COMP9318 Review

Yifang Sun @ UNSW

April 22, 2021

Data Warehousing and OLAP

- Understand the four characteristics of DW (DW vs. Data Mart)
- Differences between OLTP and OLAP
- Multidimensional data model; data cube;
  - fact, dimension, measure, hierarchies
  - cuboid, cube lattice
  - three types of schemas
  - four typical OLAP operations
  - ROLAP/MOLAP/HOLAP
- Query processing methods for OLAP servers, including the BUC cubing algorithm.

NOT needed:

- Design good DW schemas and perform ETL from operational data sources to the DW tables.

Linear Algebra

- Column vectors; Linear combination; Basis vectors; Span
- Matrix vector multiplication
- Eigenvalues and eigenvectors
- SVD: general idea.

Classification and Prediction

- Classification basics:
  - overfitting/underfitting; cross-validation
  - Classification vs prediction; vs clustering (unsupervised learning); eager learning vs. lazy learning (instance-based learning)
- Decision tree:
  - The ID3 algorithm
  - Decision tree pruning
  - Derive rules from the decision tree
  - The CART algorithm (with gini index)
- Naive Bayes classifier
  - Smoothing
  - Two ways to apply NB on text data
- Logistic regression/MaxEnt classifier; Maximum likelihood estimation of the model parameters + regularization; Gradient ascend.
- SVM: Main idea; the optimization problem in the primal form; the decision function in the dual form; kernel

Cluster Analysis

- Clustering criteria: minimize intra-cluster distance + maximize inter-cluster distance
- Distance/similarity
  - how to deal with different types of variables
  - distance functions: L_p
  - metric distance functions

Cluster Analysis /2

- Partition-based Clustering: k-Means (algorithm, advantages, disadvantages, . . .)
- Hierarchical Clustering: agglomerative, single-link / complete-link / group average hierarchical clustering
- Graph-based Clustering: Unnormalized graph laplacian and its semantics, overview of spectral clustering algorithm; embedding.

Association Rule Mining

- Concepts:
  - Input: transaction db
  - Output: (1) *frequent* itemset (via *minsup*); (2) association rules (via *minconf*)
- Apriori algorithm:
  - *Apriori property* (2 versions)
  - The Apriori algorithm
    - How to find frequent itemsets?
    - How to derive the association rules?

Association Rule Mining /2

- FP-growth algorithm:
  - How to mine the association rule using FP-trees?
- Derive association rules from the frequent itemsets.

Thanks You and Good Luck!