



BC2406: Business Analytics I

Seminar 12: Review

*Instructor: Prof. Lee Gun-woong
Nanyang Business School*

Exam Date and Format

■ Exam Date / Duration

- Date: Thursday, 1 December 2016 at 1:00pm
- Duration: 2.5 hours
- **Closed Book, Closed Laptop**

■ Examination Venue

- Will be announced

■ Format

- Total of 4 Questions (with multiple parts)
 - Principles in Business Analytics
 - Data Mining Techniques & Data Management
 - Regression
 - Classification
 - Clustering
 - Text-mining
- **Make sure to bring a calculator**

- **Read the question carefully. Read, think, and read again!**
 - Understand the focus of the question
 - How would you answer it?
- **Take a look at your answer**
 - Did you address the issues raised in the question?
 - Are your arguments logical?
 - Did you answer the question in the context of the question?
 - Is your answer complete?

Exam Tips, cont.

4

- **Attempt all questions**

- No negative marking
- Partial answers will at least earn partial credit!

- Write complete sentences – no bullet points!

- Organize your answer in well connected paragraphs

- Underline or highlight main ideas or key points

- **Write neatly and legibly!!!!!!!!!!!!!!**

How to Prepare?

- **Seminar slides, Assignments, R Exercises, R Tutorials, Past Exam**
 - Exam Exclusions: Case Articles, Project, Supplementary Slides
- **All topics are important!**
- Focus on concepts, procedures, logics and implications

Review: Principles in BA

6

■ Key Terms in BA

- Data Science
- Data Analytics
- Business Analytics
- Data Mining

■ Key Elements of BA

- Data
 - 3 Vs of Big Data
- Analytics
 - Descriptive Analytics vs. Predictive Analytics
- Business Problem
 - Decompose the overall problem into subtasks
 - Compose the solutions to the subtasks to solve the overall problem
- Insights for Decision Making

Review: Data Mining and Data Management

7

- Data Mining Techniques
 - Clustering, Association, Classification, Prediction
- Data Mining Tasks
 - Supervised Tasks vs. Unsupervised Tasks
- Non-Data mining Techniques
 - Statistics, Data Querying, Optimization
- Data Structure
 - Record Data vs. Transaction Data
- Variable Types
 - Numerical Variables
 - Continuous vs. Discrete Variables
 - Categorical Variables: Ordered vs. Unordered variables
- Sampling
- Data Treatment
 - Summary Statistics (mean, median, SD, min., max.) and correlation
 - Data Cleaning
 - Duplicate data, Outliers, Missing Values
 - Data Pre-processing:
 - Normalization, Variable Transformation

Review: Regression

8

- Regression analysis is a **probabilistic model** describing the **relationship** between a single dependent variable and one or more independent variables with unobserved errors.
 - **Simple Regression Model vs. Multiple Regression Model**
- Linear Relationship: **Intercept** and **Slope**
- Estimation Method: Ordinary Least Square (**OLS**)
- Model Evaluation
 - Model fit: R^2 vs. **adjusted R^2**
 - Model Specification: F-test, $H_0: \beta_1 = \beta_2 = \dots = 0$
- Interpretation

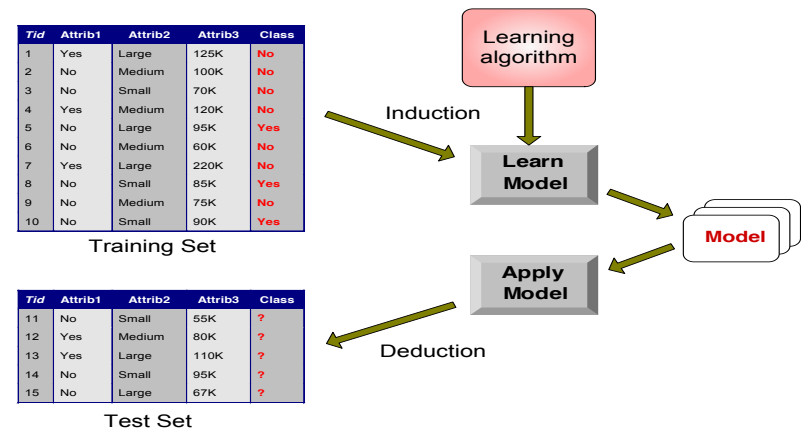
*** = 0.1%, ** = 1%, * = 5% Significance levels

Dependent Variable	Independent Variable	Estimated Price (\$)	Interpretation
Sales	Price	- 10.232***	A one-unit increase in Price decreases Sales by \$10.232 at the 0.1% significance level.
Sales	Log (Price)	- 9.389**	A one-percent increase in Price decreases Sales by \$0.09389 (=coefficient/100) at the 1% significance level.
Log (Sales)	Price	- 0.045*	A one-unit increase in Price decreases Sales by 4.5% (=coefficient*100) at the 5% significance level.
Log (Sales)	Log (Price)	- 0.032	A one-percent increase in Price decreases Sales by 0.032%

Review: Classification

9

- Build a Decision Tree
 - Select the attribute that produces the “purest” nodes and make that attribute a decision node
 - Gini Index, Entropy
 - Repeat this process for recursively for each child node
 - Stop when:
 - All (almost all) the instances have the same class attribute value
 - There are no more instances / attributes
- Evaluate model performance
 - Confusion Matrix: Accuracy, Error Rate, Sensitivity, Specificity, ROC Curve
- Overfitting and Pruning
 - Low Training Error
 - Low Generalisation Error
- Improve Model Performance
 - Boosting, Cost Matrix



Review: Clustering

10

- An **unsupervised DM task** that divides the data into cluster
 - Lower level observations can be clustered into **higher orders**
- (Dis)Similarity is measured by Distance
 - Minimise the distance between objects (**within clusters**)
 - Maximise the distance between clusters
- Data Preparation
 - Variable Standardisation (**z-score**)
- Hierarchical Clustering (Agglomerative methods)
 - Between Objects: **Euclidean Distance**
 - Between Clusters: **Min., Max., Avg., and Centroid Distance**, and **Ward's Method**
 - Output: **Dendrogram**
- K-means Clustering
 - Requires a user to specify the number of clusters
 - Minimise Within-clusters Sum of Squares (**WSS**)
 - Elbow Method

Review: Text-mining

11

- A process of identifying **hidden/meaningful information** from a collection of texts
 - The output of TM can be used for other DM techniques
- **Key Steps**
 - Create a **Corpus**
 - Clean text data
 - Parsing
 - Convert upper-case letters to lower-case letters
 - Remove unnecessary/less-important terms
 - Remove stop words
 - Remove Numbers, Remove punctuations, and symbols
 - Remove white space
 - Stemming
 - Reduce inflected words to their **word stem**
 - Create a **DTM**
 - Convert unstructured text data into structured format
 - Sparsity Rate
 - Create a frequency table
 - Use the TM outcomes for other DM techniques

- No coding! No filling in the blanks! No math!
- Try the assignments and exercises by yourself!
- Understand the logic behind the code
 - Understand key procedures in DM techniques
 - How many predictors are used in a regression analysis?
 - Which terms are removed from a text?
 - Interpret the output from functions
 - What is the sparsity rate in DTM?
 - Explain adjusted R^2

What's Next


Complete the Following for Course Project

14

- Teamwork and Interpersonal skills Evaluation using E-Rubric in Eureka (Mandatory)
 - Refer to the Student Guide posted on the course site.
 - **The evaluation should be completed latest by Friday, 11th November.**
- Project Peer Review (Optional)
 - Needs to be completed only if there's an unequal contribution by team members in a project
 - **The form should be emailed to me latest by Friday, 11th November.**

Student Feedback On Teaching

NTU Application Mail

 FollowUp. Start by Sunday, October 23, 2016. Due by Sunday, October 23, 2016.

Sent: Mon Oct, 24, 2016 12:02 AM

To: Lee Gun Woong (Asst Prof)

Dear Faculty,

The message below has been sent to your students.

Dear Student,

The Online Faculty Teaching Evaluation for Semester 1 AY16-17 is now open.

Please [click here](#) to provide the feedback. The link is valid from 24-OCT-2016 to 13-NOV-2016.

We would like to strongly encourage you to participate in this exercise. Giving teaching feedback to the School will benefit the student population as we work towards refining and improving our teaching pedagogy and course materials. Together we can develop a more nurturing and effective learning environment for every student.

Thank you and with Best Regards,

SFT Administrator

This is a computer generated letter. No signature is required.

**Thank You!!
& Good Luck!!**