Format
5 questions answer 4 questions
Q1 - Q3 compulsory

- Chapter 1 & 2
- Chapter 3
- Chapter 5 + Python coding

Q4 (Chapter 4) & Q5 (Chapter 6): 2 answer 1

40% theory + 60% application

Python coding:
Basic Python numpy array operations

Chapter 1

- Definition of data mining, machine learning and data analytics
- Applications of data mining

Definition and applications of data mining methods

- Classification
- Regression
- Clustering
- Association
- Sequence or path analysis

Chapter 2

Type of data / attributes - application and example
Curse of Dimensionality
Dimensionality Reduction and techniques, e.g. SVD, PCA etc.

Proximity Measures - similarity and distance measures
Application and problem solving: (refer tutorial questions)

- Euclidean Distance
- Minkowski/Manhattan Distance
- Mahalanobis Distance
- Simple Matching Coefficient (SMC)
- Jaccard (J) Coefficient
- Cosine Similarity

Application of exploratory data analysis and summary statistics - compute mean, median, mode, std dev, AAD, MAD, interquartile range etc. (refer tutorial questions)

Visualization: learn how to plot box plot, scatter plot and line plot.

Chapter 3

Definition and understand the key concepts in: kNN, decision tree, naive bayes, svm, ensemble methods (Adaboost and Bagging examples)
Application of problem solving (refer tutorial questions): kNN, decision tree, naive bayes
Compute (learn via examples)

- Gini Index **
- Entropy **
- Gain ratio
- Information gain
- Misclassification error

Know how to interpret and make decision on decision tree splitting based on the computed index above.

Confusion matrix: compute all the different metrics based on a given example (refer lecture example)
Know how to interpre confusion matrix
Explain cross-validation methods

- Holdout Method
- Random Subsampling
- K-fold Cross Validation
- Leave-one-out Cross validation

ROC curve, Precision-Recall curve - understand and know how to interpret

Chapter 4

Definition: regression, simple & multiple linear regression
Compute correlation between 2 datasets
Construct least squares line of regression line given a set of data

Chapter 5

Definition:

- Partitional and hierarchical clustering (agglomerative vs divisive)
- Outlier

Different types of hierarchical clustering:
Applications of k-Means clustering on a given problem (refer tutorial questions)
Application of DBScan

Chapter 6

Definition and application in an example

- Association rule
- Support
- Confidence
- Itemset
- Frequent itemset

Application of Apriori principle - example (refer tutorial questions)

Note: all key equations / formulas will be provided in appendix