

COMP20008 Exam Study Guide Weeks 1-12 Semester 1 2019

Last update: 19th May 2019

General

-you will not need to read or write Python code in the exam, but may be asked to read or write pseudo code.

Lecture 1: Introduction

-be able to explain what data wrangling is, what activities it encompasses, why it is done, why it is challenging, why it is useful. Appreciate the series of activities in data wrangling (the data wrangling pipeline). These issues are covered in [this paper](#) and [this post](#).

Lectures 2 and 3: Data formats

-Appreciate the role that relational databases play in data wrangling.
-Detailed knowledge of relational databases is not required (i.e. you do not need to know syntax of SQL, relational database system architectures, relational database system internals)
-be able to understand a regular expression using the operators

. ^ \$ * + | []

-be able to formulate a regular expression using the above operators, based on an english description

-be able to explain what is a csv file, what is a spreadsheet, what is the difference?

-be able to explain the motivation for XML and XML namespaces

-be able to explain the differences between XML and HTML

-be able to explain the difference between XML attributes and elements and describe situations in which the use of one is preferred over the other

-be able to create XML documents, based on a natural language specification

-be able to both create and understand XML documents that use XML namespace syntax

-be able to explain the purpose of XML namespaces and list reasons for why it is useful

-understand what is mean by well-formed XML and valid XML

-be able to explain the difference between XML and JSON and applications where each is suited

-be able to read and create documents using JSON

-be able to convert an XML document to JSON and vice versa

-be able to explain the purpose of using schemas for XML and JSON data

-it is not necessary to know syntax of XML Schema or JSON Schema

Workshop Week 1

This was a warm up workshop on using Python and Jupyter notebook. It is not relevant for the exam.

Lectures 4 and 5: Data cleaning: missing values and outliers detection

- be able to explain the need for and the motivation behind data preprocessing and data cleaning
- be able to explain the need for and what is involved in each of the major data pre-processing activities (data cleaning, integration, reduction and transformation)
- understand the terminologies: features, attributes, instances, objects.
- understand the difference between categorical/discrete features versus continuous features
- be able to explain the reasons why data might be missing, what are the possible causes?
- understand the difference between data missing completely at random versus data missing not completely at random
- understand what is meant by noise in data or "noisy" data
- understand the following strategies for handling missing data and their relative advantages/disadvantages (delete all instances with a missing value, manual correction, imputation)
- understand the following strategies for imputation of missing values and their relative advantages/disadvantages (fill in with zeros, fill in with mean/median value, fill in with category mean)
- be able to explain the importance of finding outliers and give concrete examples where this would be useful
- be able to explain what is an outlier
- be able to explain the difference between a global outlier and a contextual outlier
- be able to draw and read a 2-D scatter plot and visually identify outliers from it
- be able to explain how a histogram can be used to detect outliers, their relative advantages/disadvantages for this task and be able to construct and interpret a histogram for data understanding and outlier detection
- be able to construct and interpret a Tukey Boxplot and explain why it is a useful tool for data understanding and outlier detection
- it is not necessary to know Grubb's test for outlier detection

Workshop Week 2

-the material from this workshop is not examinable. Its purpose was to assist you in your project work.

Lecture 5: Recommender systems

- understand what is a recommender system
- understand why missing data is an important issue for recommender systems
- understand what is collaborative filtering and the difference between user based and item based collaborative filtering
- understand the difference between i) user based methods for collaborative filtering and ii)

item based methods for collaborative filtering

- understand how to measure user-user similarity via transformation of the Euclidean distance
- when performing user-user similarity, understand how to select neighbors and make a prediction of the missing item
- understand how to predict missing ratings for an item using item-item similarity.
- appreciate the difference between the online and offline phases for item based collaborative filtering
- the material on matrix factorisation does not need to be known

Lecture 6: Data visualisation

- be able to explain the motivation for data visualisation
- be able to draw and interpret a 2-D scatter plot
- be able to draw and interpret a histogram
- be able to interpret a heat map visualisation of a dataset
- understand the advantages and disadvantages of using parallel coordinates to visualise a dataset
- be able to interpret a parallel coordinates plot and understand why the ordering of the feature axes is important
- understand why it can be useful to normalise each feature into the range $[0, 1]$ before computing Euclidean distance between vectors

Workshop Week 3

- this workshop provided some practical experience in using Python for processing HTML, XML and JSON. The material that needs to be known is listed under lectures 2 and 3.

Lecture 7: Clustering and clustering visualisation

- be able to explain why it is useful to perform clustering on a dataset and understand the challenges involved
- understand the steps of the k-means algorithm
- be able to identify scenarios where the k-means algorithm may perform poorly
- be able to interpret a heat map visualisation of a dissimilarity matrix
- understand the steps (pseudo code) for reordering a dissimilarity matrix using the VAT algorithm
- understand why the VAT algorithm is useful and how to interpret a dissimilarity matrix that has been reordered using the VAT algorithm
- understand how VAT may be used to estimate the number of clusters in a dataset

Lecture 8: Hierarchical clustering and dimension reduction

- be able to explain the steps of (agglomerative) hierarchical clustering, using single linkage (min)
- understand how a hierarchical clustering corresponds to a tree structure (dendrogram)

- be able to explain why knowledge of clustering can assist in outlier detection
- understand the concept of a dissimilarity matrix and the steps for its construction
- appreciate the relative advantages and disadvantages of k-means versus hierarchical clustering
- appreciate the following motivations for dimensionality reduction i) reduce amount of time and memory required by data processing algorithms, ii) allow data to be more easily visualised, iii) help eliminate irrelevant features or noise
- understand the concept of dimensionality reduction of a dataset (what is the input and what is the output and what is their relationship)
- understand that dimensionality reduction may be performed by i) selecting a subset of the original features or ii) generating a small number of new features
- understand the purpose of using PCA for dimensionality reduction. Understand the potential benefits of using PCA for data visualisation
- understand the intuition of how PCA works. It is not necessary to understand the mathematical formulas used for PCA.
- Understand intuitively what is meant by "First Principal Component" and "Second Principal Component"
- understand how to interpret a 2-D visualisation of the first two principal components of a dataset

Workshop Week 4

- given a list of numbers, be able to compute the mean, median, first quartile, third quartile, interquartile range
- be able to construct and interpret a Tukey Boxplot and explain why it is a useful tool for data understanding and outlier detection
- be able to compute the Euclidean distance between two vectors (tuples).
- be able to explain scenarios where recommender systems i) are likely to make incorrect recommendations, ii) over-recommend certain items, ii) under-recommend certain items

Lecture 9: Assessing correlations

- be able to explain why identifying correlations is useful for data wrangling/analysis
- understand what is correlation between a pair of features
- understand how correlation can be identified using visualisation
- understand the concept of a linear relation, versus a non linear relation for a pair of features
- understand why the concept of correlation is important, where it is used and understand why correlation is not the same as causation
- understand the use of Euclidean distance for computing correlation between two features and its advantages/ disadvantages
- understand the use of Pearson correlation coefficient for computing correlation between two features and its advantages/ disadvantages
- understand the meaning of the variables in the Pearson correlation coefficient formula and how they can be calculated. Be able to compute this coefficient on a simple pair of features. The formula for this coefficient will be provided on the exam.

- be able to interpret the meaning of a computed Pearson correlation coefficient
- understand the advantages and disadvantages of using the Pearson correlation coefficient for assessing the degree of relationship between two features

Lecture 10: Mutual information

- understand the meaning of the variables in the (normalised) mutual information and how they can be calculated. Be able to compute this measure on a pair of features. The formula for (normalised) mutual information will be provided on the exam.
- understand the role of data discretization in computing (normalised) mutual information
- understand the meaning of the entropy of a random variable and how to interpret an entropy value. Understand its extension to conditional entropy
- be able to interpret the meaning of the (normalised) mutual information between two variables
- understand the use of (normalised) mutual information for computing correlation of some feature with a class feature and why this is useful. Understand how this provides a ranking of features, according to their predictiveness of the class
- understand that normalised mutual information can be used to provide a more interpretable measure of correlation than mutual information. The formula for normalised mutual information will be provided on the exam
- understand the advantages and disadvantages of using (normalised) mutual information for computing correlation between a pair of features. Understand the main differences between this and Pearson correlation.

Workshop Week 5

-This workshop provided some practical experience in using Python for visualisation and was intended to provide a basis for the project work. The material that needs to be known is listed under lecture 6.

Lecture 11

This was a guest lecture and is not examinable.

Lectures 12 and 13: Classification and regression techniques: decision tree and k-nearest neighbor

- understand what is meant by the task of classification and be able to identify scenarios where it is useful. Understand how it relates to data wrangling.
- understand what is meant by the task of regression and be able to identify scenarios where it is useful. Understand how it relates to data wrangling.
- understand the use of accuracy as a metric for measuring the performance of a classification method. Understand how TP,TN,FP and FN are used in the accuracy calculation. The formula for accuracy will be provided on the exam
- understand the operation and rationale of the k nearest neighbor algorithm for classification
- understand the operation and rationale of the decision tree algorithm for classification

- Understand how a decision tree may be used to make predictions about the class of a test instance
- Understand the key steps in building a decision tree. How to split the instances, how to specify the attribute test condition, how to determine the best split and how to decide when to stop splitting
- understand the advantages and disadvantages of using k nearest neighbor or decision tree for classification and the reasons for these advantages and disadvantages
- understand the use of entropy as a node impurity measure for decision tree node splitting. Understand the benefits of entropy for this task and why it is effective for assessing the goodness of a split
- appreciate why it is necessary to separate the dataset into training and testing in order to fairly evaluate the performance of a classifier
- appreciate that when distribution of testing data is quite different from distribution of training data, accuracy of classifier may be degraded
- appreciate the benefits and limitations of accuracy as a performance metric for classification

Workshop Week 6

- be able to execute the k-means algorithm and the single-linkage agglomerative hierarchical clustering algorithm, given a simple input dataset.
- understand the concept of SSE, why it is a useful criterion for measuring the quality of a clustering and what are its limitations
- understand how knowledge of SSE can help choose the number of clusters in a dataset
- This workshop provided practice with using PCA and VAT. Required knowledge about these is listed under Lectures 7-8

Lecture 14

No lecture due to Good Friday holiday.

Workshop Week 7

- be able to discretize a feature using equal frequency discretization
- the remaining questions used Pearson correlation and mutual information and the required material is listed under lectures 9 and 10.

Lecture 15-16: Blockchain and data processing

- understand the motivation for blockchain technology
- understand what problems it is trying to solve?
- understand what are the benefits of blockchains
- understand how are blocks chained together in the blockchain and what is the benefit of this chaining?
- understand why it is virtually impossible to modify content within blocks that are part of the blockchain and what are the advantages and disadvantages of this modification property

- understand how hashing is used to identify and link blocks
- understand how digital signatures be used to verify data on the blockchain
- understand how hashing can be used to make information on the blockchain private
- understand how blockchain might be used to deliver benefits in health and education applications

Workshop Week 8

- understand the difference between classification and regression
- understand the difference between training data and testing data for evaluating the performance of a classifier
- be able to compute the information gain of a feature with respect to a class label, given a simple dataset.
- understand that the information gain of a feature, with respect to a class, is equal to the mutual information between the class and the feature.
- understand why it is necessary to avoid using the test data (directly or indirectly), when constructing a model such as decision tree
- it is not necessary to know the material about decision trees and missing values
- it is not necessary to know the details of python libraries for decision trees

Lecture 17

This was a guest lecture and is not examinable.

Lectures 18-20: Data linkage and privacy

- understanding what the record (data) linkage problem is
- understand why matching a database against itself can be regarded as a data linkage task
- be able to explain where record linkage is applied and why
- appreciate why record linkage can be difficult
- be able to describe the methodology of blocking, as applied to record linkage, explain why it is useful and why it is challenging
- understand the record linkage pipeline of preparation, blocking, scoring, matching and merging
- be able to explain in what circumstances privacy is an important issue for data linkage
- understand the objective of privacy preserving data linkage
- understand the use of one way hashing for exact matching in a 2 party privacy preserving data linkage protocol
- understand the vulnerabilities of 2 party privacy preserving data linkage protocol to i) small differences in input, ii) dictionary attack
- understand the operation of the 3 party protocol for privacy preserving linkage, using hash encoding with salt for exact matching. Understand the disadvantages of this protocol
- understand when and in what circumstances dictionary attacks and frequency attacks might be mounted against 2 party and 3 party data linkage protocols
- understand the steps of the 3 party protocol for privacy preserving data linkage with approximate matching (using Bloom filters)

- understand how to compute similarity of two strings based on 2-grams and why this method is useful
- understand how a Bloom filter works (how it is created, how strings are inserted, how strings are checked for membership)
- understand how a Bloom filter provides a private representation for strings
- understand how Bloom filters can be used to compare two strings for approximate similarity and the formula for doing this (Dice similarity coefficient)

Workshop Week 9

- appreciate the organisation of blocks in a blockchain.
- In a blockchain, understand how hashes of the header and the data are used to make later blocks dependent on earlier blocks and why modifying earlier blocks necessitates the modification of later blocks
- Understand how a digital signature can be used to verify the validity of data.
- Understand the role that public and private keys play in authenticating digital signatures
- Appreciate the benefits, challenges and required considerations of using blockchain for an application such as the Running Chickens example
- The mathematical details of how public/private keys are created and the mathematics (elliptic curves) justifying their use do not to be known

Lecture 21: Public data release and individual anonymity

- appreciate that an individual's privacy may be uncovered by linking to an external dataset or by use of background or external knowledge
- understand the terms sensitive attribute, non-sensitive attribute, quasi identifier
- understand the notions of k-anonymity and l-diversity and how they protect privacy
- understand how k-anonymity is susceptible to two types of privacy attacks (background attack and homogeneity attack)
- be able to determine whether an example table satisfies a given level of k-anonymity or l-diversity
- understand the benefits of using and sharing people's location data
- understand the possible privacy concerns of people's location data being shared
- appreciate the tradeoff between privacy and utility when using location based services or when analysing location based data

Workshop Week 10

- be able to perform simple calculations for the number of record comparisons needed using blocking
- appreciate the tradeoffs in designing a record linkage system, in terms of different types of errors and different blocking strategies
- be able to explain the relationship between record linkage and data duplicate detection and the required material is listed under lecture 18

Lecture 22: Differential privacy

- for differential privacy, understand what information is being protected and what information is not being protected
- understand how to compute the global sensitivity G of counting queries
- understand the role of the privacy loss budget k
- understand the role of G and k in terms of how much noise is added to the true query result
- understand that noise is added to the real query answer and this noise-added result is what will be released to the user. Understand how this protects the privacy of an individual
- understand that the amount of noise added is dependent on the ratio G/k . Larger G allows for more noise to be added and smaller k allows for more noise added

Workshop Week 11

There were no workshop questions this week, due to oral presentations.

Lecture 23: Ethical considerations

- appreciate that there exist ethical considerations in the context of a data wrangling/data science/data analytics project
- be able to explain who are the stakeholders in big data analytics and what are their perspectives? Understand what the tensions are between their motivations and what is the degree of power balance between them.
- appreciate the difference between the following two perspectives on the definition of (big) data analytics
 - Perspective 1: The ability to collect, store, and process increasingly large and complex data sets from a variety of sources, into competitive advantage.
 - Perspective 2: Data is contributed, collected, extracted, exchanged, sold, shared, and processed for the purpose of predicting and modifying human behaviour in the production of economic or social value.
- appreciate the motivation for each of the 10 simple rules for responsible big data research. (you need not memorise this list, but should be able to comment on a rule if it is mentioned)

Lecture 24: Wrap up

This was the final lecture and did not introduce any further examinable material.

Workshop Week 12

- understand how a salt is used for 3 party protocol for privacy preserving linkage with exact matching
- understand the factors that result in a string being a false positive for a Bloom filter
- understand the derivation steps for the formula that gives the probability that a string is a false positive for a Bloom filter
- be able to explain the benefits of using Dice coefficient similarity for Bloom filter comparison, compared to using Hamming or Jaccard similarity
- appreciate the tradeoffs in designing a record linkage system, in terms of different types of

errors

- Be able to determine whether an example table satisfies a given level of k-anonymity or l-diversity
- Given a table and quasi identifiers, be able to suggest a generalisation that achieves a given level of k-anonymity
- Understand how the global sensitivity of counting queries may change according to the feature representation that is used for a dataset
- Understand how to compute the global sensitivity of a dataset for counting queries

Project

You may be asked general questions that assume knowledge of you having attempted the project.

Quizzes

You may be asked questions that assume knowledge of the quiz questions that were asked in lectures. For your convenience, a consolidated list of these is contained in the "Exam Resources" folder.

Exam

- the exam will be 2 hours in length with 15 minutes reading time. You will not be permitted to use a calculator in the exam.
- answer the question that was asked. Writing about something else won't attract any marks
- use your own words to explain answers, rather than repeating verbatim from lecture notes
- write legibly, unreadable answers cannot attract marks
- succinct answers using point form are preferred

The first two pages of the 2019-sm1 exam are reproduced below

The University of Melbourne

Semester One 2019 Exam

School: Computing and Information Systems

Subject Number: COMP20008

Subject Title: Elements of Data Processing

Exam Duration: 2 hours

Reading Time: 15 minutes

This paper has 8 pages

Authorised Materials:

No calculators may be used.

Instructions to Invigilators:

Supply students with standard script books.

Instructions to Students:

Answer all 14 questions. The maximum number of marks for this exam is 50.

Formulae and Notation

Euclidean distance: $d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$

Pearson's correlation coefficient: $r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$
 where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ and $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$

Entropy: $H(X) = -\sum_{i=1}^{|\mathcal{X}|} p_i \log_2 p_i$
 where p_i is the proportion of points in the i th category.

Conditional entropy: $H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x)$
 where \mathcal{X} is the set of all possible categories for X and $|\mathcal{X}|$ is the number of categories.

Mutual information:

$$MI(X, Y) = H(Y) - H(Y|X) = H(X) - H(X|Y)$$

Normalised mutual information:

$$NMI(X, Y) = \frac{MI(X, Y)}{\min(H(X), H(Y))}$$

Accuracy: $A = \frac{TP+TN}{TP+TN+FP+FN}$ where

TP is number of true positives

TN is number of true negatives

FP is number of false positives

FN is number of false negatives

Set Union: $X \cup Y$ is the union of sets X and Y (elements in X or in Y or in both X and Y)

Set intersection: $X \cap Y$ is the intersection of sets X and Y (elements in both X and Y)