JAMES COOK UNIVERSITY
AUSTRALIA

Practical 1
Introductory Review Questions

What are we doing?

Using what you learned from lectures (Lecture 1&2) and relevant reading materials, you will answer some review questions. These questions are for your self-review on topics covered. You will need to review lecture and reading materials or seek for other resources (e.g. Googling), in order to answer questions.

Submission:
You are required to submit one document containing your answers via the weekly-practical submission box (available on CP1407 LearnJCU)

Self-Review Questions

1. Briefly define the term “Data Science”, “Data Mining” and “Machine Learning”, and identify the difference between these terms.

   a) Data Science
   Data Science is the field of processes which extract data from the input or the insight of the user and giving out more data about the things that the user is willing to know about. It is a kind of approach which takes principles, facts, calculation and practices from different type of fields such as mathematics, statistics, artificial intelligence and computer engineering and helps businesses and scientists to analyze a large amount of raw data and find out the solution or a result or a prediction.

   b) Data Mining
   Data mining is the function which is based on the set of data which is collected by the data science and figuring out the algorithms, methods and patterns to come out with a solution or predict the upcoming factors. Data mining is a tool or a technique which help in predicting future trends and help in making decisions in businesses which are critical. Data Mining is a portion of data science which generate information mainly help in the industry of businesses and advanced analytics applications. Therefore, data mining is considered a powerful technology in businesses and modernized environments.

   c) Machine Learning
   Machine Learning is the process of a computer system which is adapting and learning on its own without the help or commands of humans by using algorithms and analyzing the pattern of data. We can also say that machine learning is a small branch of Artificial Intelligence which take action and

learn as a human based on the data which its intakes and improving the responses days and days on its own. Machine learning is a perfect computer system in improving the field of computer science. This due to the over flowing and expanding of data which will help in answering and solving businesses and complicated questions.

2. Broadly, there are two types of knowledge, shallow and deep. Shallow knowledge is simply what makes up a computer's response. If we can retrieve any answer by framing a data query (using SQL) from an existing database system (e.g. JCU student database system), the output result retrieved will constitute shallow knowledge about the data. For example, we may learn that Australian Stock Exchange generally follows the lead of Wall Street, but we wouldn't necessarily know why. Deep knowledge is the underlying reason behind such relationships. Hidden knowledge is the top layer of this deep knowledge, which normally a data mining technique can unveil. Data mining will not give us the causes or the significance, but it can point to various associations and links.

   Data query is about searching in the data when we know exactly what we are looking for. Example are:
   - A list of customers who used MasterCard to buy medicine from a pharmacy.
   - A list of employees who will reach retiring age next year.

   These are all in the domain of shallow knowledge, which can easily be obtained by simple data queries using, for instance, SQL. By contrast, let us consider the following examples:
   - Develop a profile of MasterCard holders who will take advantage of the forthcoming sale promotion at the pharmacy
   - Develop a list of employees who are likely to avail themselves of the voluntary early retirement scheme when they reach retirement age.

   These are examples of hidden knowledge whose answers cannot be obtained from data queries, although data mining techniques can unveil the information. Identify whether each of the following is a data query or a data mining task(s):

   a) A social worker is interested in learning about the proportion of males to females in the population of a particular region.

      This is data query because in order to learn about the proportion of males to females in the population of a particular region, the social worker will have to find out already existing raw data of how many male and females are in that region. And then calculate the proportion of males to females with already existing raw data. No new data is formed nor data mining is done here.

   b) A stock market analyst has been asked by his client to predict the future prices of 10 stocks three months in advance.

      This is a data mining process because the client will have to use a software which will mine data of the stock market from the last year or more and then analyse the stock market and them calculate the overcome with many algorithms, patterns and etc. Then the software will show some profitable stocks and prediction of the prices of the stocks. Therefore, in order to mine data, analyse and find out the price, to predict the future prices of 10 stocks three months in advance is a data mining process.

c) Do single men play more golf than married men?

This is a Data query process because it is just about finding raw data such as how many single men play golf and how many married men play golf and then compare the numbers of these men.

d) Determine the characteristics of a successful used car salesperson.

The following information is a data query because simple survey and interviews can identify the characteristics of a successful used car salesperson. It is just asked for the data which is already there rather than trying to make a conclusion based on the data.

e) Determine whether a credit card transaction is valid or fraudulent.

This kind of process is a data mining process because to know whether a credit card transaction is valid or fraudulent, analysing of a large amount of data is required and then identifying the patterns or algorithms to find out that it is suspicious or not.

3. Why is a fully automated data mining tool not desirable? Discuss the need for human intervention in the data mining process.

A fully automated data mining tool is possible as the machine learning is available nowadays but it is not desirable. This is due to the needs of human help in initial problem definition and transition from practical to mathematical problem. Humans must be involved in the process such as making the correct question to find out the optimal and efficient solution and transferring all the data to the computer system so that the data mining process can take place. In my point of view, currently, the data mining process need the intervention of human but after a century or more, I think every work of the data mining process can be carried out by the computer system only.

4. How can data mining help a business analyst?

Data mining process collects the reliable data of the users of the business and try to figure out the pattern or correlation of the users' interest. Data mining allows the businesses to understand their audiences and trending things and showing the pattern and useful ways to extract more value to the businesses. Data mining help businesses to know how to target the audiences and how to make profitable production and operational adjustments to get more users or audiences of the businesses. It also helps the businesses to make good and correct critical decisions.

5. Data mining is a powerful technology that can bring about positive benefits but it has also caused a certain degree of suspicion and concerns over ethical issues. Find suitable examples to highlight that such concerns are valid and reasonable.

   Data mining is a good concept when you can consistently boost or increase the server amount. But if not, it is bad due to the continuous flow of data can cause data servers to overflow and crash. To be able to mine data, you need computer and they have to be running for several hours. Since computers have to be handling huge amount of data, they are sometimes noisy. Even though, it is noisy and working several hours a day, there are data that are incomplete or distributed. Some data are sometimes inconclusive. There are still many data, that have finished or come in great form but most of them are complex. Even when we just take the best data from data mining, the performance is not that outstanding to give credit that is it perfectly good to be used. Therefore, Data mining is a powerful technology and everyone will agree with it but it is known that it is not completely trust-worthy or valid.

6. The main objectives of data mining can be broadly categorized into *classification, estimation, prediction* and *data description*.
   - Classification: Object are classified into one of a set of pre-defined classes. In order to do this, a classification model is built from a set of data examples. The accuracy of the classification of the model is then evaluated to give some degree of confidence to the result. Once a reliable classification model has been developed, it is then used to classify data records whose class outcomes are unknown.
   - Estimation: Instead of classifying an object into a discrete class, this task involves building a model (based on a set of data examples) to estimate the value of a continuous outcome variable.
   - Data Description: This task is about describing general or specific features of the selected data set. It includes summary statistics, clustering and characteristic rule mining.
   - Prediction: It overlaps significantly with the classification and estimation, but is more concerned with a future outcome of the output variable. For instance, historical data recordings on weather conditions are used to predict tomorrow's weather. Solutions for classification and estimation are widely used for prediction too.

   Categorize each of the following data mining activities as classification, estimation or description. State clearly the reason behind your decision. Can any patterns discovered be used for prediction purposes?

   a) A real-estate agency has accumulated a large number of property sale records. The properties can be studio flats, semi-detached houses, detached houses or mansion houses. The agency wants to investigate from the data set what kinds of customer are likely to purchase which types of property.

The following data mining activities is a classification because the agency classified the sale records into the studio flats, semi-detached houses, detached houses or mansion houses. Then the agency tries to find out how customer like each classified houses.

b) It is interesting for the same real-estate agency to make significant links between descriptors of the properties sold and the characteristics of their customers. For instance, customers who are married with young children may be more likely to purchase a three-bedroom, detached house with a single garage.

   The following data mining activities is a description because the customers told how he would like his properties which is a three-bedroom, detached house with a single garage as he is married and has young children.

c) In recent years, we have seen increasing amounts of toxic waste dumped into our environment. Waste water from manufacturing processes, farming land run-off and sewage water from treatment plants have broken the chemical balance of the water in our rivers. The organic matter in the water has resulted in excessive growth of algae, which in turn leads to a reduction of the oxygen level in the water. Causing the deaths of fish and other wild life. Therefore, environment agencies want to monitor closely the growth of algae in the rivers and lakes. One agency has collected water samples from a number of different sites and analysed them for various chemical substances. They have also collected algae samples at the same locations to determine the population distributions of different algae. The agency wants to use the sample data to build a model that can approximate the distribution of algae population based on amounts of the chemical substances.

   The following data mining activities is an estimation because the agency want to know the data of amounts of toxic waste so that the agency would be able to build a model that is estimate the distribution of algae population based on amounts of the chemical substances.

Laboratory Questions

1. Visit KDnuggets website and try to explore the site freely to get useful news and information in the field of Business Analytics, Data Mining, and Data Science. Try to find and list some practical applications of data mining tools. (Hint: you can refer various polls results provided by this site)

Business Analytics

a) Open Refine
   Open Refine is a software which is to clean everything before doing the process of analysing. Cleaning here does not mean deleting or getting rid of things. Cleaning means like fixing the misspelled words, solving the errors

such as spacing, capitalization and etc. Instead of maintaining a bunch of errors in a huge amount of data, Open Refine is consist of many algorithms and patterns which will make the messy work cleared in a matter of time. This can save more time and faster analytics in the field of business.

b) NodeXL and Google Fusion Table
NodeXL is a data analysis software of networks and relationships. It performs the same as the Google Fusion Table which is also the data analysis software provided by Google to manage data. NodeXL do the deep calculation and networking between two software such as the great friendship map seen in linkedin or Facebook connections. Less advance data analysis is taken care by the Google Fusion Table.

c) WolframAlpha
WolframAlpha is a hidden search engine and this is use to power Apple's Siri. This also known as the nerdy search engine because it makes a calculation with algorithm and patterns and find and provide data in detail to any search. It is mainly used for the businesses, because it can give out information required for the business as chart and graph. This search engine is well-known for high level pricing history, commodity information, and topic overviews.

Data mining

a) SAS Enterprise Miner
SAS Enterprise Miner is a mining software which performs the data mining process. This miner can collect data which help in the process of predicting and create descriptive models based on analytics. SAS Enterprise Miner is able to analyse many data, different data and complex data, and find out the similar detail and figure out the algorithm or the patterns. Then it also builds models to detect fraud, anticipate resource demands and minimize customer attrition.

b) Weka
Weka is a software which target the process of data mining. It is a collection of machine learning algorithms for data mining tasks. All kind of tools which concern with data mining such as visualization, regression, clustering, data preparation and etc. It is an open-source software with the intention of helping in collecting and organizing data. Then Weka store and manage data for farther calculation. This software is most used in businesses and workstations.

c) Rapid Miner
Rapid Miner is the enterprise-ready data science platform which have collected a lot of data from people to be able to use in farther calculation or prediction. The Rapid Miner is used mostly in the competitive places due to its high performance. It contains many data and it is always trying to get

more and more data. It is also an open-source. Businesses and workplaces use this software for their advantage in their role.

Data Science

a) Codecademy

Codecademy is a platform is a place where coding can be learned from. Why should this platform should be used rather than others is the fact that it is an online learning platform with courses and tutorials on programming, data science and web development. This is a target platform for beginners and who are trying to improve their coding skills. Python, Html, CSS, Java and more kind of coding can be learned from this platform. There are free lessons and when the coding gets advanced, there is a cost for it which is $34.99 a month, or $17.49/month if you pay annually.

b) Udemy

Udemy is a platform which people can freely upload their courses for coding. No only college, business or organizations can upload but literally anyone can upload video courses on this online platform. In this platform, you will be able to find different sorts of skills with a large about of data science skills and videos of data science. It did not end with just videos, there will be also quizzes and projects for people to learn more about the subject. All kind of data science course can be found in this platform. The price here is $15 to $30 for lifetime access to the course.

c) HackerRank

HackerRank is a website for challenges of problems in coding. It helps the users to improve their coding skills by making coding challenging and competitive. This website has many categories offering such as algorithms, data structures, mathematics, databases, and more. This is also a great place for people who want to learn python, SQL and other coding skills. This is an open source which means free to use but it is mainly designed for company usage not individual. Company use this website to look for people suitable for the place.

2. *UCI Repository* is one of popular web sites where provide a repository of databases, domain theories and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms.

   Visit the site and retrieve one data set. With this example dataset, imagine you are a data miner who uses this data set to process a data mining project to solve a real-world problem. Define the following:

   - What problem do you target to solve using this data set?

     Accelerometer data sets help in analysing the sports performance. I target to improve the performance of the player in a sport tournament using accelerometer data sets.

- What part of the data (which attributes of the data) would be used as input for your data mining model?

  Depending on the issue we're attempting to address, different characteristics of the accelerometer data might be used as input. Data from accelerometers typically comprises acceleration readings along three axes such as x, y, and z. The Euclidean norm of the acceleration values along the three axes, which can be used to calculate the magnitude of acceleration, and features derived from time-domain and frequency-domain signal analysis, such as mean, standard deviation, energy, entropy, and spectral density, are the most frequently used attributes used as input for data mining models. As additional contextual information, the model may also accept input variables related to the subject's demographics, geography, and the time of day.

- What data mining methods (e.g. classification, clustering, association rule mining etc.) can be applied?

  This data mining method can be applied which is
  Anomaly detection: Accelerometer data sets can be used to detect unusual or anomalous patterns in acceleration signals, which could indicate a fall or other abnormal activity. To identify anomalous patterns and algorithms in data, using these systems such as One-Class SVM, Local Outlier Factor, and Isolation Forest.

- What would be the output of this data mining process?

  The output of an anomaly detection algorithm would be a series of anomalous patterns in the acceleration signals, which could signify falls or other unusual actions. The output could be used to send alerts or notifications to caregivers or medical staff.