Revision

CITS4009 Computational Data Analysis

Dr. Mubashar Hassan

Department of Computer Science and Software Engineering
The University of Western Australia

Semester 2, 2024

Exam Structure

The exam contributes to 60% of the total assessment, with a total mark of 100.

- 5 Questions (each is worth 20 marks)
  - each covers a main topic of the unit
  - each contain 4 to 5 sub-questions
  - marks allocated to each sub-question are mentioned.
- Close book face to face exam.
- Approximated 1 minute per mark. Remaining 20 minutes for revision.
- One A4 page of double-sided notes is allowed.

Exam Questions Format

- Mostly short answer questions
  - requiring conceptual understanding, and/or
  - descriptions of procedures or illustrations of results interpretation
  - no need to draw diagrams
- Some coding questions
  - need to get the required function or syntax mostly right
- It is OK to use dot points, no need to provide "verbose" answers.
- Old sample exams are provided on the unit page on LMS, click "Exam info" link.

Q1 Exploratory Data Analysis (20)

- Data Science Life Cycle
- Basic R Syntax
- EDA and Visualisation
- Layered Grammar of Graphics in ggplot
- The EDA process

```
+ What are the sensible questions to ask
+ What type of geoms are most suitable for the questions
```

- Bar chart (Coding maybe required)
- Boxplot and stats (Coding maybe required)

Q2 Data Cleaning, Transformation and Visualisation (20)

Mainly coding questions. Given a data frame, write code to

- Dealing with Missing Values and Sentinel Values
- Normalisation and Scaling
- Subsetting and creating new columns through calculation
- Understand how to use functions such as mutate() and pipes for code simplification

Q3 Data Integration (20)

- Understand Relational Data (joins)
- Write code to get information out of multiple data frames

Q4 Classification (20)

Given a dataset, the questions are to test on the following

- Single Variable Models
  - What are Single Variable Models
  - Why Single Variable Models
  - How to calculate contingency table
  - How to work out the prediction value
- Understand Model Evaluation
  - Precision and Recall
  - Area under the ROC (Receiver Operating Characteritic Curve)
  - loglikelihood, deviance, AIC
  - Double Density Plots
- Classification Models
  - Know the intuition behind Decision Trees and KNN
  - Given contextualised answer when given a data frame.

(Coding maybe required)

Q5 Clustering and Regression (20)

- Clustering
  - understand the general idea of clustering
  - understand the distance measures for numerical and categorical variables
  - understand how dist() function works, contextualised to a given data frame
  - how to evaluate cluster models
  - how to select k
- Regression
  - understand what linear and logistic regression is
  - how to specify the dependent and independent variables
  - how to use double density plot in model selection
  - Relation between linear regression and logistic regression

(Coding maybe required)

Concluding Remarks

- No questions on Data Import.
- No questions on LIME and Shiny App.
- We have only scratch the surface of the R language and Machine Learning, but this establishes the fundamentals so you learn more on your own or future units.
- As an example, take a look at the powerful tools available in R Markdown and Shiny App for better communicate with your stakeholders and end users
  - R Markdown Formats: https://rmarkdown.rstudio.com/formats.html
  - R Markdown Gallary: https://rmarkdown.rstudio.com/gallery.html
  - Shiny App (for both R and Python): https://shiny.posit.co/

All the very best to your exams!