Practical 1

Introductory Review Questions

Self-Review Questions

data science is defined as the science of exploring datanature.

Data science is the science of studying data.

Data science is an integration of statistics, computing technology, and artificial intelligence (AI).
The purpose of data science is to solve scientific and business problems by extracting knowledge from data.

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

Data Science is the field of extracting insights and knowledge from data through various techniques and methods. It involves tasks such as data cleaning, data transformation, data analysis, and data visualization. The goal of Data Science is to help organizations make better decisions by providing actionable insights based on data.

Data Mining is a specific technique within Data Science that involves the automatic extraction of patterns and knowledge from large datasets. It uses various statistical and computational algorithms to identify relationships and patterns in data.

Machine Learning is a subset of Artificial Intelligence (AI) that involves the development of algorithms that can automatically learn patterns in data without being explicitly programmed. These algorithms are trained on historical data and use statistical models to identify patterns and make predictions or decisions. Machine Learning is used in a wide range of applications, such as image and speech recognition, natural language processing, and predictive modeling.

To summarize, Data Science is a broader field that encompasses Data Mining and Machine Learning. Data Mining is a specific technique within Data Science that involves the automatic extraction of patterns and knowledge from large datasets, while Machine Learning is a subset of AI that involves the development of algorithms that can automatically learn from data to make predictions or decisions.

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. —— data query

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

c) Do single men play more golf than married men? ——data query

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

e)Determine whether a credit card transaction is valid or fraudulent. ——data query Data mining

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

Fully automated data mining tools may not be ideal in practice, because they lack context, may not be able to detect errors, generate complex models, and may not be flexible enough to adapt to changing business needs or data requirements. Therefore, human intervention is important in the data mining process to ensure accurate, unbiased, and actionable results. Humans can provide context, perform quality control checks, interpret the results, and adjust the process as necessary to ensure that the results are relevant and useful.

4. How can data mining help a business analyst?

Data mining can help business analysts identify patterns and trends, segment customer types, detect fraud, develop predictive models, and improve business processes. By providing insight into complex data sets, data mining can help businesses make informed decisions and gain a competitive advantage.

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.

Here are two cases I found to be more representative:

1. Employment discrimination: Some companies are using data mining to analyze social media profiles and other personal information of job applicants to assess their suitability for the job. The practice could raise concerns about discrimination, for example, where companies could use data mining to analyze a candidate's social media posts and search history to determine their political beliefs, sexual orientation or religious affiliation. If it's screening out candidates who don't meet what the company needs, even if they're otherwise qualified for the job. This can lead to discrimination against certain groups of people and limit opportunities for different candidates.

2. Target's Pregnancy Prediction Model: Target uses data mining to analyze customers' purchasing histories and search patterns to identify customers who may be pregnant. They then sent targeted ads for baby products and maternity clothes to those customers. However, this practice raises concerns about privacy and the use of personal information without consent, and they may not be aware that their data is being used in this way.

In general, these examples of data mining may raise ethical questions. It's important for companies to be open about how they use personal data and ensure they don't discriminate against certain groups of people. Likewise, personal data requires the individual's consent to be used.

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.

Classification - This involves classifying clients based on their characteristics (income, family size, etc.) and the type of property they are likely to buy. The classification patterns can be used for predictive purposes.

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.

Description - This involves exploring the data and understanding the connections between variables. If the associations found are strong and have a temporal component, they can potentially be used for predictive purposes.

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.

Estimation - This involves estimating the relationship between chemicals and algal populations. The patterns found in this case can be used for predictive purposes, as the model can be used to estimate the algae population in new locations based on the levels of chemicals in the water.

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)

1. This article compiles the 38 top Python libraries for data science, data visualization & machine learning, as best determined

https://www.kdnuggets.com/2020/11/top-python-libraries-data-science-data-visualization-machine-learning.html

2. Data Repositories.

https://www.kdnuggets.com/datasets/index.html

3. Check out these 10 books that can help data scientists and aspiring data scientists learn machine learning today.

https://www.kdnuggets.com/2020/04/10-best-machine-learning-textbooks-data-scientists.html

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?
-What part of the data (which attributes of the data) would be used as input for your data mining model?
-What data mining methods (e.g. classification, clustering, association rule mining etc.) can be applied?
-What would be the output of this data mining process?

For this project, I chose the data set of the banking market:
https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

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

The problem I want to solve is to develop a classification model to predict whether a customer will subscribe to a fixed deposit based on the results of a marketing campaign.

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

Input attributes include age, job, marital status, education, default status, home loan, personal loan, contact communication type, month of last contact, duration of last contact, activity, previous contact, previous result , employment change rate, consumer price index, consumer confidence index, etc. The target attribute is whether the customer is subscribed to a term deposit (yes or no)

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

Choose an appropriate classification algorithm for the problem, so classification techniques are best suited for this problem since the goal is to predict a binary outcome (subscribe or not). Methods such as logistic regression, decision trees, random forests, support vector machines (SVM), k-nearest neighbors (kNN), and neural networks can be applied.

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

The output of the data mining process will be a model that can accurately predict whether a customer will subscribe to a term deposit based on the customer's profile and economic indicators. The model can help banks optimize marketing campaigns, allocate resources efficiently, and improve customer acquisition strategies.