PROCESS

- Phases in a typical Data Mining effort:

   6. Results Interpretation

      Model evaluation using key performance metrics

   7. Model Deployment

      Pilot project to integrate and run the model on operational systems

- Similar data mining methodologies developed by SAS and IBM Modeler (SPSS Clementine) are called SEMAA and CRISP-DM respectively

DATA MINING PROCESS

- Data mining techniques can be divided into Supervised Learning Methods and Unsupervised Learning Methods
- Supervised Learning
  - In supervised learning, algorithms are used to learn the function ‘f’ that can map input variables (X) into output variables (Y)

$$Y = f(X)$$

  - Idea is to approximate ‘f’ such that new data on input variables (X) can predict the output variables (Y) with minimum possible error (Ɛ)

DATA MINING PROCESS

- Supervised Learning problems can be grouped into prediction and classification problems
- Unsupervised Learning
  - In Unsupervised Learning, algorithms are used to learn the underlying structure or patterns hidden in the data
- Unsupervised Learning problems can be grouped into clustering and association rule learning problems

DATA MINING PROCESS

- Target Population
  - Subset of the population under study
  - Results are generalized to the target population
- Sample
  - Subset of the target population
- Simple Random Sampling
  - A sampling method wherein each observation has an equal chance of being selected

DATA MINING PROCESS

- Random Sampling
  - A sampling method wherein each observation does not necessarily have an equal chance of being selected
- Sampling with Replacement
  - Sample values are independent
- Sampling without Replacement
  - Sample values aren’t independent

DATA MINING PROCESS

- Sampling results in less no. of observations than the no. of total observations in the dataset
- Data Mining algorithms
  – Varying limitations on number of observations and variables
- Limitations due to computing power and storage capacity
- Limitations due to statistical software being used
- How many observations to build accurate models?

DATA MINING PROCESS

- Rare Event, e.g., low response rate in advertising by traditional mail or email
  - Oversampling of ‘success’ cases
  - Arises mainly in classification tasks
  - Costs of misclassification
    - Asymmetric costs due to more importance of ‘success’ class
  - Costs of failing to identify ‘success’ cases are generally more than costs of detailed review of all cases
  - Prediction of ‘success’ cases is likely to come at cost of misclassifying more ‘failure’ cases as ‘success’ cases than usual

DATA MINING PROCESS

- Dummy coding for categorical variables
  - Some statistical software cannot use categorical variables expressed in the label format
  - Dummy binary variables (having 0's and 1's: 0 indicating 'absence' and 1 indicating 'presence') for different classes of categorical variables are created
  - For example, if 'activity status' of individuals can be put into four mutually exclusive and jointly exhaustive classes as {student, unemployed, employed, retired}, only three dummy variables would be required

DATA MINING PROCESS

- Principle of Parsimony
  - A model or theory with less no. of assumptions and variables but with high explanatory power is generally desirable
- More no. of variables also increase the sample size requirements due to reliability of estimate
- Overfitting
  - A model built using a complex function that fits the data perfectly
  - Model ends up fitting the noise and explaining the chance variation

DATA MINING PROCESS

- Overfitting
  - More no. of iterations resulting in excessive learning of the data
  - More no. of variables in the model may lead to fitting spurious relationships
- Sample Size
  - Domain Knowledge
  - General rule of thumb: 10 × p observations, where p is the no. of predictors
  - For classification tasks: 6 × m × p observations, where m is the no. of classes in the outcome variable (Delmaster & Hancock, 2001)

DATA MINING PROCESS

- Outliers
  - A distant data point
  - Valid point or erroneous value?
  - Further review
    - Manual Inspection (Sorting, minimum and maximum values, clustering etc.)
    - Domain Knowledge
- Missing Values
  - Few records with missing values can be removed
  - Imputation

DATA MINING PROCESS

- Missing Values
  - Drop the variables having missing values
  - Replace with proxy variable
- Normalization
  - Standardization using z-score
  - Min-max normalization

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…