



COMP9318 Review

Wei Wang

`weiw AT cse.unsw.edu.au`

School of Computer Science and Engineering
Universities of New South Wales

June 3, 2009





Outline



Introduction Review

Introduction

Review

Introduction

Course Logistics

Introduction Review

■ THE formula:

$$q = q1$$

$$t = \frac{ass1 + proj1 + q}{3.0}$$

$$mark = \frac{exam \times t}{0.5 \cdot exam + 0.5 \cdot t}$$

- ass1 and proj1 will be marked ASAP and hopefully be released before the exam
- Pre-exam consultations to be announced (most likely 24–25 Jun).
- Course survey: Please participate and fill in your comments, so that we can further improve the course!

Tip: The final exam mark is important!

Review

About the Final Exam

Introduction Review

- **Time:** 1745 – 2100, 26 June 2009 (Fri), 10 minutes reading time + 3 hr closed-book exam.
- **Venue:** ElecEng G25
- **Accessories:** UNSW Approved Calculator
- Designed to test your *understanding* and familiarity of the core contents of the course.
- 100 (7 questions) + 5 *bonus* (1 question)
 - ◆ Q1: short answer (can use your own words)
 - ◆ others will requires some “calculation” (i.e., similar to tute/ass questions)
- Read the instructions carefully.

Tip: Write down intermediate steps.

About the Final Exam ...

Introduction Review

- “Advanced” Methods/algorithms/systems are not required, unless explicitly mentioned here.

Disclaimer: *We will go through the main contents of each lecture. However, note that it is by no means exhaustive.*

Introduction

Introduction Review

- DM vs. KDD
- Steps of KDD; iterative in nature; results need to be validated.
- Database (efficiency) vs. Machine learning (effectiveness) vs. Statistics (validity): e.g.?
- Descriptive/predictive data mining: e.g.?
- Able to cast a real problem into a data mining problem.

Data Warehousing and OLAP

Introduction Review

- Understand the four characteristics of DW (DW vs. Data Mart)
- Differences between OLTP and OLAP
- Multidimensional data model; data cube; simple MDX queries
 - ◆ fact, dimension, measure, hierarchies
 - ◆ cuboid, cube lattice
 - ◆ three schemas
 - ◆ four typical OLAP operations
 - ◆ ROLAP/MOLAP/HOLAP
- Query processing methods for OLAP servers.
- Able to design good DW schemas and perform ETL from operational data sources to the DW tables.

Data Preprocessing

Introduction Review

- Understand that real data is “dirty” (incomplete, noisy, inconsistent)
- How to handle missing data?
- How to handle noisy data? different binning/histogram method (including V-optimal and MaxDiff)
- Set/String distance/similarity measures (Jaccard/cosine similarity, edit distance)
- Prefix-filtering-based similarity join algorithm

Association Rule Mining

Introduction Review

- Concepts:
 - ◆ Input: transaction db
 - ◆ Output: (1) *frequent* itemset (via *minsup*); (2) association rules (via *minconf*)
 - ◆ Other “interesting” measures (e.g., lift)
- Apriori algorithm:
 - ◆ *Apriori property* (2 versions)
 - ◆ The Apriori algorithm
 - How to find frequent itemsets?
 - How to derive the association rules?
 - ◆ Push *anti-monotonic* constraints into the Apriori algorithm

Association Rule Mining /2

Introduction Review

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

Classification and Prediction

Introduction Review

- Classification vs prediction; vs clustering (unsupervised learning); eager learning vs. lazy learning (instance-based learning)
- The danger of overfitting
- Decision tree:
 - ◆ The ID3 algorithm
 - ◆ Derive rules from the decision tree
 - ◆ The SPINT algorithm (with gini index)
- Naive Bayes classifier
- k NN classifier
- Testing methods: training + testing datasets, cross-validation, leave-one-out

Cluster Analysis

Introduction Review

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

Cluster Analysis /2

Introduction Review

- Partition-based Clustering: the original k -Means (algorithm, advantages, disadvantages, ...)
- Hierarchical Clustering: agglomerative, single-link / complete / average-link hierarchical clustering
- Density-based Clustering: DBSCAN (...)