COMP6237 Data Mining

Data Mining Exam Info + Q&A

Jonathon Hare
jsh2@ecs.soton.ac.uk

The Exam I

- **Tuesday 17th May 14:30 - 16:30**
- Potentially covers ALL material in the lectures and some further reading
- (approx 20% of each question has marks for going beyond lecture material)

The Exam II

- Expect to have to **solve problems** as well as **describe approaches**
    - *Obviously problem solving will be limited to what you can do with a pen, paper and* ***calculator***
        - ***Make sure you show working if you want to get full marks***
- *You will likely also be asked to think - there will be parts of questions that won't have been covered; you'll need to* ***apply your knowledge*** *and* ***reason a sensible solution***

UNIVERSITY OF SOUTHAMPTON COMP6237W1

SEMESTER 2 EXAMINATION 2015 - 2016

DATA MINING

DURATION 120 MINS (2 Hours)

This paper contains 5 questions

Answer **ONE** question from Section A and **TWO** questions from Section B.

An outline marking scheme is shown in brackets to the right of each question.

University approved calculators MAY be used.

A foreign language dictionary is permitted ONLY IF it is a paper version of a direct Word to Word translation dictionary AND it contains no notes, additions or annotations.

10 page examination paper.

- Expect Section A to have questions related to more fundamental mathematical concepts (related to data mining).
    - These questions have many small parts

- Expect Section B to have deeper questions about data mining techniques and their application
    - These questions are likely to have ~3 parts

Useful Equations and Tables

Pearson's Correlation for a sample:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Cosine Similarity:

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{|\mathbf{A}||\mathbf{B}|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Laws of logarithms:

$$\begin{aligned}
\log_c(ab) &= \log_c(a) + log_c(b) \\
\log_c(\tfrac{a}{b}) &= \log_c(a) - log_c(b) \\
\log_a(x) &= \log_b(x)/\log_b(a) \\
b^{\log_b(x)} &= x \\
\log_b(b^x) &= x
\end{aligned}$$

Table of $\log_2$ for small numbers (rounded to 2 d.p.):

x	0.1	0.2	0.3	$0.\dot{3}$	0.4	0.5	0.6	$0.\dot{6}$	0.7	0.8	0.9	1.0
$\log_2(x)$	-3.32	-2.32	-1.74	-1.59	-1.32	-1.00	-0.74	-0.58	-0.52	-0.32	-0.15	0.00

Past papers?

- No past papers - this is a new module
- But…
  - I did teach on Adv. Machine Learning last year
  - That material was moved to data-mining

Question 1.

(a) **Provide** a description of the MapReduce framework. **Include** details of both the *programming model* and underlying *distributed data model*.

[8 marks]

(b) Assume that you have a large number of feature vectors stored on a distributed file system across a cluster of machines. **Describe** how you might apply the MapReduce programming model to distribute the problem of applying *k-means clustering* to the data. Also **describe** any *limitations* of your approach and any possible *workarounds*.

[12 marks]

Q&A Time