CS9100

THE UNIVERSITY OF WARWICK

MSc/MEng Examinations: Summer 2016

CS910: Foundations of Data Analytics

Time allowed: 2 hours.

Answer **SEVEN** questions only: **ALL FOUR** from Section A and **THREE** from Section B.

Read carefully the instructions on the answer book and make sure that the particulars required are entered on **each** answer book.

Only calculators that are approved by the Department of Computer Science are allowed to be used during the examination.

Continued

Section A Answer **ALL** questions

1. Give a brief example to illustrate each of the following concepts: [7]
   (a) A binary attribute
   (b) A categoric attribute that is *not* ordered
   (c) A way to identify outliers in data
   (d) A constraint to handle missing values in data
   (e) A functional dependency in data
   (f) A way to reduce the dimensionality of data
   (g) A pair of highly correlated attributes

2. For each of the following unix tools, briefly describe its functionality: [6]
   (a) tail
   (b) grep
   (c) sed
   (d) gnuplot
   (e) wc
   (f) ls

3. For each of the following concepts from clustering, give a brief definition. [6]
   (a) Euclidean distance
   (b) Cluster diameter
   (c) Single link clustering
   (d) Cluster centroid
   (e) ϵ-neighbourhood of a point
   (f) Confusion matrix

4. For each of the following concepts from time series analysis give a brief explanation or example. [6]
   (a) Trends
   (b) Seasonality
   (c) Cyclicity
   (d) Irregular movements
   (e) False-alarm in change detection
   (f) Detection delay in change detection

Section B Choose **THREE** questions.

5. (a) Describe what a CDF plot of a numerical attribute shows. How can you use such a plot to find the median value? [4]

   (b) Describe how you could use a PDF plot to create a rule to identify possible outlier values. [3]

   (c) Explain what is a quantile-quantile plot and how it can be used to indicate whether two distributions are similar. [3]

   (d) Contrast the relative advantages and disadvantages of generating plots with a scripting tool such as gnuplot compared to an interactive tool such as a spreadsheet. [4]

   (e) Consider the following data set, of the exam performance of 100 students from three different courses. [11]

	Pass	Merit	Distinction
MEng	12	16	12
MSc-DA	18	8	4
MSc-CSA	20	6	4

   The module organizer wants to know if there is correlation in this data between which course a student is enrolled on and their performance in the module.

   Perform a chi-square test on this data to determine if there is a correlation at the 0.01 confidence level. A table of test statistics is provided below. Show the steps of your working.

Degrees of Freedom	1	2	3	4	5	6	7	8	9	10
Confidence level										
0.05	3.84	5.99	7.82	9.49	11.07	12.59	14.07	15.51	16.92	18.31
0.01	6.64	9.21	11.35	13.28	15.09	16.81	18.48	20.09	21.67	23.21
0.001	10.83	13.82	16.27	18.47	20.52	22.46	24.32	26.13	27.88	29.59

Continued

6. (a) Describe briefly how regression can be applied to data when the explanatory and dependent variables are as follows: [6]
   i. Explanatory: numeric, dependent: numeric
   ii. Explanatory: categoric, dependent: numeric
   iii. Explanatory: numeric, dependent: categoric

   (b) Given n paired observations of two dimensional data (x_i, y_i), show from first principles how to find the least squares solution for a in the model $y = ax^2$. Express your answer in terms of expectations, such as $\mathsf{E}[X]$, $\mathsf{E}[XY^2]$ etc. [8]

   (c) Consider the following data set: [6]

x	-2	-1	1	2
y	-2	-1	1	2

   i. What is the model of the form $y = ax^2$ that minimizes the squared error for this data?
   ii. The root-mean-squared error of the optimal model $y = ax^2$ can be written as $\sqrt{\mathsf{E}[Y^2] - \frac{\mathsf{E}^2[X^2Y]}{\mathsf{E}[X^4]}}$. Find the root-mean-squared error of this model on the above data set.
   iii. Explain why this model cannot achieve zero error on this data.

   (d) Write the expression to minimize when applying Tikhonov regularization (ridge regression) to the model $y = ax^2$ with parameter α. [5]
   What is the optimal value of a when α is 0?

Continued

7. (a) Give the expression for computing the Kappa statistic for a classifier. Explain the meaning of the terms in your expression. [5]

   (b) Consider the following data set with six examples, each with three binary attributes, and a binary class: [8]

X	Y	Z	Class
1	1	0	1
0	1	1	0
0	0	1	0
0	0	0	1
1	0	1	0
1	0	0	0

   Describe the 1R (OneR) classifier that is built on this dataset. What accuracy does it achieve when applied to the training data? How does this compare to the majority class (ZeroR) classifier for this dataset?

   (c) The model for linear regression and the model for support vector machines look quite similar. Given a data point x that is a vector with d dimensions, both use a set of $d+1$ parameters $w_0 \ldots w_d$ and compute the function [5]

$$w_0 + \sum_{i=1}^{d} w_i x_i$$

   Given these similarities, explain carefully how these two models are different.

   (d) For the dataset given above, the SVM model [7]

$$1 - 2X + 2Y - 6Z$$

   is found. Show that this meets the requirements for a valid (but not necessarily optimal) SVM model, and say what are the support vectors for it.

Continued

8. (a) Explain why it is not appropriate to treat the problem of recommending items to buy based on data on past purchases as an instance of classification. [6]

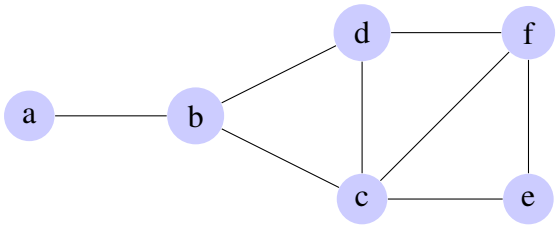
   (b) An instance of the neighborhood method for collaborative filtering computes the similarity of pairs of users' history based on the Pearson product-moment correlation coefficient. It finds the $k = 3$ most similar users, then takes the (unweighted) average of their scores for the target item. [8]

   Apply this method to the following set of binary preferences of six items to predict whether the User 1 would like Item 7 (the entry denoted by a question mark). Show your workings in detail.

Items	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
User 1	1	0	1	1	0	0	?
User 2	1	0	0	1	0	1	0
User 3	0	1	0	0	1	1	1
User 4	1	0	1	1	0	0	1
User 5	1	1	0	0	1	0	1
User 6	0	0	1	0	1	1	1
User 7	0	1	1	1	0	0	0

   (c) What features in graphs derived from social networks can be used to classify the nodes? [6]

   (d) Consider the following social network represented as a graph, where the nodes represent individuals, and edges represent a declared (symmetric) "friendship" relation between pairs. [5]

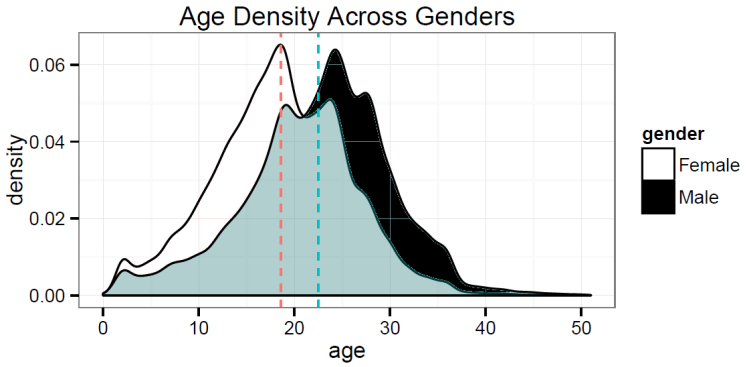


   Apply weighting based on number of common neighbours to determine which new edges are deemed most likely to occur. Explain your answer.

Continued

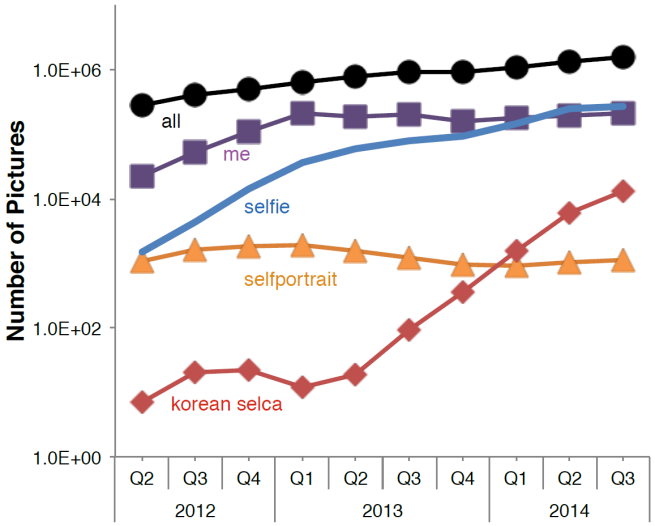
9. (a) Define the concept of "clustering coefficient" with a social network that is represented by an undirected graph. [4]

The below plot shows age distributions of users sharing "selfie" pictures on the social network site Instagram (it is taken from a recent study). The vertical dashed lines show the medians of the male and female age distributions (23 and 18, respectively), and the grey area shows the overlap of the two distributions.



(b) Give a brief summary of what you observe about the selfie-taking habits of Instagram users. You should make reference to properties of distributions such as median, mode, skewness and outliers. [6]

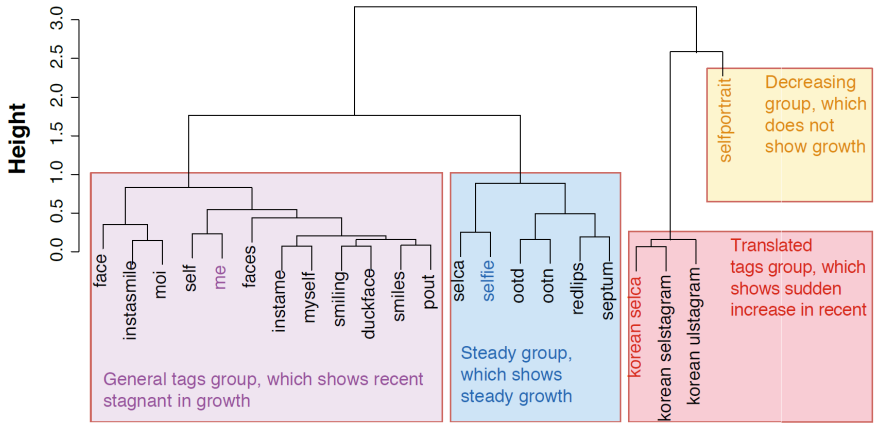
The next plot shows the popularity of various "tags" used to label photos associated with selfie pictures. These include the hashtags "me", "selfie", "self-portrait" and "korean selca" (abbreviation for self-camera), as well as all photos collected in the study.



(c) Describe in detail how you could fit a model to predict the popularity of the "korean selca" tag over 2015. How would you validate the quality of your model? [7]

The next diagram from the same study visualizes a hierarchical single-link clustering applied to different tags applied to selfies, where similarity is based on patterns of usage. The grouping shows a division into four clusters, and some comments on their patterns of behaviour.

Continued



(d) Under the distance function used in the clustering, which tag is most similar to "smiling"? Explain how you can be certain. [4]

(e) Suppose that the goal were to partition the tags into 6 clusters. Use the information in the diagram to describe what would be placed in the 6 clusters, and explain your answer. [4]

End