UNIVERSITI SAINS MALAYSIA

Second Semester Examination
2014/2015 Academic Session

June 2015

CIT553 – Business Intelligence and Data Mining
[Kecerdasan Perniagaan dan Perlombongan Data]

Duration : 2 hours
[Masa: 2 jam]

INSTRUCTIONS TO CANDIDATE:
[ARAHAN KEPADA CALON:]

- Please ensure that this examination paper contains **FOUR** questions in **NINE** printed pages before you begin the examination.

  *[Sila pastikan bahawa kertas peperiksaan ini mengandungi EMPAT soalan di dalam **SEMBILAN** muka surat yang bercetak sebelum anda memulakan peperiksaan ini.]*

- Answer **ALL** questions.

  *[Jawab **SEMUA** soalan.]*

- You may answer the questions either in English or in Bahasa Malaysia.

  [Anda dibenarkan menjawab soalan sama ada dalam bahasa Inggeris atau bahasa Malaysia.]

- In the event of any discrepancies, the English version shall be used.

  [Sekiranya terdapat sebarang percanggahan pada soalan peperiksaan, versi bahasa Inggeris hendaklah diguna pakai.]

1. (a) What are the differences between "Business Analytics" (BA) and "Business Analysis"?

(40/100)

(b) List the MicroStrategy's classification of Business Analytics tools.

(20/100)

(c) What is a Dashboard and its purpose?

(20/100)

(d) What is a Scorecard?

(20/100)

2. (a) Business Performance Management (BPM) encompasses three key components, what are they?

(25/100)

(b) List the logical and physical design of Business Performance Management (BPM) architecture.

(25/100)

(c) What is meant by Six Sigma?

(25/100)

(d) A Balance Scorecard (BSC) is designed to overcome the limitations of systems that are financially focused. The term "balance" arises because the combined set of measures are supposed to include certain indicators, what are these indicators?

(25/100)

3. (a) Discuss (shortly) whether or not each of the following activities is a data miningtask.

(i) Dividing the customers of a company according to their profitability.

(ii) Predicting the outcomes of tossing a (fair) pair of dice.

(iii) Predicting the future stock price of a company using historical record.

(iv) Monitoring the heart rate of a patient for abnormalities.

(v) Extracting the frequencies of a sound wave.

(vi) Monitoring and predicting failures in a hydropower plant.

(30/100)

(b) List **four** (**4**) goals of dimensionality reduction techniques, such as PCA - for what are they used in practice?

(20/100)

(c) Assume that you have to mine association rules for a very large transaction database which contains 9,000,000 transactions. How could sampling be used to speed up association rule mining? Give a sketch of an approach which speeds up association rule mining which uses sampling!

(10/100)

(d) What does it mean if an attribute is irrelevant for a classification problem?

(5/100)

(e) (i) Assume that you own an online book store which sells books over the internet. How can your business benefit from data mining?

(15/100)

(ii) Provide an example of a business intelligence application that benefits from data mining solutions instead of OLAP queries. Be specific with the data mining method you recommend and justify your answer.

(20/100)

4. (a) (i) Discuss issues that are important to consider when employing a Decision Tree based classification algorithm.

(10/100)

(ii) Compare the pros and cons of decision tree and neural network classification methods.

(20/100)

(b) Consider the training examples shown in table for a binary classification problem.

Customer ID	Gender	Property Model	Shirt Size	Class
1	M	Bungalow	M	C0
2	M	Terrace	M	C0
3	M	Terrace	L	C0
4	M	Terrace	L	C0
5	M	Terrace	XL	C0
6	M	Terrace	S	C0
7	F	Terrace	XL	C0
8	F	Terrace	M	C0
9	F	Double-storey	M	C1
10	F	Double-storey	M	C1
11	M	Double-storey	S	C1
12	M	Double-storey	S	C1
13	F	Bungalow	L	C1
14	F	Bungalow	L	C1
15	F	Terrace	XL	C1
16	F	Terrace	S	C1

(i) Compute the Gini index for the overall collection of training examples.

(ii) Compute the Gini index for the Gender attribute.

(iii) Compute the Gini index for the Property Type attribute using multiway split.

(iv) Compute the Gini index for the Shirt Size attribute using multiway split.

(v) Which attribute is better, Gender, PropertyType, or Shirt Size?

(20/100)

(c) An Internet marketer is interested in segmenting Internet with a clustering tool using the input attributes – top ten search key words used, top ten URLs, recent ten online purchases (vendor, product, quantity, amount), Internet usage level, heaviest access hour, and heaviest access day of a week. Answer the following questions:

(i) Can we find users with different income level? Why or why not.

(15/100)

(ii) Can we expect to find clusters differentiated based on Internet usage level? Why or why not.

(15/100)

(d) (i) Compare the major distinction of Unsupervised and Semi-supervised approaches for anomaly detection.

(10/100)

(ii) Explain what de-identication of a data set in data mining is and its possible advantages and disadvantages.

(10/100)

1. (a) Apakah perbezaan di antara "Analitik Perniagaan" dengan "Analisis Perniagaan"?

(40/100)

(b) Senaraikan klasifikasi “MicroStrategy” atas peralatan Analitik Perniagaan.

(20/100)

(c) Apakah yang dimaksudkan dengan papan pemuka (Dashboard) dan apakah tujuannya?

(20/100)

(d) Apakah yang dimaksudkan dengan kad skor (Scorecard)?

(20/100)

2. (a) Pengurusan Prestasi Perniagaan merangkumi tiga komponen utama, apakah komponen-komponen itu ?

(25/100)

(b) Senaraikan reka bentuk logik dan fizikal seni bina Pengurusan Prestasi Perniagaan (BPM).

(25/100)

(c) Apakah yang dimaksudkan dengan Enam Sigma?

(25/100)

(d) Kad Skor Seimbang (Balance Scorecard) direka untuk mengatasi keterbatasan sistem yang berfokus kewangan. Istilah "seimbang" timbul kerana set gabungan langkah-langkah yang sepatutnya mengandungi petunjuk tertentu. Apakah petunjuk-petunjuk itu?

(25/100)

3. (a) Bincangkan (dengan ringkas) sama ada setiap satu aktiviti berikut adalah perlombongan data atau tidak.

(i) Membahagikan pelanggan syarikat berdasarkan keuntungan diperoleh.

(ii) Meramalkan hasil melambung sepasang dadu yang adil.

(iii) Meramalkan harga saham masa depan syarikat dengan menggunakan rekod sejarah.

(iv) Pemantauan kadar jantung seseorang pesakit untuk keabnormalan.

(v) Mengekstrak frekuensi gelombang bunyi.

(vi) Memantau dan meramalkan kegagalan dalam loji kuasa hidro.

(30/100)

(b) Senaraikan **empat** (**4**) matlamat teknik pengurangan kematraan, seperti PCA - untuk apa yang mereka digunakan dalam amalan?

(20/100)

(c) Andaikan anda dikehendaki melombong peraturan persatuan lombong dari pangkalan data transaksi yang sangat besar iaitu 9,000,000 transaksi. Bagaimana pensampelan boleh digunakan untuk mempercepatkan perlombongan peraturan persatuan? Beri lakaran pendekatan yang mempercepatkan persatuan peraturan perlombongan yang menggunakan pensampelan.

(10/100)

(d) Apakah ertinya jika sesuatu sifat adalah tidak relevan dalam masalah pengelasan?

(5/100)

(e) (i) Andaikan anda memiliki sebuah kedai buku dalam talian yang menjual buku-buku di internet. Bagaimanakah perniagaan anda mendapat manfaat daripada perlombongan data?

(15/100)

(ii) Berikan satu contoh aplikasi perisikan perniagaan yang menerima manfaat penyelesaian perlombongan data dan bukannya pertanyaan OLAP. Kaedah perlombongan data yang disyorkan patut lebih spesifik dan wajarkan jawapan anda.

(20/100)

4. (a) (i) Bincangkan isu-isu yang penting untuk dipertimbang apabila menggunakan algoritma pengelasan berasaskan Pohon Berkeputusan.

(10/100)

(ii) Bandingkan kebaikan dan keburukan pohon keputusan dan kaedah klasifikasi rangkaian neural.

(20/100)

(b) Pertimbangkan contoh data latihan masalah pengelasan binary seperti ditunjukkan di dalam jadual.

ID Pelanggan	Jantina	Model Rumah	Saiz Kemeja	Kelas
1	L	Bungalow	M	C0
2	L	Teres	M	C0
3	L	Teres	L	C0
4	L	Teres	L	C0
5	L	Teres	XL	C0
6	L	Teres	S	C0
7	P	Teres	XL	C0
8	P	Teres	M	C0
9	P	Dua-tingkat	M	C1
10	P	Dua-tingkat	M	C1
11	L	Dua-tingkat	S	C1
12	L	Dua-tingkat	S	C1
13	P	Bungalow	L	C1
14	P	Bungalow	L	C1
15	P	Teres	XL	C1
16	P	Teres	S	C1

(i) Kira nilai index Gini bagi keseluruhan koleksi contoh data latihan.

(ii) Kira nilai index Gini bagi attribut Jantina.

(iii) Kira nilai index Gini bagi atribut Model Rumah dengan menggunakan pecahan pelbagai cara.

(iv) Kira nilai index Gini bagi atribut Saiz Kemeja dengan menggunakan pecahan pelbagai cara.

(v) Atribut mana yang lebih baik; Jantina, Model Rumah atau Saiz Kemeja.

(20/100)

(c) Pemasar Internet berminat menggunakan atribut input seperti; sepuluh carian utama yang digunakan, sepuluh URL teratas, sepuluh pembelian terbaru dalam talian (vendor, produk, kuantiti, harga), tahap penggunaan Internet, capaian jam tertinggi dan capaian harian dalam seminggu yang tertingggi untuk mengelaskan Internet menggunakan alat pengelompokan. Jawab soalan-soalan berikut:

(i) Bolehkah kita mencari pengguna yang mempunyai tahap pendapatan yang berbeza? Kenapa ya atau kenapa tidak?

(15/100)

(ii) Bolehkah kita mencari kelompok yang berbeza dari segi tahap penggunaan Internet? Kenapa ya atau kenapa tidak?

(15/100)

(d) (i) Bagi teknik pengesanan anomali, bandingkan perbezaan utama antara Tanpa Pegawasan dan Separa Pengawasan.

(10/100)

(ii) Bagi satu set data dalam perlombongan data, terangkan maksud penyah-pengenalan bagi satu set data.

(10/100)

- oooOooo -