Study Guide – Midterm Exam

DATA 3300

The following is a list of topics that will be covered on the midterm exam. Please see the Canvas page “Module 7: Midterm Exam” for details about the exam and study materials.

Data Analytics & Data Science

- What data science is and how it’s used
- Common sources of data:
    a. Social networks
    b. Traditional business systems
    c. Internet of Things
- Different types of data analytics and their application(s):
    a. Diagnostic
    b. Descriptive
    c. Predictive
    d. Prescriptive
- The sequence of steps in the CRISP-DM process (and the importance of each):
    a. Business Understanding
    b. Data Understanding
    c. Data Preparation
    d. Modeling
    e. Evaluation
    f. Deployment

Data Quality & Preparation

- What ETL is and why it is important in data analytics
- Five data quality characteristics:
    a. Accuracy
    b. Uniqueness
    c. Completeness
    d. Consistency
    e. Time-Appropriateness
- Common forms of “dirty data” (and the threats they pose to data analysis):
    a. Errors (typos, misspellings)
    b. Inconsistent Data
    c. Absence of Data
    d. Contradicting Data
    e. Reused Primary Keys
- Common steps in data cleansing (and how long data cleansing takes as part of the overall data mining process):
    a. Parsing
    b. Correcting

Study Guide – Midterm Exam

DATA 3300

c. Standardizing
d. Matching
e. Consolidating

Data Understanding

- Familiarize yourself with the differences between qualitative and quantitative variable types:
  a. Qualitative:
     i. Nominal
     ii. Ordinal
  b. Qualitative
     i. Ratio
     ii. Interval
- Understand how each of the metrics below relate to data exploration (data distribution, central tendency, and data dispersion)
  a. Understand each of the following descriptive statistics and their purpose:
     i. Mean
     ii. Median
     iii. Mode
     iv. Variance
     v. Standard deviation
     vi. Interquartile range
     vii. Outliers
  b. Understand and identify:
     i. Skewness
     ii. Kurtosis
- Basic principles of visualization (and how to interpret visualizations), including most common chart types and their purpose (and how to create visualizations that clearly communicate the data, not just reinforce prior beliefs):
  a. Histograms
  b. Line charts
  c. Box plots
  d. Pie Charts
  e. Stacked column chart

Modeling Foundations

- What data mining is, its appropriate applications, and common data mining tasks
- What is meant by the terms:
  a. Data instance/Record/Case/Observation
  b. Attributes/Variables
     i. Target attribute/Dependent variable

Study Guide – Midterm Exam

DATA 3300

- The difference between supervised and unsupervised data mining
- The difference between classification and regression types of supervised data mining
    a. Classification → When the DV is categorical
    b. Regression → When the DV is numerical

Association Rules Analysis

- What association analysis is, the type of data it requires, and the types of business questions it can answer
- What an association rule looks like:
    a. Itemsets and their role in association rules
    b. Antecedents and consequents
- Understand **how to calculate** and **interpret**:
    a. Support
    b. Confidence
    c. Lift
- Know the tradeoffs between adjusting the minimum support and confidence thresholds and the resulting association rules generated

Clustering Analysis

- What clustering analysis is, the type of data it requires, and the types of business questions it can answer
- What *k*-means clustering is:
    a. Understand the steps involved
        i. Why we sometimes normalize data in cluster analysis
            1. z-score normalization
        ii. Impact of potentially over-weighting variables which are measuring the same thing in different ways
    b. Know what *k* stands for and how it is determined
        i. Post-hoc evaluation
        ii. Elbow rule
        iii. Tractability
    c. Understand the basics of how the algorithm works
        i. How are clusters determined?
        ii. How is a centroid value determined?
        iii. What is the relationship between a cluster and a centroid?
            1. How do you interpret results with centroid values?
- Ways to calculate similarity/dissimilarity between cases and how to interpret the distance
    a. Euclidian Distance (know and interpret formula)
    b. Intra-class similarity
    c. Inter-class similarity
- Understand how to interpret a cluster analysis through a centroid table and plot.

Study Guide – Midterm Exam

DATA 3300

Statistical Correlation

- What correlation analysis is, the type of data it requires, and the types of business questions it can answer
- Know how to identify and interpret:
    a. Correlation coefficient
    b. Correlation analysis results
    c. Convergent validity (and when to use it)
    d. Coefficient of determination (know how to calculate it)
- What scatter plots look like for strong versus no relationship between two variables
- Assumptions and limitations of correlation analysis:
    a. Homoscedasticity
    b. Normal distribution of data
    c. Impact of outliers

Good Luck!