COMP9318 Review

June 4, 2018

Course Logisitics

- **THE** formula:

$$mark = 0.55 \cdot exam + 0.15 \cdot (ass1 + proj1 + lab)$$

$$mark = \mathbf{FL}, \text{ if } exam < 40$$

$$lab = avg(best\_of\_3(lab1, lab2, lab3, lab4, lab5))$$

-
-
  - 18 Jun: 1200–1400, K17-508
- Course feedback: via comments in the course sur
  messages to me on the forum. We are particularly i
  aspects such as **coverage**, **difficulty le**
  python/Jupyter, **project**, and **background required**.

Note

(1) The final exam mark is important and you must achieve at least 40!
(2) Supplementary exam is only for those who cannot attend the final exam.

About the Final Exam

- **Time**: 1345 – 1600, 19 Jun 2016 (Tue), 10 minutes reading time + 2 hr closed-book exam.
- **Accessories**: *UNSW Approved Calculator*. Note: watches are prohibited.
- Designed to test your *understanding* and familiarity of the core
- https://eduassistpro.github.io/ry.

  similar to tute/ass questions)

About the Final Exam /2

- Read the instructions carefully.
- Use your time wisely. Don't spend too much time if stuck on one question or writing excessively long answers on Q1.

Tips

(1) Wr n your calcu
new pa

Disclaimer

We will go through the main contents of each lecture. Ho that it is by no means exhaustive

Introduction

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

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
- cubing algorithm.

NOT needed:

- Design good DW schemas and perform ETL fro sources to the DW tables.

Linear Algebra

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

Data Preprocessing

- Understand that real data is "dirty" (incomplete, noisy, inconsistent)
- How to handle missing data?
- How to normalize the data?
- How to handle noisy data? different binning/histogram method (including V-optimal and MaxDiff)
-
- NOT
-
  t-SNE)

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:
  -

- Naive Bayes classifier
  - Smoothing
  - Two ways to apply NB on text data
- Logistic regression/MaxEnt classifier; Max 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
  -

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

Association Rule Mining

- Concepts:
  - Input: transaction db
  - Output: (1) *frequent* itemset (via *minsup*); (2) association rules (via *minconf*)
- Apriori algorithm:
  -

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!