



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# INTRODUCTION

## LECTURE 01

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# INTRODUCTION

- What is Business Analytics?
  - “Business analytics is comprised of solutions used to build analysis models and simulations to create scenarios, understand realities and predict future states.” - Gartner IT Glossary
  - Includes data mining, predictive analytics, applied analytics and statistics, and is delivered as an application suitable for a business user
- Analytics can be classified as:
  - Descriptive analytics
  - Predictive analytics
  - Prescriptive analytics



# INTRODUCTION

- Descriptive analytics involve gathering, organizing, tabulating, and depicting data and then describing the characteristics of what you are studying.
  - Also called reporting in managerial lingo
  - First phase of analytics
  - Though useful, it doesn't inform you about why the results happen or what can happen in future.



# INTRODUCTION

- Predictive analytics use the past to predict the future.
  - Identify associations among different variables and predict the likelihood of a phenomenon reoccurring on the basis of those relationships
- Correlation vs. Causation
- Prescriptive analytics suggest a course of action.
  - Recommends decisions entailing mathematical and computational models
  - Final phase of analytics



# INTRODUCTION

- Methods from statistics, forecasting, data mining, experimental design are used in Business Analytics
- What is Data Mining?
  - “Extracting useful information from large datasets” - Hand et al. 2001
  - “The process of discovering meaningful correlations, patterns and trends by sifting through large amounts of data stored in repositories. Data mining employs pattern recognition technologies, as well as statistical and mathematical techniques.” - Gartner IT Glossary



# INTRODUCTION

- Where is Data Mining Used?
  - Variety of domains:
    - Predicting the response of a drug or a medical treatment on the patient suffering from a serious disease
    - Predicting whether an intercepted communication is about a potential terror attack
    - Predicting whether a packet of network data can pose a cybersecurity threat



# INTRODUCTION

- Where is Data Mining Used?
  - Common business questions:
    - Which customers are most likely to respond to the marketing or promotional offer?
    - Which customers are most likely to default on loan?
    - Which customers are most likely to subscribe to a magazine?



# INTRODUCTION

- Data Mining Genesis
  - An interdisciplinary subfield of computer science
  - Originates from the fields of machine learning and statistics
  - Data mining as “statistics at scale and speed” – Pregibon 1999
  - Extension: “statistics at scale, speed, and simplicity” – Shmueli et al. 2010



# INTRODUCTION

## Classical Statistical Setting

- Data scarcity and Computational difficulty
- Same sample is used to compute an estimate and to check its reliability
- Logic of inference: confidence intervals and hypothesis tests  
(Inference is determining whether a pattern or result might have happened by chance)

## Data Mining Paradigm

- Large datasets and fast computing powers
- Fitting a model with one sample and evaluating the performance using another sample
- Machine learning techniques, such as trees and neural networks are less structured and more computationally intensive in comparison to statistical techniques



# INTRODUCTION

- Rapid Growth of Data
  - Millions of transactions on a daily basis
    - Organized retailers such as Shoppers Stop, Big Bazaar, and Pantaloons
    - E-commerce retailers such as Flipkart, Amazon, and Snapdeal
  - Growing economy and Internet growth
  - Decreasing cost and increasing availability of automatic data capture mechanisms, e.g., Bar codes, POS devices, click-stream data, GPS data
  - Operational databases to data warehouse and data marts
  - Constant declining cost of data storage and improving processing capabilities



# INTRODUCTION

- Core of this course focuses on
  - Predictive Analytics consisting of tasks of
    - Prediction,
    - Classification,
    - Association rules
- In Data mining, typically several different methods are applied for a particular goal and the most useful is selected



# INTRODUCTION

- Usefulness of a method
  - Goal of the analysis
  - Underlying assumptions of the method
  - Size of the dataset
  - Types of pattern in the dataset
- Dataset Example: Sedan Car owner
  - Goal: Income level and Household Area is used to classify whether a household owns a sedan car



# INTRODUCTION

- Dataset format
  - Tabular or matrix format: variables in columns and observations in rows
  - Each row represents a household (unit of analysis) in SedanCar dataset
- R and RStudio
  - R is a programming language and software environment for statistical computing and graphics.
  - It is widely used by statisticians and data miners
  - RStudio is the most commonly used integrated development environment (IDE) for R.



# INTRODUCTION

- Key Terms
  - Algorithm
    - A specific sequence of actions or set of rules to be followed to perform a task.
    - Algorithms are used to implement data mining techniques such as trees, neural networks etc.
  - Model
    - By model, we mean data mining model here
    - A data mining model is an application of a data mining technique on dataset



# INTRODUCTION

- Key Terms
  - Variable
    - Operationalized way of representing a characteristic of an object, event, or phenomenon
    - A variable can take different values in different situations.
  - Input variable, Independent variable, Feature, Field, Attribute, or Predictor
    - Input variable is an input to the model



# INTRODUCTION

- Key Terms
  - Output variable, Outcome variable , Dependent variable, Target variable, or Response
    - Output variable is an output of the model
  - Record, observation, case, row
    - Observation is the unit of analysis on which the variable measurements are taken such as a customer, a household, an organization, an industry etc.



# INTRODUCTION

- In Data Mining and related domains, generally two types of variables are used:
  - Categorical
    - Nominal
    - Ordinal
  - Continuous
    - Interval
    - Ratio



# INTRODUCTION

- Understanding the type of variables in a dataset is important
  - To identify an appropriate statistical or data mining technique
  - Proper interpretation of the data analysis results
- Data of these variable types are either quantitative or qualitative in nature
  - Quantitative data measure numeric values and are expressed in number
  - Qualitative data measure types and are expressed by a label, or a numeric code



# INTRODUCTION

- Structure of these variable types increases from nominal to ratio in a hierarchical fashion
- Nominal
  - Values indicate distinct types, e.g., gender, nationality, religion, PIN code, employee ID
  - Only two operations = and ≠ are supported



# INTRODUCTION

- **Ordinal**
  - Values indicate a natural order or sequence, e.g., academic grades, Likert scale, quality of a food item
  - Four additional operations  $<$ ,  $\leq$ ,  $>$ ,  $\geq$  are supported
- **Interval**
  - Difference between two values is also meaningful
  - Values may be in reference to a somewhat arbitrary zero point
  - Celsius temperature, Fahrenheit temperature, location variables: Distance from landmarks, geographical coordinates (latitude & longitude), calendar dates



# INTRODUCTION

- Interval
  - Two additional operations +, - are supported
- Ratio
  - Ratio of two values is also meaningful. Values are in reference to an absolute zero point
  - Kelvin temperature, age, length, weight, height, income
  - Two additional operations  $\times$ ,  $\div$  are supported



# INTRODUCTION

- Conversion from one variable type to other
  - High structure variable type can be converted into low structure variable type
  - For example, a ratio variable ‘age’ can be converted into an ordinal variable ‘age group’



# INTRODUCTION

- Course Roadmap
  - Module I: General Overview of Data Mining and its Components
  - Module II: Data Preparation and Exploration
  - Module III: Performance Metrics and Assessment
  - Module IV: Supervised Learning Methods
  - Module V: Unsupervised Learning Methods
  - Module VI: Time Series Forecasting
  - Module VII: Conclusion



# INTRODUCTION

- Supplementary Lectures
  - Introduction to R
  - Basic Statistical Methods



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE

# Key References

- HBR Video (Business Analytics Defined by Thomas H. Davenport)
- Gartner IT Glossary
- 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



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 from internal and external sources

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

- Sample



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE

# 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



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE

# 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 ( $\epsilon$ )



# 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:  $10 \times p$  observations, where  $p$  is the no. of predictors
  - For classification tasks:  $6 \times m \times p$  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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# INTRODUCTION TO R

## LECTURE 03

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# INTRODUCTION to R

- Installation Steps for Windows PC or Laptop
  - Install R
    - Download Link: <https://cran.r-project.org/bin/windows/base/>
  - Install RStudio Desktop
    - Download Link: <https://www.rstudio.com/products/rstudio/download/>
  - Install JAVA if not already installed
    - Download Link: <https://www.java.com/en/download/>
- Open RStudio
  - Installing R packages



# INTRODUCTION to R

- R Graphical User Interface (GUI)
  - Command-line interface (CLI)
  - Similar to BASH shell in LINUX or interactive version of scripting language Python
  - RStudio is a popular GUI for R and it has been used to write R scripts for this course



# INTRODUCTION to R

- RStudio has four main window sections
  - Top-Left Section: To write and save R code (Script section)
  - Bottom-Left Section: To execute R code and output (Console section)
  - Top-Right Section: To manage datasets and variables (Data section)
  - Bottom-Right Section: To display plots and seek help on R functions (Plot and Help Section)



# INTRODUCTION to R

- Dataset Import
  - In this course, datasets are either imported from Excel files or created in RStudio
- Open Rstudio
  - R Basics



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE

# Thanks...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# Basic Statistics Using R

## LECTURE 04

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# Basic Statistics

- Descriptive Statistics
  - Open RStudio
- Hypothesis Testing
  - Formulate an assertion and test it using data
    - Comparing populations, e.g., comparing performance of students in exams for two different class sections
    - Testing the difference of the means from two data samples
  - A common technique to assess the difference or significance of the same



# Basic Statistics

- Common assumption in Hypothesis testing
  - No difference between two samples
  - Referred as NULL Hypothesis  $H_0$
  - Alternative Hypothesis ( $H_A$ ): There is difference between two samples
- Example:
  - $H_0$ : Students from class A and B had same performance in the examinations
  - $H_A$ : Students from class A performed better than students from class B



# Basic Statistics

- Hypothesis test leads to:
  - Either rejection of the null hypothesis in favor of the alternative
  - Or acceptance of the null hypothesis
- Examples:
  - $H_0$ : New data mining model does not predict better than existing model
  - $H_A$ : New data mining model predicts better than existing model



# Basic Statistics

- Examples:
  - $H_0$ : Regression coefficient is zero, i.e., variable has no impact on outcome
  - $H_A$ : Regression coefficient is nonzero, i.e., variable has an impact on outcome
- A typical hypothesis test is comparing the means of two populations
- Normal Distribution
  - A common continuous probability distribution and useful due to Central limit theorem



# Basic Statistics

- Difference of Means
  - Drawing inferences on two populations: P1 and P2
  - Compare means:  $\mu_1$  and  $\mu_2$
  - $H_0: \mu_1 = \mu_2$
  - $H_A: \mu_1 \neq \mu_2$
  - Basic approach: compare observed sample means:  $\bar{x}_1$  and  $\bar{x}_2$
- Student's t-test
  - Assumptions: Two population distributions (P1 and P2) have equal but unknown variances
  - Two samples of  $n_1$  and  $n_2$  observations drawn randomly and independently from P1 and P2, respectively



# Basic Statistics

- Student's t-test
  - If P1 and P2 are normally distributed with same mean and variance
  - Then t-statistic follows a t-distribution with  $n_1+n_2-2$  degrees of freedom

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\text{Where } S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$



# Basic Statistics

- Student's t-test
  - $S_p$  is pooled standard deviation,  $S_1$  and  $S_2$  are sample standard deviation
  - Shape of t-distribution is similar to normal distribution and becomes identical to normal distribution as degrees of freedom reach 30 or more
  - Numerator of t is the difference of the sample means
    - Observed t value of 0 indicates the sample results are exactly equal to  $H_0$
    - Observed t value being far enough from 0 and t-distribution indicating a low enough probability ( $<0.05$ ) will lead to rejection of  $H_0$
    - t-value falling in corresponding areas in the curve less than 5% of the time



# Thanks...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# Basic Statistics Using R Part-2

## LECTURE 05

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# Basic Statistics

- Student's t-test
  - If P1 and P2 are normally distributed with same mean and variance
  - Then t-statistic follows a t-distribution with  $n_1+n_2-2$  degrees of freedom

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\text{Where } S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$



# Basic Statistics

- Student's t-test
  - $S_p$  is pooled standard deviation,  $S_1$  and  $S_2$  are sample standard deviation
  - Shape of t-distribution is similar to normal distribution and becomes identical to normal distribution as degrees of freedom reach 30 or more
  - Numerator of t is the difference of the sample means
    - Observed t value of 0 indicates the sample results are exactly equal to  $H_0$
    - Observed t value being far enough from 0 and t-distribution indicating a low enough probability ( $<0.05$ ) will lead to rejection of  $H_0$
    - t-value falling in corresponding areas in the curve less than 5% of the time



# Basic Statistics

- Student's t-test
  - For a low probability,  $\alpha = 0.05$ , known as significance level of the test
  - $t^*$  is determined such that  $p(|t| \geq t^*) = \alpha$
  - $H_0$  is rejected if observed value of  $t$  is such that  $|t| \geq t^*$
- Significance level of a statistical test is the probability of rejecting the null hypothesis
  - If null hypothesis is true and  $\alpha = 0.05$ , the observed magnitude of  $t$  would exceed  $t^*$  5% of the time



# Basic Statistics

- p-value is sum of  $p(t \leq -|\text{observed t-value}|)$  and  $p(t \geq |\text{observed t-value}|)$
- Open Rstudio
- Welch's t-test
  - Used when assumption of equal population variance is not reasonable

$$t_w = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$



# Basic Statistics

- Welch's t-test
  - Assumption of random samples drawn from two normal populations with the same mean is still applicable
  - t-distribution
- Open RStudio
- Confidence Interval
  - Provide interval estimate of a population parameter using sample data
  - Indicates uncertainty associated with a point estimate
  - How close  $\bar{x}$  is to  $\mu$



# Basic Statistics

- Confidence Interval
  - A 95% confidence interval estimate for a population mean straddles the true unknown mean 95% of the time

$$\mu \in \bar{x} \pm \frac{2\sigma}{\sqrt{n}}$$

- Type I and Type II Errors

	$H_0$ is true	$H_0$ is false
$H_0$ accepted		Type II error
$H_0$ rejected	Type I error	



# Basic Statistics

- Type I and Type II Errors
  - Significance level = type I error (Denoted by  $\alpha$ )
    - Can be managed using appropriate significance level
  - Type II error (Denoted by  $\beta$ )
    - Can be managed using appropriate sample size
- Power of a test
  - Correctly rejecting  $H_0$
  - $1 - \beta$
  - Used to determine the sample size



# Basic Statistics

- ANOVA
  - Used for more than two populations or groups instead of performing multiple t-tests
  - Generalization of hypothesis testing that is used for the difference of two group means
  - For  $n$  groups,  $n(n-1)/2$  t-tests would be required
  - Multiple t-tests
    - Cognitively difficult
    - Increased probability of type I error



# Basic Statistics

- ANOVA
  - $H_0$ : All the population means are equal
  - $H_A$ : At least one pair of the population means is not equal
  - Assumption: Each population is normally distributed with same variance
  - Test whether different population clusters are more tightly grouped or spread across all the populations



# Basic Statistics

- ANOVA
  - Between-groups mean sum of squares ( $S_B^2$ )
    - An estimate of between-groups variance

$$S_B^2 = \frac{1}{k-1} \sum_{i=1}^k n_i (\bar{x}_i - \bar{x}_0)^2$$

Where k=no. of groups,  $n_i$  is no. of observations in ith group,  $\bar{x}_0$  is mean of all the groups,  $\bar{x}_i$  is mean of ith group

- Within-group mean sum of squares ( $S_W^2$ )
  - An estimate of within-group variance



# Basic Statistics

- ANOVA
  - Within-group mean sum of squares ( $S_W^2$ )
$$S_W^2 = \frac{1}{n - k} \sum_{i=1}^k \sum_{j=1}^{n_i} n_i(x_{ij} - \bar{x}_i)^2$$
  - If  $S_B^2 > S_W^2$ , some of the population means are different
  - F-test statistic
- Open RStudio



# Thanks...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# PARTITIONING PROCESS

## LECTURE 6

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# DATA MINING PROCESS

- Partitioning
  - Using same data for model building and model evaluation introduces bias
  - Selection of best model from several candidate models could be due to
    - Genuine superiority of the final model over other candidate models
    - Chance occurrence leading to better match between final model and data
  - Many data-driven techniques can end up producing the latter situation due to overfitting



# DATA MINING PROCESS

- Partitioning
  - Partitioning of dataset into two or three parts can solve this problem
  - Typically three partitions- training, validation, and test sets are created following a predetermined proportions for each set and records are randomly assigned to different partitions
  - Sometimes records are assigned based on a relevant variable



# DATA MINING PROCESS

- Partitioning
  - Training Partition
    - Usually largest
    - To build the candidate models
  - Validation Partition
    - To evaluate the candidate models
    - Or to fine-tune and improve the model
  - Test Partition
    - To evaluate the final model



# DATA MINING PROCESS

- Types of Datasets
  - Cross-Sectional Data
    - Observations on variables related to many subjects (individuals, firms, industries, or countries)
    - Observed at same point of time (snapshot)
    - Unit of analysis is specified
    - Each observation represents a distinct subject
    - Main idea is to compare differences among the subjects



# DATA MINING PROCESS

- Types of Datasets
  - Time Series Data
    - Observations on a variable related to one subject
    - Observed over a successive equally spaced points in time
    - Each observation represents a distinct time period
    - Main idea is to examine changes in the subject over time



# DATA MINING PROCESS

- Types of Datasets
  - Panel Data or Longitudinal Data
    - Observations on variables related to same subjects over a successive equally spaced points in time
    - Main idea is to compare differences among the subjects and to examine changes in the subjects over time
    - Cross-sections with time order



# DATA MINING PROCESS

- Types of Datasets
  - Pooled Cross-Sectional Data
    - Observations on variables related to subjects at different time periods
    - Main idea is to examine the impact on subjects due to environmental changes caused by certain events or policies
    - Independent cross-sections from different time periods



# DATA MINING PROCESS

- Model Building
  - An example with Linear Regression
  - Open RStudio



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE

# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# VISUALIZATION TECHNIQUES

## LECTURE 07

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# VISUALIZATION TECHNIQUES

- “A picture is worth a thousand words”
  - A popular proverb
- Primary Role in the Data Mining Process is during the Data Exploration, and Conditioning phase
  - Understanding data structure
  - Identifying gaps or erroneous values
  - Identifying outliers
  - Finding patterns



# VISUALIZATION TECHNIQUES

- Primary Role in the Data Mining Process is during the Data Exploration, and Conditioning phase
  - Finding missing values
  - Identifying duplicate rows and columns
  - Variable selection, transformation and derivation
    - Appropriate bin sizes for converting continuous variable into categorical variable
    - Combining categories
    - Usefulness of variables and metrics



# VISUALIZATION TECHNIQUES

- Data Exploration and Conditioning
  - Required preliminary step before formal analysis
  - Visual analysis
    - A free-form data exploration
    - Main idea is to support the data mining goal and subsequent formal analysis
    - Techniques range from basic plots to interactive visualizations
    - Features such filtering, zooming, color and multiple panels
  - Usage of Visualization Techniques depends on
    - Different data mining tasks such as classification, prediction, clustering etc.
    - Different data mining techniques such as CART, HAC etc.



# VISUALIZATION TECHNIQUES

- Basic Charts
  - Display one or two variables at a time
  - Useful to understand the structure of the data, variable types, and missing values in the dataset
  - For Supervised learning methods, main focus is on outcome variable
    - Typically plotted on y-axis



# VISUALIZATION TECHNIQUES

- Line Charts or Graphs
  - Used mainly to display time series data
  - Overall level and Changes over time
  - Open RStudio
- Bar Charts
  - For comparing groups using a single statistic
  - X-axis is used for categorical variable
  - 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# VISUALIZATION TECHNIQUES Part-2

## LECTURE 08

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# VISUALIZATION TECHNIQUES

- Scatterplot
  - Useful for prediction tasks
    - Focus is on finding meaningful relationships between numerical variables
  - Useful for unsupervised learning tasks such as clustering
    - Focus is on finding information overlap
  - Both the axis are used for numerical variable
  - Open RStudio



# VISUALIZATION TECHNIQUES

- Distribution Plots
  - Histogram and Boxplot
    - Distribution of a numerical variable
    - Directions for new variable derivations
    - Directions for binning of a numerical variable
  - Useful in supervised learning, specifically prediction tasks
    - Variable transformation in case of a skewed distribution
    - Selection of appropriate data mining method



# VISUALIZATION TECHNIQUES

- Boxplots
  - Display entire distribution
  - Side-by-side boxplots for comparing groups
    - Importance of numerical predictors in classification tasks
  - Series of boxplots for changes in distributions over time
  - Open RStudio



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE

# VISUALIZATION TECHNIQUES

- Histograms
  - Display frequencies covering all the values
  - Vertical Bars are used
  - Open RStudio
- Heatmaps
  - Display numeric variables using graphics based on 2-D tables
    - Color schemes are used to indicate values
  - Useful to visualize correlation and missing values
    - Specially, in case of large no. of values



# VISUALIZATION TECHNIQUES

- Multidimensional Visualization
  - Multiple panels
  - Color
  - Size and shape
  - Animation
  - Aggregation, rescaling, and Interactivity
  - Main idea is to help build visual perception to support the subsequent analysis
- Open RStudio



# VISUALIZATION TECHNIQUES

- Multidimensional Visualization
  - Trend Lines
  - In-plot labels
  - Scaling Up
  - Multivariate plots
- Specialized Visualization
  - Network graphs (Network data)
  - Treemaps (Hierarchical data)
  - Map Charts (Geographical data)
- 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# VISUALIZATION TECHNIQUES Part-3

## LECTURE 09

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# VISUALIZATION TECHNIQUES

- Histograms
  - Display frequencies covering all the values
  - Vertical Bars are used
  - Open RStudio
- Heatmaps
  - Display numeric variables using graphics based on 2-D tables
    - Color schemes are used to indicate values
  - Useful to visualize correlation and missing values
    - Specially, in case of large no. of values



# VISUALIZATION TECHNIQUES

- Multidimensional Visualization
  - Multiple panels
  - Color
  - Size and shape
  - Animation
  - Aggregation, rescaling, and Interactivity
  - Main idea is to help build visual perception to support the subsequent analysis
- 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# VISUALIZATION TECHNIQUES Part-4

## LECTURE 10

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# VISUALIZATION TECHNIQUES

- Multidimensional Visualization
  - Multiple panels
  - Color
  - Size and shape
  - Animation
  - Aggregation, rescaling, and Interactivity
  - Main idea is to help build visual perception to support the subsequent analysis
- 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# VISUALIZATION TECHNIQUES Part-5

## LECTURE 11

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# VISUALIZATION TECHNIQUES

- Multidimensional Visualization
  - Trend Lines
  - In-plot labels
  - Scaling Up
  - Multivariate plots
- Specialized Visualization
  - Network graphs (Network data)
  - Treemaps (Hierarchical data)
  - Map Charts (Geographical data)
- 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# VISUALIZATION TECHNIQUES Part-6

## LECTURE 12

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# VISUALIZATION TECHNIQUES

- Multidimensional Visualization
  - Trend Lines
  - In-plot labels
  - Scaling Up
  - Multivariate plots
- Specialized Visualization
  - Network graphs (Network data)
  - Treemaps (Hierarchical data)
  - Map Charts (Geographical data)
- 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# DIMENSION REDUCTION TECHNIQUES

## LECTURE 13

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# DIMENSION REDUCTION TECHNIQUES

- Large no. of variables
  - Subsets of variables might be highly correlated
  - Computational issues
  - Costs of data preparation, exploration, and conditioning
  - Dimensionality (Principle of Parsimony)
- Dimension Reduction is also called as factor selection or feature extraction in some domains



# DIMENSION REDUCTION TECHNIQUES

- Dimension Reduction Techniques
  - Domain Knowledge
  - Data Exploration Techniques
  - Data Conversion Techniques
  - Automated reduction Techniques
  - Data Mining Techniques



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE

# DIMENSION REDUCTION TECHNIQUES

- Domain Knowledge
  - Identifying key variables for the data mining task
  - Removing redundant variables
  - Identifying erroneous variables
  - Measurement issues for variables



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE

# DIMENSION REDUCTION TECHNIQUES

- Data Exploration Techniques
  - Descriptive statistics
    - Summary statistics
    - Pivot tables
    - Correlation analysis
  - Visualization Techniques
- Open RStudio



# DIMENSION REDUCTION TECHNIQUES

- Data Conversion Techniques
  - Combining categories
  - Converting a categorical variable into a numerical variable
- RStudio
- Automated reduction Techniques
  - Principal Component Analysis (PCA)



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE

# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# DIMENSION REDUCTION TECHNIQUES Part 2

## LECTURE 14

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# DIMENSION REDUCTION TECHNIQUES

- Principal Component Analysis (PCA)
  - Used for reducing the no. of predictors
  - Used for quantitative variables
  - Highly correlated variable subsets
  - Main idea is to find a set of new variables that contains most of the information of original variables
  - Eliminating covariation and multicollinearity
  - Redistribution of variability
- Open RStudio



# DIMENSION REDUCTION TECHNIQUES

- Principal Component Analysis (PCA)
  - Data Mining Process
    - Apply PCA to the training partition
    - Predictors would now be principal score columns
    - Apply the principal weights obtained from training partition to the variables in the validation partition to obtain the scores
  - Relationship between predictors and output variable is ignored



# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# DIMENSION REDUCTION TECHNIQUES Part-3

## LECTURE 15

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES



# DIMENSION REDUCTION TECHNIQUES

- Principal Component Analysis (PCA)
  - Used for reducing the no. of predictors
  - Used for quantitative variables
  - Highly correlated variable subsets
  - Main idea is to find a set of new variables that contains most of the information of original variables
  - Eliminating covariation and multicollinearity
  - Redistribution of variability
- Open RStudio



# DIMENSION REDUCTION TECHNIQUES

- Principal Component Analysis (PCA)
  - Data Mining Process
    - Apply PCA to the training partition
    - Predictors would now be principal score columns
    - Apply the principal weights obtained from training partition to the variables in the validation partition to obtain the scores
  - Relationship between predictors and output variable is ignored



# DIMENSION REDUCTION TECHNIQUES

- Data Mining Techniques
  - Subset selection procedures using Regression models
    - Linear regression for prediction
    - Logistic regression for classification
    - Regression models can also be used for combining categories (using p-values)
  - Classification and Regression Tree (CART)
    - Classification tree for classification
    - Regression tree for prediction (Using tree diagram)



# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# PERFORMANCE METRICS

## LECTURE 16

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# PERFORMANCE METRICS

- Need for performance metrics
  - Usefulness of a model
  - Comparison of candidate models
- Classification Performance
  - Probability of misclassification
  - Naïve Rule: most prevalent class
    - Serves as benchmark
  - Class Separation
- Open RStudio



# PERFORMANCE METRICS

- Performance Metrics based on Naïve Rule
  - Multiple R<sup>2</sup>
    - Distance between fit of model to data and fit of naïve rule to data
- Naïve rule equivalent for prediction
  - Sample mean
- Classification Matrix
  - $n_{i,j}$ : no. of class i cases classified as class j cases

Classification Matrix		
	Predicted Class	
Actual Class	1	0
1	$n_{1,1}$	$n_{1,0}$
0	$n_{0,1}$	$n_{0,0}$



# PERFORMANCE METRICS

- Classification Performance
  - Validation partition classification matrix
  - Comparison of training partition classification matrix with validation partition classification matrix
    - Detect overfitting
- Performance Metrics based on classification matrix
  - Misclassification rate or error
  - Accuracy



# PERFORMANCE METRICS

- Performance Metrics based on classification matrix

$$\text{err} = \frac{n_{0,1} + n_{1,0}}{n}$$

$$\text{accuracy} = 1 - \text{err} = \frac{n_{0,0} + n_{1,1}}{n}$$

- Open RStudio
- Cutoff probability value
  - Accuracy for all the classes is important
    - A case is assigned to the class with the highest probability as estimated by the model

# PERFORMANCE METRICS

- Cutoff probability value
  - Accuracy for a particular class of interest is important
    - A case is assigned to the class of interest if probability for the class is above cutoff value
  - Default cutoff value for a two class model is 0.5 (principally similar to naïve rule)
- Open Excel
  - One-variable table



# PERFORMANCE METRICS

- Why change cutoff probability value from 0.5?
  - Class of interest
  - Asymmetric misclassification cost
- When to incorporate change in cutoff value?
  - After final model selection
  - Before model derivation



# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# PERFORMANCE METRICS Part-2

## LECTURE 17

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# PERFORMANCE METRICS

- Performance Metrics in presence of a class of interest

$$\text{sensitivity} = \frac{n_{11}}{n_{10} + n_{11}} = \text{true positive fraction}$$

$$\text{specificity} = \frac{n_{00}}{n_{00} + n_{01}} = \text{true negative fraction}$$

- ROC (receiver operating characteristic) curve
  - Used to plot {sensitivity, 1-specificity} points as the cutoff value increases
  - Top left corner points reflect wanted performance



# PERFORMANCE METRICS

- Open Excel and RStudio
- Rank Ordering of records for class of interest
  - Based on estimated probabilities of class membership
- Lift curve is used to display the effectiveness of the model in rank ordering of cases
  - Constructed using validation partition scores



# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# PERFORMANCE METRICS Part-3

## LECTURE 18

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# PERFORMANCE METRICS

- Performance Metrics in presence of a class of interest

$$\text{sensitivity} = \frac{n_{11}}{n_{10} + n_{11}} = \text{true positive fraction}$$

$$\text{specificity} = \frac{n_{00}}{n_{00} + n_{01}} = \text{true negative fraction}$$

- ROC (receiver operating characteristic) curve
  - Used to plot {sensitivity, 1-specificity} points as the cutoff value increases
  - Top left corner points reflect wanted performance



# PERFORMANCE METRICS

- Open Excel and RStudio
- Rank Ordering of records for class of interest
  - Based on estimated probabilities of class membership
- Lift curve is used to display the effectiveness of the model in rank ordering of cases
  - Constructed using validation partition scores



# PERFORMANCE METRICS

- Cumulative lift curve or gains chart
  - Used to plot cumulative no. of cases on x-axis and cumulative no. of true positive cases on y-axis
  - Plot displays the lift value of the model for a given no. of cases w.r.t the random selection (probability value of class membership determines the reference line)
- Open Excel and RStudio
- Decile Chart
  - Alternative plot to convey the same information as gains chart



# PERFORMANCE METRICS

- Open RStudio
- Asymmetric Misclassification Costs
  - When misclassification error for a class of interest is more costly than for the other class
  - Example, misclassifying a customer as false positive who is actually likely to respond to the promotional offering
    - Opportunity cost of foregone sale vs. costs of making an offer (profit of ₹20 for a ₹100 item vs. ₹1 scenario)
  - Misclassification rate is not appropriate metric in this case



# PERFORMANCE METRICS

- Asymmetric Misclassification Costs
  - Other considerations
    - Costs of analyzing data
    - Actual net value impact per record
    - New Goal : minimization of costs or maximization of profits
- Open Excel
- How to improve actual classifications by incorporating asymmetric misclassification costs?
  - Change the rules of classification e.g. cutoff value



# PERFORMANCE METRICS

- Performance Metrics based on asymmetric misclassification costs

$$\text{average misclassification cost} = \frac{c_0 n_{0,1} + c_1 n_{1,0}}{n}$$

- Measures average cost of misclassification per observation
- Where  $c_i$  is cost of misclassifying a class  $i$  observation



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE

# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# PERFORMANCE METRICS Part-4

## LECTURE 19

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# PERFORMANCE METRICS

- Open RStudio
- Asymmetric Misclassification Costs
  - When misclassification error for a class of interest is more costly than for the other class
  - Example, misclassifying a customer as false positive who is actually likely to respond to the promotional offering
    - Opportunity cost of foregone sale vs. costs of making an offer (profit of ₹20 for a ₹100 item vs. ₹1 scenario)
  - Misclassification rate is not appropriate metric in this case



# PERFORMANCE METRICS

- Asymmetric Misclassification Costs
  - Other considerations
    - Costs of analyzing data
    - Actual net value impact per record
    - New Goal : minimization of costs or maximization of profits
- Open Excel
- How to improve actual classifications by incorporating asymmetric misclassification costs?
  - Change the rules of classification e.g. cutoff value



# PERFORMANCE METRICS

- Performance Metrics based on asymmetric misclassification costs

$$\text{average misclassification cost} = \frac{c_0 n_{0,1} + c_1 n_{1,0}}{n}$$

- Measures average cost of misclassification per observation
- Where  $c_i$  is cost of misclassifying a class  $i$  observation



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE

# PERFORMANCE METRICS

- Ratio of costs ( $c_0/c_1$ )
- Future misclassification costs
  - Prior Probabilities ( $p_0/p_1$ )
  - $(p_0/p_1)^*$  ( $c_0/c_1$ )
- Lift curve incorporating costs
- Open RStudio
- Lift vs.
  - No. of records or cutoff value?



# PERFORMANCE METRICS

- Asymmetric misclassification costs for m classes ( $m > 2$ )
  - Classification matrix will be ' $m \times m$ '
  - $m$  prior probabilities
  - $m(m-1)$  misclassification costs
  - Matrix for misclassification costs becomes complicated
  - Lift chart not usable for multiclass scenario



# PERFORMANCE METRICS

- Oversampling of rare class members
  - Simple random sampling vs. stratified sampling
- Oversampling approach
  1. Sample more rare class observations (equivalent of oversampling without replacement)
    - Lack of adequate no. of rare class observations
    - Ratio of costs is difficult to determine
  2. Replicate existing rare class observations (equivalent of oversampling with replacement)



# PERFORMANCE METRICS

- Typical solution adopted by analysts
  - Sample equal no. of members from both the classes
- Oversampling adjustment for performance evaluation
  - Score
    1. Validation partition without oversampling
    2. Oversampled validation partition and then remove the oversampling effects by adjusting weights



# PERFORMANCE METRICS

- Typical steps in rare class scenario
  1. Build the candidate models on training partition with 50% class 1 observations and 50% class 0 observations
  2. Validate the models with the validation partition drawn using simple random sample taken from original dataset
- Detailed steps
  1. Separate the class 1 and class 0 observations into two strata (distinct sets)
  2. Half the records from class 1 stratum are randomly selected into training partition



# PERFORMANCE METRICS

- Detailed steps
  3. Remaining class 1 records are reserved for validation partition
  4. Randomly select class 0 records for training partition equal to no. of class 1 records in step 2
  5. Randomly select class 0 records to maintain the original ratio of class 0 to class 1 records for validation partition
  6. For test partition, a random sample can be taken from validation partition



# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# PERFORMANCE METRICS Part-5

## LECTURE 20

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# PERFORMANCE METRICS

- Oversampling of rare class members
  - Simple random sampling vs. stratified sampling
- Oversampling approach
  1. Sample more rare class observations (equivalent of oversampling without replacement)
    - Lack of adequate no. of rare class observations
    - Ratio of costs is difficult to determine
  2. Replicate existing rare class observations (equivalent of oversampling with replacement)



# PERFORMANCE METRICS

- Typical solution adopted by analysts
  - Sample equal no. of members from both the classes
- Oversampling adjustment for performance evaluation
  - Score
    1. Validation partition without oversampling
    2. Oversampled validation partition and then remove the oversampling effects by adjusting weights



# PERFORMANCE METRICS

- Typical steps in rare class scenario
  1. Build the candidate models on training partition with 50% class 1 observations and 50% class 0 observations
  2. Validate the models with the validation partition drawn using simple random sample taken from original dataset
- Detailed steps
  1. Separate the class 1 and class 0 observations into two strata (distinct sets)
  2. Half the records from class 1 stratum are randomly selected into training partition



# PERFORMANCE METRICS

- Detailed steps
  3. Remaining class 1 records are reserved for validation partition
  4. Randomly select class 0 records for training partition equal to no. of class 1 records in step 2
  5. Randomly select class 0 records to maintain the original ratio of class 0 to class 1 records for validation partition
  6. For test partition, a random sample can be taken from validation partition



# PERFORMANCE METRICS

- When ‘Validation partition without oversampling’ is not useful
  - Due to very few class 1 records
  - Second approach of  
‘Using oversampled validation partition for evaluation as well and adjusting the weights to get rid of oversampling effects’  
is taken
    - Adjustment of validation partition classification matrix and lift curve is performed to get reliable accuracy measures



# PERFORMANCE METRICS

- Lift Curve on oversampled validation partition
  - Multiply the net value of a record with proportion of class 1 records in original data
- In a two-class scenario, records which are difficult to classify by the model, can be labeled with a third class option
  - ‘cannot say’
  - Expert judgment can be used for such cases



# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# PREDICTION PERFORMANCE

## LECTURE 21

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# PERFORMANCE METRICS

- Prediction Performance
  - Continuous outcome variable
  - Predictive accuracy vs. goodness of fit
  - E.g., goodness of fit measures in regression modeling
    - $R^2$  and std. err estimate
    - Residual analysis
  - Prediction Error on validation partition
  - Benchmark criterion: average



# PERFORMANCE METRICS

- Prediction Error
  - For a record i, prediction error = actual value - predicted value
  - $e_i = y_i - \hat{y}_i$
- Predictive Accuracy Measures
  - Average Error

$$\frac{1}{n} \sum_{i=1}^n e_i$$

- On average, indicates over or under prediction



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE

# PERFORMANCE METRICS

- Predictive Accuracy Measures
  - Mean Absolute Error (MAE) or Mean Absolute Deviation (MAD)
$$\frac{1}{n} \sum_{i=1}^n |e_i|$$
  - On average, magnitude of error
  - Mean Absolute Percentage Error (MAPE)

$$100\% \times \frac{1}{n} \sum_{i=1}^n |e_i/y_i|$$

- On average, percentage deviation from actual values



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE

# PERFORMANCE METRICS

- Predictive Accuracy Measures
  - Root Mean Squared Error (RMSE)

$$\sqrt{\frac{1}{n} \sum_{i=1}^n e_i^2}$$

- Similar to std. err estimate computed on validation partition
- Measured in same unit as the outcome variable



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE

# PERFORMANCE METRICS

- Predictive Accuracy Measures
  - Total Sum of Squared Errors (Total SSE or SSE)

$$\sum_{i=1}^n e_i^2$$

- These Predictive Accuracy Measures are used to
  - Compare the candidate models
  - Degree of prediction accuracy
  - Outlier issues



# PERFORMANCE METRICS

- Outlier influence in accuracy measures
  - By comparing median based measures and mean based measures
  - Histogram or boxplot of residuals
- Model with high predictive accuracy may or may not be same as
  - model with best fit of data
- Evaluation using visualization techniques
  - Lift curve
    - Relevant when records with highest predicted values are sought



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE

# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# MULTIPLE LINEAR REGRESSION

## LECTURES 22

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# MULTIPLE LINEAR REGRESSION

- Most popular model
- Idea is to fit a linear relationship between
  - A quantitative outcome variable ( $Y$ ) and
  - A set of  $p$  predictors  $\{X_1, X_2, X_3, \dots, X_p\}$
- Assumption: relationship as expressed in the following model equation holds true for the target population

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$$

Where  $\beta_0, \dots, \beta_p$  are coefficients and  $\varepsilon$  is the noise or unexplained part



# MULTIPLE LINEAR REGRESSION

- Objectives:
  - Understanding the relationships between outcome variable and predictors
    - Followed in statistical approach
  - Predicting values of outcome variable for new records
    - Followed in data mining approach
- Applications in data mining
  - Predicting credit card spending, life of an equipment, sales etc.



# MULTIPLE LINEAR REGRESSION

- Model Building and Results Interpretation phases differ depending on the objective:
  - Explanatory (predicting the impact of promotional offer on sales)
  - Predictive (predicting sales)
- Selection of suitable data mining techniques depends on the goal itself



# MULTIPLE LINEAR REGRESSION

## Explanatory Modeling

- Fits the data closely
- Full sample is used to estimate best-fit model
- Performance metrics measure how close model fits the data

## Predictive Modeling

- Predicts new records accurately
- Sample is partitioned into training, validation, and test sets and training partition is used to estimate the model
- Performance metrics measure how well model predicts new observations



# MULTIPLE LINEAR REGRESSION

## Explanatory Modeling

- Model might not have best predictive accuracy
- Statistical techniques with assumed or hypothesized relationships and scarce data (primary data )

## Predictive Modeling

- Model might not be best-fit of data
- Machine learning techniques with no assumed structure and large datasets (secondary data)



# MULTIPLE LINEAR REGRESSION

- Estimates for target population
  - Coefficients:  $\beta_0, \dots, \beta_p$  and
  - $\sigma$ , std. deviation of noise ( $\varepsilon$ )
  - Cannot be measured directly due to unavailability of data on entire population
- Estimation technique:
  - Ordinary least squares (OLS)
    - Computes the sample estimates which minimize the sum of squared deviations between actual values and predicted values



# MULTIPLE LINEAR REGRESSION

- Ordinary least squares (OLS)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p$$

- Unbiased predictions (on average, closer to actual values)
- Smallest average squared error

Given following assumptions hold true

- Noise follows a normal distribution
- Linear relationship holds true
- Observations are independent
- Homoskedasticity: variability in the outcome variable is same irrespective of the values of the predictors



# MULTIPLE LINEAR REGRESSION

- Partitioning in data mining modeling allows relaxation from the first assumption
- In statistical modeling, same sample is used to fit the model and assess its reliability
  - Predictions of new records lack reliability
  - First assumption is required to derive confidence intervals for predictions
- Example: 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# MULTIPLE LINEAR REGRESSION Part-2

## LECTURES 23

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# MULTIPLE LINEAR REGRESSION

- Ordinary least squares (OLS)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p$$

- Unbiased predictions (on average, closer to actual values)
- Smallest average squared error

Given following assumptions hold true

- Noise follows a normal distribution
- Linear relationship holds true
- Observations are independent
- Homoskedasticity: variability in the outcome variable is same irrespective of the values of the predictors



# MULTIPLE LINEAR REGRESSION

- Partitioning in data mining modeling allows relaxation from the first assumption
- In statistical modeling, same sample is used to fit the model and assess its reliability
  - Predictions of new records lack reliability
  - First assumption is required to derive confidence intervals for predictions
- Example: 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# MULTIPLE LINEAR REGRESSION Part-3

## LECTURES 24

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# MULTIPLE LINEAR REGRESSION

- Ordinary least squares (OLS)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p$$

- Unbiased predictions (on average, closer to actual values)
- Smallest average squared error

Given following assumptions hold true

- Noise follows a normal distribution
- Linear relationship holds true
- Observations are independent
- Homoskedasticity: variability in the outcome variable is same irrespective of the values of the predictors



# MULTIPLE LINEAR REGRESSION

- Partitioning in data mining modeling allows relaxation from the first assumption
- In statistical modeling, same sample is used to fit the model and assess its reliability
  - Predictions of new records lack reliability
  - First assumption is required to derive confidence intervals for predictions
- Example: Open RStudio



# MULTIPLE LINEAR REGRESSION

- Variable Selection
  - Availability of large no. of variables for selecting a set of predictors
  - Main idea is to select most useful set of predictors for a given outcome variable of interest
  - Selecting all the variables in the model is not recommended
    - Data collection issues in future
    - Measurement accuracy issues for some variables
    - Missing values
    - Parsimony



# MULTIPLE LINEAR REGRESSION

- Variable Selection
  - Selecting all the variables in the model is not recommended
    - Multicollinearity: two or more predictors sharing the same linear relationship with the outcome variable
    - Sample size issues: Rule of thumb
$$n > 5*(p+2)$$
Where n=no. of observations  
And p=no. of predictors
      - Variance of predictions might increase due to inclusion of predictors which are uncorrelated with the outcome variable
      - Average error of predictions might increase due to exclusion of predictors which are correlated with the outcome variable



# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# MULTIPLE LINEAR REGRESSION Part-4

## LECTURES 25

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# MULTIPLE LINEAR REGRESSION

- Variable Selection
  - Availability of large no. of variables for selecting a set of predictors
  - Main idea is to select most useful set of predictors for a given outcome variable of interest
  - Selecting all the variables in the model is not recommended
    - Data collection issues in future
    - Measurement accuracy issues for some variables
    - Missing values
    - Parsimony



# MULTIPLE LINEAR REGRESSION

- Variable Selection
  - Selecting all the variables in the model is not recommended
    - Multicollinearity: two or more predictors sharing the same linear relationship with the outcome variable
    - Sample size issues: Rule of thumb
$$n > 5*(p+2)$$
Where n=no. of observations  
And p=no. of predictors
      - Variance of predictions might increase due to inclusion of predictors which are uncorrelated with the outcome variable
      - Average error of predictions might increase due to exclusion of predictors which are correlated with the outcome variable



# MULTIPLE LINEAR REGRESSION

- Bias-variance trade-off
  - too few vs. too many predictors
    - Few predictors -> higher bias -> lower variance
  - Drop variables with ‘coefficient < std. dev. of noise’ and with moderate or high correlation with other variables
    - Lower variance
- Steps to reduce the no. of predictors
  - Domain knowledge
  - Practical reasons



# MULTIPLE LINEAR REGRESSION

- Steps to reduce the no. of predictors
  - Summary statistics and graphs
  - Statistical methods using computational power
    - Exhaustive search: all possible combinations
    - Partial-iterative search: algorithm based
- Exhaustive Search
  - Large no. of subsets
  - Criteria to compare models
    - Adjusted R<sup>2</sup>



# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# MULTIPLE LINEAR REGRESSION-PART V

## EXHAUSTIVE SEARCH

### LECTURE 26

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# MULTIPLE LINEAR REGRESSION

- Steps to reduce the no. of predictors
  - Summary statistics and graphs
  - Statistical methods using computational power
    - Exhaustive search: all possible combinations
    - Partial-iterative search: algorithm based
- Exhaustive Search
  - Large no. of subsets
  - Criteria to compare models
    - Adjusted R<sup>2</sup>



# MULTIPLE LINEAR REGRESSION

- Adjusted R<sup>2</sup>

$$R^2_{adj} = 1 - \frac{n-1}{n-p-1} (1 - R^2)$$

Where R<sup>2</sup> is proportion of explained variability in the model

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

- R<sup>2</sup> is called coefficient of determination



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE

# MULTIPLE LINEAR REGRESSION

- $R^2$  would be equal to squared correlation in a single predictor model, that is how  $R^2$  gets its name
- Adjusted  $R^2$  introduces a penalty on the no. of predictors to trade-off between artificial increase vs. amount of information
- High adjusted  $R^2$  values  $\rightarrow$  low  $\hat{\sigma}^2$



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE

# MULTIPLE LINEAR REGRESSION

- Exhaustive Search
  - Criteria to compare models
    - Mallow's  $C_p$
- Mallow's  $C_p$

$$C_p = \frac{SSR}{\hat{\sigma}_f^2} + 2(p + 1) - n$$

Where  $\hat{\sigma}_f^2$  is estimated value of  $\sigma^2$  in the full model

$$\text{and } SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$



# MULTIPLE LINEAR REGRESSION

- Mallow's  $C_p$ 
  - Assumption: full model with all predictors is unbiased
    - Predictors elimination would reduce the variability
  - Best subset model would have  $C_p \sim p+1$  and  $p$  would be a small value
  - Requires high  $n$  value for the training partition relative to  $p$
- Open RStudio



# MULTIPLE LINEAR REGRESSION

- Partial-iterative search
  - Computationally cheaper
  - Best subset is not guaranteed
    - Potential of missing “good” sets of predictors
  - Produce close-to-best subsets
  - Preferred approach for large no. of predictors
  - For moderate no. of predictors, exhaustive search is better
- Trade-off between computation cost vs. potential of finding best subset



# MULTIPLE LINEAR REGRESSION

- Partial-iterative search algorithms
  - Forward selection
    - Add predictors one by one
    - Strength as a single predictor is used
  - Backward elimination
    - Drop predictors one by one
  - Stepwise regression
    - Add predictors one by one and consider dropping insignificant ones
- 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# MULTIPLE LINEAR REGRESSION-PART VI

## Partial Iterative Search

### LECTURE 27

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# MULTIPLE LINEAR REGRESSION

- Partial-iterative search
  - Computationally cheaper
  - Best subset is not guaranteed
    - Potential of missing “good” sets of predictors
  - Produce close-to-best subsets
  - Preferred approach for large no. of predictors
  - For moderate no. of predictors, exhaustive search is better
- Trade-off between computation cost vs. potential of finding best subset



# MULTIPLE LINEAR REGRESSION

- Partial-iterative search algorithms
  - Forward selection
    - Add predictors one by one
    - Strength as a single predictor is used
  - Backward elimination
    - Drop predictors one by one
  - Stepwise regression
    - Add predictors one by one and consider dropping insignificant ones
- 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



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



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# MACHINE LEARNING TECHNIQUE k-NEAREST NEIGHBORS (k-NN)- Part 2

## LECTURE 29

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# 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



# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

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

## LECTURE 30

DR. GAURAV 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  $\equiv$  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



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE

# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# NAÏVE BAYES

## LECTURE 31

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# NAÏVE BAYES

- Complete or Exact Bayes for classification
  - Search for records in training partition having same predictors' values as the new observation to be classified
  - Find the most prevalent class of the outcome variable among the records
  - Assign this class to the new observation
- Class of interest
  - User specified cut off value for the class of interest



# NAÏVE BAYES

- Class of interest
  - Search for records in training partition having same predictors' values as the new observation to be classified
  - Find the probability of a record belonging to the class of interest among the records
  - If computed probability value > cut off value, assign the new observation to the class of interest



# NAÏVE BAYES

- Concept of conditional probability
  - For an outcome variable with m classes  $\{C_1, C_2, \dots, C_m\}$  and p predictors  $\{x_1, x_2, \dots, x_p\}$ , we are interested in the following probability value:

$$P(C_i | x_1, x_2, \dots, x_p) = \frac{P(x_1, x_2, \dots, x_p | C_i)P(C_i)}{P(x_1, x_2, \dots, x_p | C_1)P(C_1) + \dots + P(x_1, x_2, \dots, x_p | C_m)P(C_m)}$$

- Assign the new observation to the class with highest probability value
- Or, if the probability value for the class of interest > cut off value for the same, assign the new observation to the class of interest



# NAÏVE BAYES

- Bayes Model for classification
  - Predictors should also be categorical
  - Numerical variables will have to be converted into categorical variables through binning
- Open Excel
- Complete or Exact Bayes Limitations
  - For a model even with small no. of predictors, many new observations to be classified might not get exact matches
  - Probability of a match might reduce significantly on adding just one variable to the set of predictors



# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# NAÏVE BAYES PART-2

## LECTURE 32

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# NAÏVE BAYES

- Bayes Model for classification
  - Predictors should also be categorical
  - Numerical variables will have to be converted into categorical variables through binning
- Open Excel
- Complete or Exact Bayes Limitations
  - For a model even with small no. of predictors, many new observations to be classified might not get exact matches
  - Probability of a match might reduce significantly on adding just one variable to the set of predictors



# NAÏVE BAYES

- Instead of Complete or Exact Bayes, switch to Naïve Bayes
  - In Naïve Bayes, all the records are used instead of relying on just the matching records
- Naïve Bayes Modification
  - For class  $i$  of outcome variable, compute the probabilities ( $P_1, P_2, \dots, P_p$ ) of belonging to class  $i$  for each predictor's value ( $x_1, x_2, \dots, x_p$ ) taken by the new observation to be classified
  - Compute  $P_1 \times P_2 \times \dots \times P_p \times P(C_i)$
  - Execute previous two steps for all the classes



# NAÏVE BAYES

- Naïve Bayes Modification
  - To compute the probability of the new observation belonging to class i, divide the value computed in step 2 by the summation of values computed in step 2 for all the classes
  - Execute previous step for all the classes
  - Classify the new observation to the class with the highest probability value



# NAÏVE BAYES

- Naïve Bayes formula

$$P(C_i | x_1, x_2, \dots, x_p) = \frac{[P(x_1 | C_i)P(x_2 | C_i) \dots P(x_p | C_i)]P(C_i)}{[P(x_1 | C_1)P(x_2 | C_1) \dots P(x_p | C_1)]P(C_1) + \dots + [P(x_1 | C_m)P(x_2 | C_m) \dots P(x_p | C_m)]P(C_m)}$$

- Naïve Bayes formula is directly derived from the exact Bayes formula after making following assumption:
- Predictors' values  $\{x_1, x_2, \dots, x_p\}$  occur independent of each other for a given class

$$P(x_1, x_2, \dots, x_p | C_i) \equiv P(x_1 | C_i)P(x_2 | C_i) \dots P(x_p | C_i)$$



# NAÏVE BAYES

- Naïve Bayes formula
  - For classification, naïve Bayes formula works quite well
  - Since we don't require probabilities values to be accurate in absolute term, rather just a reasonably accurate rank ordering of these values
  - For the same reason, we should use the numerator only and drop the denominator which is common for all the classes
- Steps when we have a class of interest
  - User specified cut off value for the class of interest



# NAÏVE BAYES

- Steps when we have a class of interest
  - Compute the probabilities ( $P_1, P_2, \dots, P_p$ ) of belonging to class of interest for each predictor's value ( $x_1, x_2, \dots, x_p$ ) taken by the new observation to be classified
  - Compute  $P_1 \times P_2 \times \dots \times P_p \times P(\text{Class of interest})$
  - Execute previous two steps for all the classes
  - To compute the probability of the new observation belonging to class of interest, divide the value computed in step 2 by the summation of values computed in step 2 for all the classes



# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# NAÏVE BAYES PART-3

## LECTURE 33

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# NAÏVE BAYES

- Bayes Model for classification
  - Predictors should also be categorical
  - Numerical variables will have to be converted into categorical variables through binning
- Open Excel
- Complete or Exact Bayes Limitations
  - For a model even with small no. of predictors, many new observations to be classified might not get exact matches
  - Probability of a match might reduce significantly on adding just one variable to the set of predictors



# NAÏVE BAYES

- Instead of Complete or Exact Bayes, switch to Naïve Bayes
  - In Naïve Bayes, all the records are used instead of relying on just the matching records
- Naïve Bayes Modification
  - For class  $i$  of outcome variable, compute the probabilities ( $P_1, P_2, \dots, P_p$ ) of belonging to class  $i$  for each predictor's value ( $x_1, x_2, \dots, x_p$ ) taken by the new observation to be classified
  - Compute  $P_1 \times P_2 \times \dots \times P_p \times P(C_i)$
  - Execute previous two steps for all the classes



# NAÏVE BAYES

- Naïve Bayes Modification
  - To compute the probability of the new observation belonging to class i, divide the value computed in step 2 by the summation of values computed in step 2 for all the classes
  - Execute previous step for all the classes
  - Classify the new observation to the class with the highest probability value



# NAÏVE BAYES

- Naïve Bayes formula

$$P(C_i | x_1, x_2, \dots, x_p) = \frac{[P(x_1 | C_i)P(x_2 | C_i) \dots P(x_p | C_i)]P(C_i)}{[P(x_1 | C_1)P(x_2 | C_1) \dots P(x_p | C_1)]P(C_1) + \dots + [P(x_1 | C_m)P(x_2 | C_m) \dots P(x_p | C_m)]P(C_m)}$$

- Naïve Bayes formula is directly derived from the exact Bayes formula after making following assumption:
- Predictors' values  $\{x_1, x_2, \dots, x_p\}$  occur independent of each other for a given class

$$P(x_1, x_2, \dots, x_p | C_i) \equiv P(x_1 | C_i)P(x_2 | C_i) \dots P(x_p | C_i)$$



# NAÏVE BAYES

- Naïve Bayes formula
  - For classification, naïve Bayes formula works quite well
  - Since we don't require probabilities values to be accurate in absolute term, rather just a reasonably accurate rank ordering of these values
  - For the same reason, we should use the numerator only and drop the denominator which is common for all the classes
- Steps when we have a class of interest
  - User specified cut off value for the class of interest



# NAÏVE BAYES

- Steps when we have a class of interest
  - Compute the probabilities ( $P_1, P_2, \dots, P_p$ ) of belonging to class of interest for each predictor's value ( $x_1, x_2, \dots, x_p$ ) taken by the new observation to be classified
  - Compute  $P_1 \times P_2 \times \dots \times P_p \times P(\text{Class of interest})$
  - Execute previous two steps for all the classes
  - To compute the probability of the new observation belonging to class of interest, divide the value computed in step 2 by the summation of values computed in step 2 for all the classes



# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# NAÏVE BAYES PART-4

## LECTURE 34

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# NAÏVE BAYES

- Naïve Bayes formula
  - For classification, naïve Bayes formula works quite well
  - Since we don't require probabilities values to be accurate in absolute term, rather just a reasonably accurate rank ordering of these values
  - For the same reason, we should use the numerator only and drop the denominator which is common for all the classes
- Steps when we have a class of interest
  - User specified cut off value for the class of interest



# NAÏVE BAYES

- Steps when we have a class of interest
  - Compute the probabilities ( $P_1, P_2, \dots, P_p$ ) of belonging to class of interest for each predictor's value ( $x_1, x_2, \dots, x_p$ ) taken by the new observation to be classified
  - Compute  $P_1 \times P_2 \times \dots \times P_p \times P(\text{Class of interest})$
  - Execute previous two steps for all the classes
  - To compute the probability of the new observation belonging to class of interest, divide the value computed in step 2 by the summation of values computed in step 2 for all the classes



# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# NAÏVE BAYES PART-5

## LECTURE 35

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# NAÏVE BAYES

- Further Comments on Naïve Bayes
  - Good performance despite assumption of independent predictors' values being far from true
  - Requires large no. of records
  - Few classes of predictors might not be represented in the training partition records
    - Zero probability is assumed
  - Good for classification but not for estimating probabilities of class membership



# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# CLASSIFICATION & REGRESSION TREES

## LECTURE 36

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# CLASSIFICATION & REGRESSION TREES

- CART
  - A data-driven method
    - Based on separating observations into homogeneous subgroups by creating splits on predictors
  - Used for both prediction and classification tasks
  - Model is represented by a tree diagram
    - Easy to interpret logical rules
    - CART algorithm grows binary trees
  - Adoption across domains



# CLASSIFICATION & REGRESSION TREES

- Classification Trees
  - Recursive partitioning
    - About partitioning  $p$ -dimensional space of predictors using training partition, where  $p$  is no. of predictors
  - Pruning
    - About pruning the built tree using validation data



# CLASSIFICATION & REGRESSION TREES

- Recursive Partitioning
  - Partitioning p-dimensional space of predictors into non-overlapping multi-dimensional rectangles
  - The partitioning process is recursive in nature
    - Applied on the results of previous partitions
- Steps for Recursive Partitioning
  - An optimal combination of one of the predictors,  $x_i$  and its value  $v_i$  is selected to create first split of p-dimensional space into two parts
    - Part I:  $x_i \leq v_i$
    - Part II:  $x_i > v_i$



# CLASSIFICATION & REGRESSION TREES

- Steps for Recursive Partitioning
  - Step 1 is applied again on the two parts and process continues to create more rectangular parts
  - The partitioning process continues till we reach pure homogeneous parts
    - All the observations in the part belong to just one of the classes
- 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# CLASSIFICATION & REGRESSION TREES PART-2

## LECTURE 37

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# CLASSIFICATION & REGRESSION TREES

- Impurity Measures

- Gini index and Entropy measure

- Gini Index

For an outcome variable with  $m$  classes, Gini impurity index for a rectangular part is defined as

$$gini = 1 - \sum_{k=1}^m P_k^2$$

Where  $P_k$  is the proportion of rectangular part observations belonging to class  $k$



# CLASSIFICATION & REGRESSION TREES

- Gini Index
  - Gini values lie in the range  $\{0, (m-1)/m\}$  for m-class scenario and  $\{0, 0.5\}$  for two-class scenario
- Entropy Measure

For an outcome variable with m classes, entropy for a rectangular part is defined as

$$\text{Entropy} = - \sum_{k=1}^m P_k \log_2(P_k)$$



# CLASSIFICATION & REGRESSION TREES

- Entropy Measure
  - Entropy values lie in the range  $\{0, \log_2(m)\}$  for m-class scenario and  $\{0, 1\}$  for two-class scenario
- Open RStudio
- Tree diagram or tree structure
  - Each split of p-dimensional space into two parts can be depicted as a split of a node in a decision tree into two child nodes
  - First split creates branches of root node



# CLASSIFICATION & REGRESSION TREES

- Two types of nodes in tree structure
  - Decision node: Depicted with a circle
  - Terminal or leaf node: Depicted with a rectangle
    - Correspond to Final rectangular parts
- Steps to classify a new observation using tree based models
  - New observation to be classified is dropped down the tree starting from root node
  - At each decision node, the appropriate branch is taken until we reach a leaf node



# CLASSIFICATION & REGRESSION TREES

- Steps to classify a new observation using tree based models
  - At leaf node, majority class is assigned to the new observation
    - For class of interest scenario, proportion of records belonging to the class of interest is compared with the user specified cut off value for the same
- 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# CLASSIFICATION & REGRESSION TREES PART-3

## LECTURE 38

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# CLASSIFICATION & REGRESSION TREES

- Two types of nodes in tree structure
  - Decision node: Depicted with a circle
  - Terminal or leaf node: Depicted with a rectangle
    - Correspond to Final rectangular parts
- Steps to classify a new observation using tree based models
  - New observation to be classified is dropped down the tree starting from root node
  - At each decision node, the appropriate branch is taken until we reach a leaf node



# CLASSIFICATION & REGRESSION TREES

- Steps to classify a new observation using tree based models
  - At leaf node, majority class is assigned to the new observation
    - For class of interest scenario, proportion of records belonging to the class of interest is compared with the user specified cut off value for the same
- Open RStudio



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE

# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# CLASSIFICATION & REGRESSION TREES PART-4

## LECTURE 39

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# CLASSIFICATION & REGRESSION TREES

- CART
  - A data-driven method
    - Based on separating observations into homogeneous subgroups by creating splits on predictors
  - Used for both prediction and classification tasks
  - Model is represented by a tree diagram
    - Easy to interpret logical rules
    - CART algorithm grows binary trees
  - Adoption across domains



# CLASSIFICATION & REGRESSION TREES

- Classification Trees
  - Recursive partitioning
    - About partitioning  $p$ -dimensional space of predictors using training partition, where  $p$  is no. of predictors
  - Pruning
    - About pruning the built tree using validation data



# CLASSIFICATION & REGRESSION TREES

- CART example has been discussed in the lecture video



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE

# CLASSIFICATION & REGRESSION TREES

- Further Comments on CART
  - Can be used as a variable selection approach
  - No variable transformation is required
  - Robust to outliers
  - Non-linear and non-parametric technique
  - Handle missing values
  - Sensitive to sample data changes
  - Predictor's strength as a single variable is modeled and not as part of a group of predictors



# CLASSIFICATION & REGRESSION TREES

- Further Comments on CART
  - Might not fit linear structures or relationships between predictors
    - New predictors based on hypothesized relationships can be used
  - Require a large dataset
  - High computation time



# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# CLASSIFICATION & REGRESSION TREES PART-5

## LECTURE 40

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# CLASSIFICATION & REGRESSION TREES

- CART example has been discussed in the lecture video



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE

# CLASSIFICATION & REGRESSION TREES

- Further Comments on CART
  - Can be used as a variable selection approach
  - No variable transformation is required
  - Robust to outliers
  - Non-linear and non-parametric technique
  - Handle missing values
  - Sensitive to sample data changes
  - Predictor's strength as a single variable is modeled and not as part of a group of predictors



# CLASSIFICATION & REGRESSION TREES

- Further Comments on CART
  - Might not fit linear structures or relationships between predictors
    - New predictors based on hypothesized relationships can be used
  - Require a large dataset
  - High computation time



# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# CLASSIFICATION & REGRESSION TREES PART-6

## LECTURE 41

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# CLASSIFICATION & REGRESSION TREES

- Steps to classify a new observation using tree based models
  - At leaf node, majority class is assigned to the new observation
    - For class of interest scenario, proportion of records belonging to the class of interest is compared with the user specified cut off value for the same
- Open RStudio



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE

# CLASSIFICATION & REGRESSION TREES

- Pruning
  - Avoid overfitting
    - Full grown tree leads to complete overfitting of data
    - Poor performance on new data
  - Overall error of tree models
    - Expected to decrease until the point where relationships between outcome variable and predictors are fitted
    - Then tree models start fitting to the noise and overall error starts increasing
      - Due to splits involving small number of observations



# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# Pruning Process

## LECTURE 42

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# CLASSIFICATION & REGRESSION TREES

- Pruning
  - Avoid overfitting
    - Full grown tree leads to complete overfitting of data
    - Poor performance on new data
  - Overall error of tree models
    - Expected to decrease until the point where relationships between outcome variable and predictors are fitted
    - Then tree models start fitting to the noise and overall error starts increasing
      - Due to splits involving small number of observations



# CLASSIFICATION & REGRESSION TREES

- Pruning
  - Stop tree growth before it starts overfitting data or fitting noise
    - No. of splits or tree depth level
    - No. of observations in a node to attempt the split
    - Accepted level of reduction in impurity
    - Difficulties in determining the stopping point for such rules
  - Prune the full grown tree back to a level where it doesn't overfit data or fit noise
    - Use validation partition to prune the tree modeled with training partition
    - Idea is to remove the tree branches which don't reduce the error rate further



# CLASSIFICATION & REGRESSION TREES

- Pruning
  - Prune the full grown tree back to a level where it doesn't overfit data or fit noise
    - Find the point where error rate on validation partition starts to increase
    - Cost complexity parameter or complexity parameter (CP) in CART algorithm
$$CP = Err + PF * TL$$
Where Err is misclassification error, PF is penalty factor for tree length (TL)
  - Minimum error tree
    - Tree with minimum misclassification error on validation partition



# CLASSIFICATION & REGRESSION TREES

- Pruning
  - Best pruned tree
    - Adjustment for sampling error on minimum error tree
    - Smallest tree in the pruning sequence which lies within one std. err. (of error rate) of minimum error tree
- Open RStudio
- Classification Rules
  - Each terminal node in a tree model is equivalent to a classification rule
  - Simplify and remove redundant rules



# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# Pruning Process Part-2

## LECTURE-43

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# CLASSIFICATION & REGRESSION TREES

- Pruning
  - Avoid overfitting
    - Full grown tree leads to complete overfitting of data
    - Poor performance on new data
  - Overall error of tree models
    - Expected to decrease until the point where relationships between outcome variable and predictors are fitted
    - Then tree models start fitting to the noise and overall error starts increasing
      - Due to splits involving small number of observations



# CLASSIFICATION & REGRESSION TREES

- Pruning
  - Stop tree growth before it starts overfitting data or fitting noise
    - No. of splits or tree depth level
    - No. of observations in a node to attempt the split
    - Accepted level of reduction in impurity
    - Difficulties in determining the stopping point for such rules
  - Prune the full grown tree back to a level where it doesn't overfit data or fit noise
    - Use validation partition to prune the tree modeled with training partition
    - Idea is to remove the tree branches which don't reduce the error rate further



# CLASSIFICATION & REGRESSION TREES

- Pruning
  - Prune the full grown tree back to a level where it doesn't overfit data or fit noise
    - Find the point where error rate on validation partition starts to increase
    - Cost complexity parameter or complexity parameter (CP) in CART algorithm
$$CP = Err + PF * TL$$
Where Err is misclassification error, PF is penalty factor for tree length (TL)
  - Minimum error tree
    - Tree with minimum misclassification error on validation partition



# CLASSIFICATION & REGRESSION TREES

- Pruning
  - Best pruned tree
    - Adjustment for sampling error on minimum error tree
    - Smallest tree in the pruning sequence which lies within one std. err. (of error rate) of minimum error tree
- Open RStudio
- Classification Rules
  - Each terminal node in a tree model is equivalent to a classification rule
  - Simplify and remove redundant rules



# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# Pruning Process Part-3

## LECTURE-44

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# CLASSIFICATION & REGRESSION TREES

- Pruning
  - Avoid overfitting
    - Full grown tree leads to complete overfitting of data
    - Poor performance on new data
  - Overall error of tree models
    - Expected to decrease until the point where relationships between outcome variable and predictors are fitted
    - Then tree models start fitting to the noise and overall error starts increasing
      - Due to splits involving small number of observations



# CLASSIFICATION & REGRESSION TREES

- Pruning
  - Stop tree growth before it starts overfitting data or fitting noise
    - No. of splits or tree depth level
    - No. of observations in a node to attempt the split
    - Accepted level of reduction in impurity
    - Difficulties in determining the stopping point for such rules
  - Prune the full grown tree back to a level where it doesn't overfit data or fit noise
    - Use validation partition to prune the tree modeled with training partition
    - Idea is to remove the tree branches which don't reduce the error rate further



# CLASSIFICATION & REGRESSION TREES

- Pruning
  - Prune the full grown tree back to a level where it doesn't overfit data or fit noise
    - Find the point where error rate on validation partition starts to increase
    - Cost complexity parameter or complexity parameter (CP) in CART algorithm
$$CP = Err + PF * TL$$
Where Err is misclassification error, PF is penalty factor for tree length (TL)
  - Minimum error tree
    - Tree with minimum misclassification error on validation partition



# CLASSIFICATION & REGRESSION TREES

- Pruning
  - Best pruned tree
    - Adjustment for sampling error on minimum error tree
    - Smallest tree in the pruning sequence which lies within one std. err. (of error rate) of minimum error tree
- Open RStudio
- Classification Rules
  - Each terminal node in a tree model is equivalent to a classification rule
  - Simplify and remove redundant rules



# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# REGRESSION TREES

## LECTURE 45

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# CLASSIFICATION & REGRESSION TREES

- Regression Trees
  - Outcome variable should be numerical
  - Steps to build tree model are similar to that of classification trees
  - Prediction step, impurity measures and performance metrics are different
- Prediction step
  - Value of a leaf node is predicted value for a new observation that fell in that leaf node
  - Value of a leaf node is computed by taking average of training partition records constituting that leaf node



# CLASSIFICATION & REGRESSION TREES

- Impurity Measures
  - Sum of squared deviations from mean of leaf node
    - Equivalent to squared errors since mean value of leaf node is predicted value
  - Lowest impurity is zero when all the observations that fell in a leaf node have same actual value of outcome variable



# CLASSIFICATION & REGRESSION TREES

- Further Comments on CART
  - Can be used as a variable selection approach
  - No variable transformation is required
  - Robust to outliers
  - Non-linear and non-parametric technique
  - Handle missing values
  - Sensitive to sample data changes
  - Predictor's strength as a single variable is modeled and not as part of a group of predictors



# CLASSIFICATION & REGRESSION TREES

- Further Comments on CART
  - Might not fit linear structures or relationships between predictors
    - New predictors based on hypothesized relationships can be used
  - Require a large dataset
  - High computation time



# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# LOGISTIC REGRESSION

## LECTURE 46

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# LOGISTIC REGRESSION

- Equivalent of linear regression for categorical outcome variable
  - Predictors can be categorical or continuous
- Applied in following tasks
  - Classification task
    - Predicting the class of a new observation
  - Profiling
    - Understanding similarities and differences among groups



# LOGISTIC REGRESSION

- Steps for logistic regression
  - Estimate probabilities of class memberships
  - Classify observations using probabilities values
    - Most probable class method: assign the observation to the class with highest probability value
      - Equivalently, for a two-class case, cutoff value of 0.5 can be used
    - Class of interest: user specified cutoff value
      - For a two-class case, typically a value greater than average probability value for class of interest, but less than 0.5 can be used



# LOGISTIC REGRESSION

- Logistic Regression Model
  - Used typically in cases when structured model is preferred over data-driven models for classification tasks
  - Categorical outcome variable cannot be directly modeled as a linear function of predictors
    - Inability to apply various mathematical operators
    - Variable type mismatches
    - Range reasonability issues
      - LHS range={0, ..., m-1}
      - RHS range=(-∞, ∞)



# LOGISTIC REGRESSION

- Logistic Regression Model
    - Instead of using outcome variable ( $Y$ ) in the model, a function of  $Y$ , called *logit* is used
  - Logit
    - Think about modeling probability value as a linear function of predictors, specifically in a two-class case
- If  $P$  is the probability of class 1 membership

$$P = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Where  $p$  is the no. of predictors



# LOGISTIC REGRESSION

- Logit
  - LHS range improves from {0, 1} to [0, 1], however still cannot match RHS
  - Can we bring RHS range to [0,1]?
    - Nonlinear approach
  - Typically, a nonlinear function of the following form is used to perform the required transformation

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}}$$

This function is called *logistic response function*



# LOGISTIC REGRESSION

- Logit
  - Rearrange the previous equation as below:

$$\frac{P}{1 - P} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}$$

LHS is expression for *odds*, another measure of class membership

$$odds = \frac{P}{1 - P}$$

- Odds of belonging to a class is defined as ratio of probability of class 1 membership to probability of class 0 membership
  - This metric is popular in sports, horse racing, gambling, and many other areas



# LOGISTIC REGRESSION

- Logit

- Previous equation can be rewritten as

$$odds = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}$$

- Range is now  $(0, \infty)$
  - Take log on both sides of previous equation

$$\log(odds) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

- Standard logistic model
  - Now, LHS and RHS both have same range  $(-\infty, \infty)$
- Log(odds) is called logit
  - Logit is used as the outcome variable in the model instead of categorical Y



# LOGISTIC REGRESSION

- Odds and logit can be written as a function of probability of class 1 membership
  - Open RStudio
- In logistic regression model, we predict the logit values and therefore corresponding probability of a categorical outcome
  - Predicted probabilities values become the basis for classification
  - A prediction model for classification task



# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# LOGISTIC REGRESSION- PART 2

## LECTURE 47

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# LOGISTIC REGRESSION

- Logit
  - LHS range improves from {0, 1} to [0, 1], however still cannot match RHS
  - Can we bring RHS range to [0,1]?
    - Nonlinear approach
  - Typically, a nonlinear function of the following form is used to perform the required transformation

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}}$$

This function is called *logistic response function*



# LOGISTIC REGRESSION

- Logit
  - Rearrange the previous equation as below:

$$\frac{P}{1 - P} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}$$

LHS is expression for *odds*, another measure of class membership

$$odds = \frac{P}{1 - P}$$

- Odds of belonging to a class is defined as ratio of probability of class 1 membership to probability of class 0 membership
  - This metric is popular in sports, horse racing, gambling, and many other areas



# LOGISTIC REGRESSION

- Logit

- Previous equation can be rewritten as

$$odds = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}$$

- Range is now  $(0, \infty)$
  - Take log on both sides of previous equation

$$\log(odds) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

- Standard logistic model
  - Now, LHS and RHS both have same range  $(-\infty, \infty)$
- Log(odds) is called logit
  - Logit is used as the outcome variable in the model instead of categorical Y



# LOGISTIC REGRESSION

- Odds and logit can be written as a function of probability of class 1 membership
  - Open RStudio
- In logistic regression model, we predict the logit values and therefore corresponding probability of a categorical outcome
  - Predicted probabilities values become the basis for classification
  - A prediction model for classification task



# LOGISTIC REGRESSION

- Estimation Technique
  - Least squares method used in multiple linear regression cannot be used
    - Non-linear formulation of logistic regression
  - Maximum likelihood method is used
    - Estimates are optimized in order to maximize the likelihood of obtaining the observations used in training the model
    - Less robust than estimation techniques used in linear regression
    - Reliability of estimates
      - Outcome variable categories should have adequate proportion
      - Adequate sample size w.r.t no. of estimates



# LOGISTIC REGRESSION

- Estimation Technique
  - Maximum likelihood method is used
    - Collinearity issues similar to linear regression
- Interpretation of Results
  - Logit model
    - Additive factor ( $\beta$ )
      - If  $\beta < 0$ , increase in  $x \Rightarrow$  decrease in logit values
      - If  $\beta > 0$ , increase in  $x \Rightarrow$  increase in logit values
    - For any value of  $x$ , interpretative statements of results are same



# Thanks...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# LOGISTIC REGRESSION PART-3

## LECTURE 48

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# LOGISTIC REGRESSION

- Odds and odds ratios
  - Odds is a ratio of two probability values (prob. of class 1/prob. Of Class 0)
  - Odds ratio is ratio of two odds (odds of class m1/odds of class m2)
    - Odds ratio  $> 1 \Rightarrow$  odds of class m1 are higher than class m2
- 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# LOGISTIC REGRESSION PART-4

## LECTURE 49

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# LOGISTIC REGRESSION

- Odds and logit can be written as a function of probability of class 1 membership
  - Open RStudio
- In logistic regression model, we predict the logit values and therefore corresponding probability of a categorical outcome
  - Predicted probabilities values become the basis for classification
  - A prediction model for classification task

# LOGISTIC REGRESSION

- Estimation Technique
  - Least squares method used in multiple linear regression cannot be used
    - Non-linear formulation of logistic regression
  - Maximum likelihood method is used
    - Estimates are optimized in order to maximize the likelihood of obtaining the observations used in training the model
    - Less robust than estimation techniques used in linear regression
    - Reliability of estimates
      - Outcome variable categories should have adequate proportion
      - Adequate sample size w.r.t no. of estimates



# LOGISTIC REGRESSION

- Estimation Technique
  - Maximum likelihood method is used
    - Collinearity issues similar to linear regression
- Interpretation of Results
  - Logit model
    - Additive factor ( $\beta$ )
      - If  $\beta < 0$ , increase in  $x \Rightarrow$  decrease in logit values
      - If  $\beta > 0$ , increase in  $x \Rightarrow$  increase in logit values
    - For any value of  $x$ , interpretative statements of results are same



# LOGISTIC REGRESSION

- Interpretation of Results
  - Odds model
    - Multiplicative factor ( $e^{\beta}$ )
      - If  $\beta < 0$ , increase in  $x \Rightarrow$  decrease in odds
      - If  $\beta > 0$ , increase in  $x \Rightarrow$  increase in odds
    - For any value of  $x$ , interpretative statements of results are same
  - Probability model
    - For a unit increase in a particular predictor, corresponding change in the probability value is not a constant, while holding all other predictors constant
      - Depends on the specific values of the predictor
    - Interpretative statements of results depend on specific values of  $x$



# LOGISTIC REGRESSION

- Odds and odds ratios
  - Odds is a ratio of two probability values (prob. of class 1/prob. Of Class 0)
  - Odds ratio is ratio of two odds (odds of class m1/odds of class m2)
    - Odds ratio  $> 1 \Rightarrow$  odds of class m1 are higher than class m2
- Open RStudio



# LOGISTIC REGRESSION

- Linear Regression for a categorical outcome variable?
  - Can be done by treating the outcome variable as continuous and coding it numerically
  - However, anomalies will lead to spurious modeling
    - Predictions can take any value, not just dummy values {0,1}
    - Outcome variable or residuals don't follow normal distribution
      - binomial distribution
    - Variance of outcome variable is not constant across all records (violation of homoscedasticity)
      - $np(1-p)$



# LOGISTIC REGRESSION

- Logistic Regression for Profiling Task
  - Apart from model performance on validation partition
  - Model's fit to data is assessed on training partition
    - However, still avoid overfitting
    - Usefulness of predictors is examined
  - Goodness of fit metrics
    - Overall fit of the model
      - Deviance (equivalent to SSE in linear regression)
      - $1 - \text{Deviance}/\text{Null Deviance}$  (equivalent to multiple  $R^2$  in linear regression)
    - Single predictors



# LOGISTIC REGRESSION

- Outcome variable with  $m$  classes ( $m > 2$ )
  - Multinomial logistic regression
    - Separate binary logistic regression model for  $m-1$  classes (one class is treated as reference class)
  - Ordinal logistic regression
    - Large no. of ordinal classes: treat ordinal variable as continuous variable and apply multiple linear regression



# LOGISTIC REGRESSION

- Outcome variable with m classes (m>2)
  - Ordinal logistic regression
    - Small no. of ordinal classes: Proportional odds or cumulative logit method
      - Separate binary logistic regression model for m-1 cumulative probabilities
- For a three class case: C1, C2, and C3 and a single predictor x1
$$\text{logit}(C1) = \alpha_0 + \beta_1 x_1$$
$$\text{logit}(C1 \text{or } C2) = \beta_0 + \beta_1 x_1$$
- 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# LOGISTIC REGRESSION PART-5

## LECTURE 50

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# LOGISTIC REGRESSION

- Interpretation of Results
  - Odds model
    - Multiplicative factor ( $e^{\beta}$ )
      - If  $\beta < 0$ , increase in  $x \Rightarrow$  decrease in odds
      - If  $\beta > 0$ , increase in  $x \Rightarrow$  increase in odds
    - For any value of  $x$ , interpretative statements of results are same
  - Probability model
    - For a unit increase in a particular predictor, corresponding change in the probability value is not a constant, while holding all other predictors constant
      - Depends on the specific values of the predictor
    - Interpretative statements of results depend on specific values of  $x$



# LOGISTIC REGRESSION

- Odds and odds ratios
  - Odds is a ratio of two probability values (prob. of class 1/prob. Of Class 0)
  - Odds ratio is ratio of two odds (odds of class m1/odds of class m2)
    - Odds ratio  $> 1 \Rightarrow$  odds of class m1 are higher than class m2
- Open RStudio



# LOGISTIC REGRESSION

- Linear Regression for a categorical outcome variable?
  - Can be done by treating the outcome variable as continuous and coding it numerically
  - However, anomalies will lead to spurious modeling
    - Predictions can take any value, not just dummy values {0,1}
    - Outcome variable or residuals don't follow normal distribution
      - binomial distribution
    - Variance of outcome variable is not constant across all records (violation of homoscedasticity)
      - $np(1-p)$



# LOGISTIC REGRESSION

- Logistic Regression for Profiling Task
  - Apart from model performance on validation partition
  - Model's fit to data is assessed on training partition
    - However, still avoid overfitting
    - Usefulness of predictors is examined
  - Goodness of fit metrics
    - Overall fit of the model
      - Deviance (equivalent to SSE in linear regression)
      - $1 - \text{Deviance}/\text{Null Deviance}$  (equivalent to multiple  $R^2$  in linear regression)
    - Single predictors



# LOGISTIC REGRESSION

- Outcome variable with  $m$  classes ( $m > 2$ )
  - Multinomial logistic regression
    - Separate binary logistic regression model for  $m-1$  classes (one class is treated as reference class)
  - Ordinal logistic regression
    - Large no. of ordinal classes: treat ordinal variable as continuous variable and apply multiple linear regression



# LOGISTIC REGRESSION

- Outcome variable with m classes (m>2)
  - Ordinal logistic regression
    - Small no. of ordinal classes: Proportional odds or cumulative logit method
      - Separate binary logistic regression model for m-1 cumulative probabilities
- For a three class case: C1, C2, and C3 and a single predictor x1
$$\text{logit}(C1) = \alpha_0 + \beta_1 x_1$$
$$\text{logit}(C1 \text{or } C2) = \beta_0 + \beta_1 x_1$$
- 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# LOGISTIC REGRESSION PART-6

## LECTURE 51

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# LOGISTIC REGRESSION

- Equivalent of linear regression for categorical outcome variable
  - Predictors can be categorical or continuous
- Applied in following tasks
  - Classification task
    - Predicting the class of a new observation
  - Profiling
    - Understanding similarities and differences among groups



# LOGISTIC REGRESSION

- Steps for logistic regression
  - Estimate probabilities of class memberships
  - Classify observations using probabilities values
    - Most probable class method: assign the observation to the class with highest probability value
      - Equivalently, for a two-class case, cutoff value of 0.5 can be used
    - Class of interest: user specified cutoff value
      - For a two-class case, typically a value greater than average probability value for class of interest, but less than 0.5 can be used



# LOGISTIC REGRESSION

- Logistic Regression Model
  - Used typically in cases when structured model is preferred over data-driven models for classification tasks
  - Categorical outcome variable cannot be directly modeled as a linear function of predictors
    - Inability to apply various mathematical operators
    - Variable type mismatches
    - Range reasonability issues
      - LHS range={0, ..., m-1}
      - RHS range=(-∞, ∞)



# LOGISTIC REGRESSION

- Logistic Regression Model
    - Instead of using outcome variable ( $Y$ ) in the model, a function of  $Y$ , called *logit* is used
  - Logit
    - Think about modeling probability value as a linear function of predictors, specifically in a two-class case
- If  $P$  is the probability of class 1 membership

$$P = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Where  $p$  is the no. of predictors



# LOGISTIC REGRESSION

- Logit
  - LHS range improves from {0, 1} to [0, 1], however still cannot match RHS
  - Can we bring RHS range to [0,1]?
    - Nonlinear approach
  - Typically, a nonlinear function of the following form is used to perform the required transformation

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}}$$

This function is called *logistic response function*



# LOGISTIC REGRESSION

- Logit
  - Rearrange the previous equation as below:

$$\frac{P}{1 - P} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}$$

LHS is expression for *odds*, another measure of class membership

$$odds = \frac{P}{1 - P}$$

- Odds of belonging to a class is defined as ratio of probability of class 1 membership to probability of class 0 membership
  - This metric is popular in sports, horse racing, gambling, and many other areas



# LOGISTIC REGRESSION

- Logit

- Previous equation can be rewritten as

$$odds = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}$$

- Range is now  $(0, \infty)$
  - Take log on both sides of previous equation

$$\log(odds) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

- Standard logistic model
  - Now, LHS and RHS both have same range  $(-\infty, \infty)$
- Log(odds) is called logit
  - Logit is used as the outcome variable in the model instead of categorical Y



# LOGISTIC REGRESSION

- Odds and logit can be written as a function of probability of class 1 membership
  - Open RStudio
- In logistic regression model, we predict the logit values and therefore corresponding probability of a categorical outcome
  - Predicted probabilities values become the basis for classification
  - A prediction model for classification task



# LOGISTIC REGRESSION

- Estimation Technique
  - Least squares method used in multiple linear regression cannot be used
    - Non-linear formulation of logistic regression
  - Maximum likelihood method is used
    - Estimates are optimized in order to maximize the likelihood of obtaining the observations used in training the model
    - Less robust than estimation techniques used in linear regression
    - Reliability of estimates
      - Outcome variable categories should have adequate proportion
      - Adequate sample size w.r.t no. of estimates



# LOGISTIC REGRESSION

- Estimation Technique
  - Maximum likelihood method is used
    - Collinearity issues similar to linear regression
- Interpretation of Results
  - Logit model
    - Additive factor ( $\beta$ )
      - If  $\beta < 0$ , increase in  $x \Rightarrow$  decrease in logit values
      - If  $\beta > 0$ , increase in  $x \Rightarrow$  increase in logit values
    - For any value of  $x$ , interpretative statements of results are same



# LOGISTIC REGRESSION

- Interpretation of Results
  - Odds model
    - Multiplicative factor ( $e^{\beta}$ )
      - If  $\beta < 0$ , increase in  $x \Rightarrow$  decrease in odds
      - If  $\beta > 0$ , increase in  $x \Rightarrow$  increase in odds
    - For any value of  $x$ , interpretative statements of results are same
  - Probability model
    - For a unit increase in a particular predictor, corresponding change in the probability value is not a constant, while holding all other predictors constant
      - Depends on the specific values of the predictor
    - Interpretative statements of results depend on specific values of  $x$



# LOGISTIC REGRESSION

- Odds and odds ratios
  - Odds is a ratio of two probability values (prob. of class 1/prob. Of Class 0)
  - Odds ratio is ratio of two odds (odds of class m1/odds of class m2)
    - Odds ratio  $> 1 \Rightarrow$  odds of class m1 are higher than class m2
- Open RStudio



# LOGISTIC REGRESSION

- Linear Regression for a categorical outcome variable?
  - Can be done by treating the outcome variable as continuous and coding it numerically
  - However, anomalies will lead to spurious modeling
    - Predictions can take any value, not just dummy values {0,1}
    - Outcome variable or residuals don't follow normal distribution
      - binomial distribution
    - Variance of outcome variable is not constant across all records (violation of homoscedasticity)
      - $np(1-p)$



# LOGISTIC REGRESSION

- Logistic Regression for Profiling Task
  - Apart from model performance on validation partition
  - Model's fit to data is assessed on training partition
    - However, still avoid overfitting
    - Usefulness of predictors is examined
  - Goodness of fit metrics
    - Overall fit of the model
      - Deviance (equivalent to SSE in linear regression)
      - $1 - \text{Deviance}/\text{Null Deviance}$  (equivalent to multiple  $R^2$  in linear regression)
    - Single predictors



# LOGISTIC REGRESSION

- Outcome variable with  $m$  classes ( $m > 2$ )
  - Multinomial logistic regression
    - Separate binary logistic regression model for  $m-1$  classes (one class is treated as reference class)
  - Ordinal logistic regression
    - Large no. of ordinal classes: treat ordinal variable as continuous variable and apply multiple linear regression



# LOGISTIC REGRESSION

- Outcome variable with m classes (m>2)
  - Ordinal logistic regression
    - Small no. of ordinal classes: Proportional odds or cumulative logit method
      - Separate binary logistic regression model for m-1 cumulative probabilities
- For a three class case: C1, C2, and C3 and a single predictor x1
$$\text{logit}(C1) = \alpha_0 + \beta_1 x_1$$
$$\text{logit}(C1 \text{or } C2) = \beta_0 + \beta_1 x_1$$
- 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# LOGISTIC REGRESSION PART-7

## LECTURE 52

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# LOGISTIC REGRESSION

- Linear Regression for a categorical outcome variable?
  - Can be done by treating the outcome variable as continuous and coding it numerically
  - However, anomalies will lead to spurious modeling
    - Predictions can take any value, not just dummy values {0,1}
    - Outcome variable or residuals don't follow normal distribution
      - binomial distribution
    - Variance of outcome variable is not constant across all records (violation of homoscedasticity)
      - $np(1-p)$



# LOGISTIC REGRESSION

- Logistic Regression for Profiling Task
  - Apart from model performance on validation partition
  - Model's fit to data is assessed on training partition
    - However, still avoid overfitting
    - Usefulness of predictors is examined
  - Goodness of fit metrics
    - Overall fit of the model
      - Deviance (equivalent to SSE in linear regression)
      - $1 - \text{Deviance}/\text{Null Deviance}$  (equivalent to multiple  $R^2$  in linear regression)
    - Single predictors



# LOGISTIC REGRESSION

- Outcome variable with  $m$  classes ( $m > 2$ )
  - Multinomial logistic regression
    - Separate binary logistic regression model for  $m-1$  classes (one class is treated as reference class)
  - Ordinal logistic regression
    - Large no. of ordinal classes: treat ordinal variable as continuous variable and apply multiple linear regression



# LOGISTIC REGRESSION

- Outcome variable with m classes (m>2)
  - Ordinal logistic regression
    - Small no. of ordinal classes: Proportional odds or cumulative logit method
      - Separate binary logistic regression model for m-1 cumulative probabilities
- For a three class case: C1, C2, and C3 and a single predictor x1
$$\text{logit}(C1) = \alpha_0 + \beta_1 x_1$$
$$\text{logit}(C1 \text{or } C2) = \beta_0 + \beta_1 x_1$$
- 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# ARTIFICIAL NEURAL NETWORKS

## LECTURE 53

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# ARTIFICIAL NEURAL NETWORKS

- Based on
  - Human learning and memory properties
  - Capacity to generalize from particulars
  - Biological activity of brain, where interconnected neurons learn from experience
- Can model complex relationships between outcome variable and set of predictors
  - Applications in Finance (credit card fraud) and engineering disciplines (autonomous vehicle movement)



# ARTIFICIAL NEURAL NETWORKS

- Can model complex relationships between outcome variable and set of predictors
  - Flexible data driven model
    - Not required to specify the form of relationship
    - Useful technique, when functional form of relationship is complicated or unknown
    - Linear and logistic regressions can be conceptualized as special cases
- Neural Network Architectures
  - Multilayer feedforward networks



# ARTIFICIAL NEURAL NETWORKS

- Multilayer feedforward networks
  - Fully connected networks
    - Comprising of multiple layers of nodes
    - With one-way flow and no cycles
  - Input layer
    - First layer of the network
  - Hidden layers
    - Layers between input and output layer



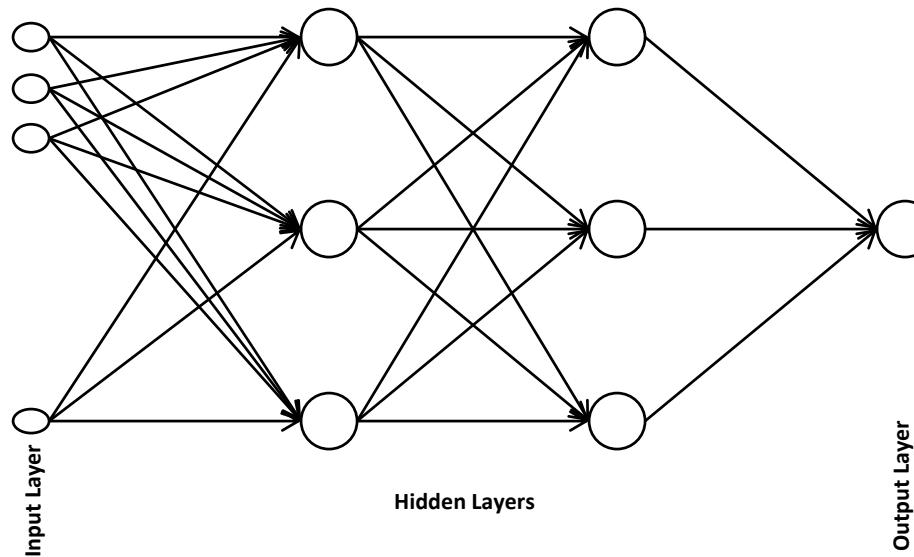
# ARTIFICIAL NEURAL NETWORKS

- Multilayer feedforward networks
  - Output layer
    - Last layer of the network
  - Nodes receive feed from previous layer and forward it to next layer after applying a particular function
  - Function used to map input values (received feed) to output values (forwarded feed) at a node is typically different for each type of layers



# ARTIFICIAL NEURAL NETWORKS

- Multilayer feedforward networks



# ARTIFICIAL NEURAL NETWORKS

- Multilayer feedforward networks
  - Each arrow from node  $i$  to node  $j$  has a value  $w_{ij}$  indicating weight of the connection
  - Each node in the hidden and output layers also has a bias value,  $\theta_j$  (equivalent to intercept term)
- Computing output values at nodes of each layer type
  - Input layer nodes
    - No. of nodes are typically equal to no. of predictors,  $p$
    - Each node will receive input values from its corresponding predictor
    - Output is same as input, that is, predictor's value



# ARTIFICIAL NEURAL NETWORKS

- Computing output values at nodes of each layer type
  - Hidden layer nodes
    - Sum of bias value and weighted sum of input values received from previous layer is computed
$$\theta_j + \sum_{i=1}^p w_{ij}x_i$$
    - Function  $g$  (referred as transfer function) is applied on this sum to produce the output values
    - Transfer function could be a monotone function, for example:
      - Linear function:  $g(x) = bx$
      - Exponential function:  $g(x) = e^{bx}$
      - Logistic or sigmoidal function:  $g(x) = 1/(1+e^{-bx})$



# ARTIFICIAL NEURAL NETWORKS

- Computing output values at nodes of each layer type
  - Hidden layer nodes
    - $\theta_j$  and  $w_{ij}$  are typically initialized to small random values in the range  $0.0 \pm 0.05$
    - Network updates these values after learning from data during each iteration or round of training
  - Output layer nodes
    - Steps are same as for hidden layer nodes, except the fact that input values are received from last hidden layer
    - Output values produced by nodes are used as
      - Predictions in a prediction task
      - Scores to be used to classify a record in a classification task



# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# ARTIFICIAL NEURAL NETWORK PART-2

## LECTURE 54

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# ARTIFICIAL NEURAL NETWORKS

- Computing output values at nodes of each layer type
  - Hidden layer nodes
    - $\theta_j$  and  $w_{ij}$  are typically initialized to small random values in the range  $0.0 \pm 0.05$
    - Network updates these values after learning from data during each iteration or round of training
  - Output layer nodes
    - Steps are same as for hidden layer nodes, except the fact that input values are received from last hidden layer
    - Output values produced by nodes are used as
      - Predictions in a prediction task
      - Scores to be used to classify a record in a classification task



# ARTIFICIAL NEURAL NETWORKS

- Open RStudio
- Neural Network training process
  - Steps to compute neural network output values are repeated for all the records in the training partition
  - Prediction errors are used for learning after each iteration
- Linear and Logistic regression as special cases
  - A neural network with single output node and no hidden layers would approximate the linear and logistic regression models



# ARTIFICIAL NEURAL NETWORKS

- Linear and Logistic regression as special cases
  - If a linear transfer function ( $g(x) = bx$ ) is used, output would be

$$y = \theta + \sum_{i=1}^p w_i x_i$$

- A formulation equivalent to multiple linear regression equation
- However, estimation method (least squares) is different from neural network (back propagation)



# ARTIFICIAL NEURAL NETWORKS

- Linear and Logistic regression as special cases
  - If a logistic transfer function ( $g(x) = 1/(1+e^{-bx})$ ) is used, output would be

$$P(y = 1) = \frac{1}{1 + e^{\theta + \sum_{i=1}^p w_i x_i}}$$

- A formulation equivalent to logistic regression equation
- However, estimation method (maximum-likelihood method) is different from neural network (back propagation)



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE

# ARTIFICIAL NEURAL NETWORKS

- Normalization
  - Scale of [0,1] is typically recommended for neural network models for performance purposes
  - For numeric variables,

$$V_{norm} = \frac{V - \min(V)}{\max(V) - \min(V)}$$



# ARTIFICIAL NEURAL NETWORKS

- Normalization
  - Binary variables (categorical variables with two classes)
    - Create dummy variables: set of values {0, 1}
  - Nominal variables with  $m (>2)$  classes
    - Create  $m-1$  dummy variables: set of values {0, 1}
  - Ordinal variables with  $m (>2)$  classes
    - Map the values to the set {0,  $1/(m-1)$ ,  $2/(m-1)$ , ...,  $(m-2)/(m-1)$ , 1}



# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# ARTIFICIAL NEURAL NETWORK PART-3

## LECTURE 55

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# ARTIFICIAL NEURAL NETWORKS

- Normalization
  - Scale of [0,1] is typically recommended for neural network models for performance purposes
  - For numeric variables,

$$V_{norm} = \frac{V - \min(V)}{\max(V) - \min(V)}$$



# ARTIFICIAL NEURAL NETWORKS

- Normalization
  - Binary variables (categorical variables with two classes)
    - Create dummy variables: set of values {0, 1}
  - Nominal variables with  $m (>2)$  classes
    - Create  $m-1$  dummy variables: set of values {0, 1}
  - Ordinal variables with  $m (>2)$  classes
    - Map the values to the set {0,  $1/(m-1)$ ,  $2/(m-1)$ , ...,  $(m-2)/(m-1)$ , 1}



# ARTIFICIAL NEURAL NETWORKS

- Other transformations
  - Transformations which could spread the values more symmetrically can be done for performance purposes
    - Log transform of a right-skewed variable
- Estimation method
  - Least squares and maximum likelihood methods use a global metric of errors (e.g., SSE) to estimate the parameters



# ARTIFICIAL NEURAL NETWORKS

- Estimation method
  - Neural networks use error values of each observation to update the parameters in an iterative fashion (referred as learning)
    - Error for the output node (prediction error) is distributed across all the hidden layer nodes
    - All hidden layer nodes share responsibility for part of the error (referred as node-specific error)
    - Node-specific errors are used to update the connection weights and bias values



# ARTIFICIAL NEURAL NETWORKS

- Back Propagation
  - An algorithm to update weights and bias values of a neural network
  - Error values are computed from output layer back to hidden layers
    - All hidden layer and output layer nodes and all connection weights become part of learning process
  - Node-specific error for output node,  
$$err = \text{correction factor} \times (\text{actual value} - \text{predicted value})$$
$$\theta_{new} = \theta_{old} + \text{learning rate} \times err$$
$$w_{new} = w_{old} + \text{learning rate} \times err$$
  - Learning rate controls the rate of change from previous iteration
    - Value is typically a constant in the range [0,1]



# ARTIFICIAL NEURAL NETWORKS

- Back Propagation
  - Node-specific error for hidden nodes
    - Based on *err* value of output node instead of prediction error
    - Steps are same as those used for output node
- Methods for updating weight and bias values
  - Case updating
    - Updating is done after each case or record is run through the network (referred as a trial)
    - When all the records are run through the network, it is referred as ***one epoch, or sweep through the data***
    - Many epochs could be used to train the network



# ARTIFICIAL NEURAL NETWORKS

- Methods for updating weight and bias values
  - Batch updating
    - Updating is done after all the records are run through the network
    - In place of prediction error of the record, sum of prediction errors for all records is used
    - Many epochs could be used to train the network
  - Case updating vs. batch updating
    - Case updating yields more accurate results
      - With a longer runtime



# ARTIFICIAL NEURAL NETWORKS

- Stopping Criteria for updating
  - Small incremental change in bias and weight values from previous iteration
  - Rate of change of error function values reaches a required threshold
  - Limit on no. of runs is reached
- Open RStudio



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE

# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# ARTIFICIAL NEURAL NETWORK PART-4

## LECTURE 56

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# ARTIFICIAL NEURAL NETWORKS

- Normalization
  - Scale of [0,1] is typically recommended for neural network models for performance purposes
  - For numeric variables,

$$V_{norm} = \frac{V - \min(V)}{\max(V) - \min(V)}$$



# ARTIFICIAL NEURAL NETWORKS

- Normalization
  - Binary variables (categorical variables with two classes)
    - Create dummy variables: set of values {0, 1}
  - Nominal variables with  $m (>2)$  classes
    - Create  $m-1$  dummy variables: set of values {0, 1}
  - Ordinal variables with  $m (>2)$  classes
    - Map the values to the set {0,  $1/(m-1)$ ,  $2/(m-1)$ , ...,  $(m-2)/(m-1)$ , 1}



# ARTIFICIAL NEURAL NETWORKS

- Other transformations
  - Transformations which could spread the values more symmetrically can be done for performance purposes
    - Log transform of a right-skewed variable
- Estimation method
  - Least squares and maximum likelihood methods use a global metric of errors (e.g., SSE) to estimate the parameters



# ARTIFICIAL NEURAL NETWORKS

- Estimation method
  - Neural networks use error values of each observation to update the parameters in an iterative fashion (referred as learning)
    - Error for the output node (prediction error) is distributed across all the hidden layer nodes
    - All hidden layer nodes share responsibility for part of the error (referred as node-specific error)
    - Node-specific errors are used to update the connection weights and bias values



# ARTIFICIAL NEURAL NETWORKS

- Back Propagation
  - An algorithm to update weights and bias values of a neural network
  - Error values are computed from output layer back to hidden layers
    - All hidden layer and output layer nodes and all connection weights become part of learning process
  - Node-specific error for output node,  
$$err = \text{correction factor} \times (\text{actual value} - \text{predicted value})$$
$$\theta_{new} = \theta_{old} + \text{learning rate} \times err$$
$$w_{new} = w_{old} + \text{learning rate} \times err$$
  - Learning rate controls the rate of change from previous iteration
    - Value is typically a constant in the range [0,1]



# ARTIFICIAL NEURAL NETWORKS

- Back Propagation
  - Node-specific error for hidden nodes
    - Based on *err* value of output node instead of prediction error
    - Steps are same as those used for output node
- Methods for updating weight and bias values
  - Case updating
    - Updating is done after each case or record is run through the network (referred as a trial)
    - When all the records are run through the network, it is referred as ***one epoch, or sweep through the data***
    - Many epochs could be used to train the network



# ARTIFICIAL NEURAL NETWORKS

- Methods for updating weight and bias values
  - Batch updating
    - Updating is done after all the records are run through the network
    - In place of prediction error of the record, sum of prediction errors for all records is used
    - Many epochs could be used to train the network
  - Case updating vs. batch updating
    - Case updating yields more accurate results
      - With a longer runtime



# ARTIFICIAL NEURAL NETWORKS

- Stopping Criteria for updating
  - Small incremental change in bias and weight values from previous iteration
  - Rate of change of error function values reaches a required threshold
  - Limit on no. of runs is reached
- Open RStudio



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE

# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# ARTIFICIAL NEURAL NETWORK PART-5

## LECTURE 57

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# ARTIFICIAL NEURAL NETWORKS

- Estimation method
  - Neural networks use error values of each observation to update the parameters in an iterative fashion (referred as learning)
    - Error for the output node (prediction error) is distributed across all the hidden layer nodes
    - All hidden layer nodes share responsibility for part of the error (referred as node-specific error)
    - Node-specific errors are used to update the connection weights and bias values



# ARTIFICIAL NEURAL NETWORKS

- Back Propagation
  - An algorithm to update weights and bias values of a neural network
  - Error values are computed from output layer back to hidden layers
    - All hidden layer and output layer nodes and all connection weights become part of learning process
  - Node-specific error for output node,  
 $err = \text{correction factor} \times (\text{actual value} - \text{predicted value})$   
 $\theta_{new} = \theta_{old} + \text{learning rate} \times err$   
 $w_{new} = w_{old} + \text{learning rate} \times err$
  - Learning rate controls the rate of change from previous iteration
    - Value is typically a constant in the range [0,1]



# ARTIFICIAL NEURAL NETWORKS

- Back Propagation
  - Node-specific error for hidden nodes
    - Based on *err* value of output node instead of prediction error
    - Steps are same as those used for output node
- Methods for updating weight and bias values
  - Case updating
    - Updating is done after each case or record is run through the network (referred as a trial)
    - When all the records are run through the network, it is referred as ***one epoch, or sweep through the data***
    - Many epochs could be used to train the network



# ARTIFICIAL NEURAL NETWORKS

- Methods for updating weight and bias values
  - Batch updating
    - Updating is done after all the records are run through the network
    - In place of prediction error of the record, sum of prediction errors for all records is used
    - Many epochs could be used to train the network
  - Case updating vs. batch updating
    - Case updating yields more accurate results
      - With a longer runtime



# ARTIFICIAL NEURAL NETWORKS

- Stopping Criteria for updating
  - Small incremental change in bias and weight values from previous iteration
  - Rate of change of error function values reaches a required threshold
  - Limit on no. of runs is reached
- Open RStudio



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE

# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# ARTIFICIAL NEURAL NETWORK PART-6

## LECTURE 58

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# ARTIFICIAL NEURAL NETWORKS

- Estimation method
  - Neural networks use error values of each observation to update the parameters in an iterative fashion (referred as learning)
    - Error for the output node (prediction error) is distributed across all the hidden layer nodes
    - All hidden layer nodes share responsibility for part of the error (referred as node-specific error)
    - Node-specific errors are used to update the connection weights and bias values



# ARTIFICIAL NEURAL NETWORKS

- Back Propagation
  - An algorithm to update weights and bias values of a neural network
  - Error values are computed from output layer back to hidden layers
    - All hidden layer and output layer nodes and all connection weights become part of learning process
  - Node-specific error for output node,  
 $err = \text{correction factor} \times (\text{actual value} - \text{predicted value})$   
 $\theta_{new} = \theta_{old} + \text{learning rate} \times err$   
 $w_{new} = w_{old} + \text{learning rate} \times err$
  - Learning rate controls the rate of change from previous iteration
    - Value is typically a constant in the range [0,1]



# ARTIFICIAL NEURAL NETWORKS

- Back Propagation
  - Node-specific error for hidden nodes
    - Based on *err* value of output node instead of prediction error
    - Steps are same as those used for output node
- Methods for updating weight and bias values
  - Case updating
    - Updating is done after each case or record is run through the network (referred as a trial)
    - When all the records are run through the network, it is referred as ***one epoch, or sweep through the data***
    - Many epochs could be used to train the network



# ARTIFICIAL NEURAL NETWORKS

- Methods for updating weight and bias values
  - Batch updating
    - Updating is done after all the records are run through the network
    - In place of prediction error of the record, sum of prediction errors for all records is used
    - Many epochs could be used to train the network
  - Case updating vs. batch updating
    - Case updating yields more accurate results
      - With a longer runtime



# ARTIFICIAL NEURAL NETWORKS

- Stopping Criteria for updating
  - Small incremental change in bias and weight values from previous iteration
  - Rate of change of error function values reaches a required threshold
  - Limit on no. of runs is reached
- Open RStudio



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE

# ARTIFICIAL NEURAL NETWORKS

- A complete modeling is discussed in the lecture video based on this topic using data of used cars record



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE

# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# Discriminant Analysis

## LECTURE 59

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# Discriminant Analysis

- Statistical technique
  - Used for classification and profiling tasks
  - Model-based approach
  - Idea is
    - To find a separating line or hyperplane equidistant from centroids of different classes
- Or
- Classification procedure is based on distance based metrics
  - Based on the distance of a record from each class



# Discriminant Analysis

- Classification
  - Best separation between items is found by measuring their distance from each class
  - An item is classified to the closest class
- Euclidean distance metric
  - Distance of a record  $(x_1, \dots, x_p)$  from centroid  $(\bar{x}_1, \dots, \bar{x}_p)$  of a class is computed

$$D_{eu}(x, \bar{x}) = \sqrt{(x_1 - \bar{x}_1)^2 + \dots + (x_p - \bar{x}_p)^2}$$

Where centroid  $\bar{x}$  is a vector of means of p predictors



# Discriminant Analysis

- Issues with Euclidean distance metric
  - Distance values depend on the unit of a measurement
  - Based on mean and doesn't account for variance
    - Variability plays an important role in determining the closeness of a record to a particular class
  - Distance should be computed using std. dev. (z-scores) instead of unit of measurement
  - Correlation between variables is ignored



# Discriminant Analysis

- “Statistical distance” (or Mahalanobis distance) can be used to overcome issues with Euclidean distance metric

$$D_{ml}(x, \bar{x}) = [x - \bar{x}]' S^{-1} [x - \bar{x}]$$

Where  $[x - \bar{x}]'$  is transpose matrix of  $[x - \bar{x}]$

- Column vectors are turned into row vectors
- and  $S^{-1}$  is inverse matrix of  $S$  (covariance matrix between p predictors)
- Can be considered as p-dimensional extension of division operation



# Discriminant Analysis

- Linear Classification Functions
  - Used as basis for separation of records into classes
    - Compute classification score measuring closeness of a record to each class
    - Highest classification score is equivalent of smallest statistical distance
  - Main idea is
    - To find linear functions of predictors that maximize ratio of between-class variability to within-class variability
- 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE



IIT ROORKEE



NPTEL ONLINE  
CERTIFICATION COURSE

# Discriminant Analysis Part-2

## LECTURE 60

DR. GAURAV DIXIT  
DEPARTMENT OF MANAGEMENT STUDIES



# Discriminant Analysis

- Statistical technique
  - Used for classification and profiling tasks
  - Model-based approach
  - Idea is
    - To find a separating line or hyperplane equidistant from centroids of different classes
- Or
- Classification procedure is based on distance based metrics
  - Based on the distance of a record from each class



# Discriminant Analysis

- Classification
  - Best separation between items is found by measuring their distance from each class
  - An item is classified to the closest class
- Euclidean distance metric
  - Distance of a record  $(x_1, \dots, x_p)$  from centroid  $(\bar{x}_1, \dots, \bar{x}_p)$  of a class is computed

$$D_{eu}(x, \bar{x}) = \sqrt{(x_1 - \bar{x}_1)^2 + \dots + (x_p - \bar{x}_p)^2}$$

Where centroid  $\bar{x}$  is a vector of means of p predictors



# Discriminant Analysis

- Issues with Euclidean distance metric
  - Distance values depend on the unit of a measurement
  - Based on mean and doesn't account for variance
    - Variability plays an important role in determining the closeness of a record to a particular class
  - Distance should be computed using std. dev. (z-scores) instead of unit of measurement
  - Correlation between variables is ignored



# Discriminant Analysis

- “Statistical distance” (or Mahalanobis distance) can be used to overcome issues with Euclidean distance metric

$$D_{ml}(x, \bar{x}) = [x - \bar{x}]' S^{-1} [x - \bar{x}]$$

Where  $[x - \bar{x}]'$  is transpose matrix of  $[x - \bar{x}]$

- Column vectors are turned into row vectors
- and  $S^{-1}$  is inverse matrix of  $S$  (covariance matrix between p predictors)
- Can be considered as p-dimensional extension of division operation



# Discriminant Analysis

- Linear Classification Functions
  - Used as basis for separation of records into classes
    - Compute classification score measuring closeness of a record to each class
    - Highest classification score is equivalent of smallest statistical distance
  - Main idea is
    - To find linear functions of predictors that maximize ratio of between-class variability to within-class variability
- Open RStudio



# Discriminant Analysis

- Assumptions and other issues
  - Predictors follow multivariate normal distribution for all classes
    - Given adequate sample points for all classes, relatively robust to violations of normality assumption
  - Correlation structure between predictors for each class should be same
  - Sensitive to outliers



# Discriminant Analysis

- Further Comments on discriminant analysis
  - Application and performance aspects are similar to multiple linear regression
  - In discriminant analysis, coefficients of linear discriminant are optimized w.r.t class separation
    - In linear regression, coefficients are optimized w.r.t outcome variable
  - Estimation technique is least squares
    - Same as linear regression



# 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...



IIT ROORKEE



NPTEL  
ONLINE  
CERTIFICATION COURSE