UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR11007 DATA MINING AND EXPLORATION

Thursday 1st May 2014

14:30 to 16:30

INSTRUCTIONS TO CANDIDATES

Answer QUESTION 1 and ONE other question.

Question 1 is COMPULSORY.

All questions carry equal weight.

CALCULATORS MAY NOT BE USED IN THIS EXAMINATION

MSc Courses

Convener: B. Franke
External Examiners: A. Burns, S. Denham, P. Healey, T. Norman

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

1. **You MUST answer this question.**

(a) You are at the start of a project to assess the potential benefits of targeting different sales avenues (i.e. potential places or methods of selling goods) for a national clothing retailer who sells items through different stores, catalogues, online sources and discount selling to other retailers. The retailer would like to know what sales avenues might be the most worth targeting advertising for. The retailer is aware that targeting advertising toward one sales avenue also has knock on effects for sales in other sales avenues. The retailer has historic data regarding monthly sales from all outlets, as well as details of what adverts (text content/transcription) were run when (what month). The historic advertising is substantial and diverse, and relates to particular sales avenues. Describe the initial approaches you would consider in visualising the data for this problem. State at least two potential approaches you might consider for solving the problem and how your visualisations may help to assess their validity. [*7 marks*]

(b) What is the basic principle of projection pursuit? Outline a projection pursuit algorithm, and contrast it with Principal Component Analysis. [*5 marks*]

(c) Their are two types of visualization: visualization for data exploration and data visualization for presentation of results. State the two most important fundamental principles of each type of visualization. [*4 marks*]

(d) In the context of finding association rules, define the frequency (or support) of an itemset and the accuracy (or confidence) of an association rule. What does the APRIORI algorithm do, and how does it achieve this? [*6 marks*]

(e) Consider the sales data overleaf. Are the blank cells indicative of zero sales or missing data? Justify your belief with at least two reasons. [*3 marks*]

QUESTION CONTINUES ON NEXT PAGE

Monthly Sales by Country

Year-Month in yyyy-mm format

	1996-07	1996-08	1996-09	1996-10	1996-11	1996-12	1997-01	1997-02	1997-03	1997-04	1997-05	1997-06
Argentina	$3,489						$319	$443		$226	$110	
Austria	$3,598		$2,709		$15,682	$6,431	$9,119	$1,792	$439	$8,623	$4,180	$551
Belgium	$3,963	$4,782	$350		$814	$10,240	$5,978	$3,891	$2,218		$946	
Brazil				$5,141		$2,232	$12,856	$914	$2,578	$1,627	$1,947	$1,275
Canada				$353	$834	$1,766	$11,188	$1,078		$1,277		$3,105
Denmark	$347	$1,376		$1,393			$1,055	$3,077	$2,518		$2,698	$835
Finland	$2,270	$539	$1,810	$2,669	$10,011	$73	$4,106	$6,248	$3,048	$3,682	$3,630	$2,237
France	$6,905	$9,519	$2,552	$7,553	$6,888	$1,990	$3,009	$6,691	$2,230	$16,504	$19,797	$2,875
Germany			$4,407	$2,036		$2,680	$1,441			$4,430		$5,716
Ireland		$372	$608				$1,833	$517	$235			$2,519
Italy	$101	$1,269	$1,043	$1,303	$972				$1,249	$5,715	$1,941	$489
Mexico						$1,058				$200		$2,920
Norway						$459						
Poland				$1,453	$136	$717		$851	$2,427		$1,677	
Portugal		$242	$1,616	$982	$136		$338				$683	$155
Spain	$696	$2,102		$1,810		$2,326		$1,704			$5,415	$4,346
Sweden	$3,047					$1,118	$2,098			$2,314	$1,824	
Switzerland		$479	$517	$384	$4,902	$2,991	$3,063	$1,039	$3,137	$352	$5,529	$2,620
UK	$1,226	$3,392	$9,719	$10,789	$4,056	$8,924	$4,456	$6,698	$12,268	$6,096	$1,019	$6,719
USA	$2,221	$1,415	$1,051	$1,649	$1,168	$2,235	$400	$1,539	$6,200	$1,987	$2,385	
Venezuela												
Total	$27,862	$25,485	$26,381	$37,516	$45,600	$45,240	$61,258	$38,484	$38,547	$53,033	$53,781	$36,363

Historic data taken from http://www.fmsinc.com/microsoftaccess/query/crosstab-report/

2. (a) Describe qualitatively, but in detail, the TrueSkill ranking system. Given an example how the latent "skill" of a player can change under TrueSkill even while a player does not play any additional games. [*8 marks*]

   (b) What are the interpretation issues for receiver operator characteristic (ROC) curves in situations with a rare true positive class? [*2 marks*]

   (c) Describe the six parts of the Cross Industry Standard Process for Data Mining, that describe the phases of the full lifecycle of a data mining project. [*6 marks*]

   (d) On a data mining study for a client you discover a statistical discrepancy between two datasets that should have the same broad statistical properties. There is no obvious reason for this, but the most likely reason is some data collection bias on one or other study. You are not sure about the exact data collection methods used for each study. You need to determine which dataset is the most likely to be biased. What is the first procedure you would use to try to ascertain this? [*3 marks*]

   (e) Write down the simple Page Rank algorithm. Note a particular problem with this simple Page Rank, and give an improved method, showing how this alleviates the problem. Discuss how you could actually compute the Page Rank. [*6 marks*]

3. (a) Explain how Wikipedia can be used as a knowledge base for document clustering. [*7 marks*]

   (b) Describe hierarchical agglomerative clustering and three different set distance measures that can be used. [*5 marks*]

   (c) Suppose you needed to provide emergency resources for people, and wished to ensure people had sufficiently close geographic access to those resources. You wished to examine the budgetary costs associated with different distances to resources, and planned to use a hierarchical agglomerative clustering of geographic home locations as part of your cost benefit analysis. What between-set distance measure would you choose and why? [*4 marks*]

   (d) Explain the process the Support Vector Machine with a linear Kernel uses for handling data that is not linearly separable, and contrast that with how logistic regression achieves this. [*5 marks*]

   (e) Give an example of a situation where you would expect missing data to be missing at random. Give another example where you would expect missing data to not be missing at random. [*4 marks*]