DATA MINING (INFO-H-423)
Mahmoud SAKR
Intended Learning Outcomes

Introduction & Decision Tree

- Explain the steps of supervised learning.
- Describe classification as one of the DM tasks.
- Explain/Illustrate the concepts of Entropy, Gain, Gain-ratio.
- Explain and Apply the ID3 Algorithm to a given dataset.

Data Preparation and Distance Measures

- Illustrate and compare the different model validation methods: holdout, cross validation, bootstrap.
- Describe the different data preparation tasks, their goals, and few examples. The details of the methods are not required. It is enough to know the general idea.
- Describe without detail the concepts and methods of data cleaning, scaling, normalization, and distance measures.
- Explain and apply the Lp-norm distance, and the Edit distance.

Clustering

- Explain clustering as an unsupervised learning method.
- Apply the K-Means, the K-Medians, and the DBSCAN Algorithms to given datasets.
- Compare K-Means, K-Medians and K-Medoids Algorithms.
- Compare between the different families of clustering Algorithms (representative-based, density based, probabilistic model-based).
- Analyze the complexity of the studied clustering Algorithms.
- Explain how to assess clustering quality

Note: the details of the probabilistic model-based clustering are not required for this exam.
Only the general concept of fuzzy clustering is required.

Timeseries Forecasting

- Explain the different types of forecasting models
- Illustrate with examples the trend, seasonality, and cycle components of a time series
- Apply classical timeseries decomposition for a given timeseries
- Illustrate timeseries forcasting with decomposition
- Describe mathematical concepts for analyzing time series, including concepts of white noise, stationarity, and autocorrelation
- Explain AR model, and parameters
- Explain MA model, and parameters
- Explain ARMA, ARIMA models, and parameters
- Explain the Prophet model

Outlier Mining

- Illustrate with examples the types of outliers: global, contextual, and collective
- Illustrate the use of grids in speeding up distance based outlier detection
- Discuss in general terms the different methods of outlier mining: distance based-clustering based, classification based.

Frequent itemsets & Association rules Mining

- Illustrate the uses of frequent pattern mining.
- Explain and compare the measures: support, confidence, lift, correlation analysis.

• Explain the Apriori property, and its applications in optimizing the search for frequent itemsets.
• Apply the Apriori and the FP-Growth Algorithms to given datasets.
• Compare the frequent pattern mining Algorithms: brute force enumeration of the itemset lattice, Apriori, and FP-Growth Algorithms.
• Apply your understanding of the different frequent itemset mining methods to reason about further optimizations.
• Extract association rules from frequent itemsets, and assess their quality.

Classification

• Explain the naive Bayes classifier, and apply it to a given dataset.
• Explain the concept of the confusion matrix, and use it to define the different classification quality measures.
• Motivate ensemble learning, as a method to improve classification accuracy.
• Illustrate the difference between bagging and boosting.
• Describe random forest induction, and the tuning of parameters L, D.

Stream Mining

• Describe and illustrate Bloom filter, Count-Min, Flajolet-Martin, and hyperloglog
• Apply these methods to query real data streams

Note: No need to memorize equations and Algorithms. They will be given if needed.

With my best wishes.