COMP6237 Data Mining

Data Mining Exam Info + Q&A

Jonathon Hare
jsh2@ecs.soton.ac.uk

The Exam I

- **Wednesday 31st May 14:30 - 16:30**
- Potentially covers ALL material in the lectures
  - will require 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***

The Exam III

- You will likely be asked to **draw diagrams** to explain how certain things work
    - if the question asks you to draw something, then consider that you won't get full marks if you don't draw anything!
- You don't need to write essays to answer questions
    - Bullet points, etc are fine

UNIVERSITY OF SOUTHAMPTON COMP6237W1

SEMESTER 2 EXAMINATION 2016 - 2017

DATA MINING

DURATION 120 MINS (2 Hours)

This paper contains 5 questions

Answer **ONE** question from Section A and **TWO** questions from Section B. You are advised to spend no longer than 40 minutes per question.

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

Each question is worth 33 marks. A maximum of 99 marks are available for the paper.

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?

- Last years paper is available
- Also…
  - I did teach on Adv. Machine Learning two years ago
  - That material (hadoop and big data) was moved to data-mining

Q&A Time