PAPER CODE NO.	EXAMINER : Dr. Danushka Bollegala
COMP527	DEPARTMENT : Computer Science Tel. No. 0151 7954283

UNIVERSITY OF LIVERPOOL

Resit Examinations 2015/16

Data Mining and Visualisation

TIME ALLOWED : Two and a Half Hours

INSTRUCTIONS TO CANDIDATES

Answer FOUR questions.

If you attempt to answer more questions than the required number of questions (in any section), the marks awarded for the excess questions answered will be discarded (starting with your lowest mark).

Question 1 Assume that you are given the dataset shown in Table 1 about heights (measured in centimeters) and weights (measured in kilograms) of a group of 5 students. Answer the following questions related to this dataset.

Table 1: Heights and weights of 10 students

Student no.	Height (cm)	Weight (kg)
1	169	60
2	171	70
3	150	80
4	180	75
5	150	60

A. Heights and weights are in different ranges. Propose a method to scale both those values into the same [0,1] range. **(2 marks)**

$\hat{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$

B. Using the method that you proposed in part **A** above scale the weights of the five students to [0,1] range **(5 marks)**

min = 60 and max = 80. Therefore, $s_1 = 0, s_2 = 0.5, s_3 = 1.0, s_4 = 0.75, s_5 = 0.$

C. Assuming that there exists a strong correlation between the height and the weight of a person, propose a method to detect potentially incorrect weight measurements. **(4 marks)**

Because there is a strong correlation between the height and the weight of a person, we can learn a regression model such as a liner regression model. Next, we can use trained regression model to predict the weight of each student given that student's height. If the predicted value of the weight differs significantly from the measured value of the weight, then it is likely that the weight measurement was incorrect.

D. Assume that you are planning to learn a naive Bayes classifier to predict whether a student is obese using their height and weight measurements. Will it be a problem if you accidentally took duplicate (repeated) measurements for some of the students? Explain your answer. **(4 marks)**

Yes. Duplicate data points will be a problem to the naive Bayes classifier because you would overestimate the counts for a paricular feature or a class.

E. Assume that you are planning to learn a support vector machine to predict whether a student is obese using their height and weight measurements. Will it be a problem if you accidentally took duplicate (repeated) measurements for some of the students? Explain your answer. **(4 marks)**

No. Duplicate instances will be represented by the same point in the feature space. Therefore, duplicates will not affect the classification hyperplane.

F. It turns out that 90% of the students are not obese. Why would classification accuracy be an inappropriate evaluation measure to evaluate the performance of the binary obese classifier we intend to learn from this dataset? **(3 marks)**

Because 90% of the students are not obese, by predicting non-obese for all the students (majority baseline) we will get a classification accuracy of 90%.

G. Suggest a better classifier evaluation measure for the scenario discussed in part **(F)** above. **(3 marks)**

Area under the ROC curve (AUC).

UNIVERSITY OF LIVERPOOL

Question 2 Assume we are given a training dataset consisting of four instances $(\boldsymbol{x}_1, +1), (\boldsymbol{x}_2, +1), (\boldsymbol{x}_3, -1)$, and $(\boldsymbol{x}_4, -1)$. Here, the $\boldsymbol{x}_1 = (1, 0)^\top$, $\boldsymbol{x}_2 = (0, 1)^\top$, $\boldsymbol{x}_3 = (-1, 0)^\top$, and $\boldsymbol{x}_4 = (0, -1)^\top$. We would like to train a binary k-nearest neighbour classifier from this dataset. Answer the following questions about this task.

A. Compute the Euclidean distance between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. **(3 marks)**

$\sqrt{(1-0)^2 + (0-1)^2} = \sqrt{2}$. *1 mark will be awarded for writing the formula for Euclidean distance even if the computation is incorrect.*

B. Compute the Manhattan distance between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. **(3 marks)**

$|1-0| + |0-1| = 2$. *1 mark will be awarded for writing the formula for Manhattan distance even if the computation is incorrect.*

C. Using Euclidean distance as the distance measure and $k = 1$, classify an instance $\boldsymbol{x}^* = (0.5, 0)$. Explain your answer. **(3 marks)**

The Euclidean distances between $\boldsymbol{x}^*$ and the four data points $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$, and $\boldsymbol{x}_4$ are respectively $0.5, \sqrt{5}/2, 1.5, \sqrt{5}/2$. Therefore, the label of the nearest neighbour of $\boldsymbol{x}^*$ will be 1,. Therefore, $\boldsymbol{x}^*$ will be classified as positive. The answers that do not show the reasoning will get only 1 mark.

D. Using Euclidean distance as the distance measure and $k = 3$, classify an instance $\boldsymbol{x}^* = (0.5, 0)$. Explain your answer. **(3 marks)**

The Euclidean distances between $\boldsymbol{x}^*$ and the four data points $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$, and $\boldsymbol{x}_4$ are respectively $0.5, \sqrt{5}/2, 1.5, \sqrt{5}/2$. Therefore, the labels of the nearest neighbours of $\boldsymbol{x}^*$ will be $1, 1, -1$. Therefore, $\boldsymbol{x}^*$ will be classified as positive. The answers that do not show the reasoning will get only 1 mark.

E. Propose a strategy for determining the label of $\boldsymbol{x}^* = (0.5, 0)$ when $k = 2$? **(4 marks)**

The Euclidean distances between $\boldsymbol{x}^*$ and the four data points $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$, and $\boldsymbol{x}_4$ are respectively $0.5, \sqrt{5}/2, 1.5, \sqrt{5}/2$. We have two sets of neighbours in this case $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ or $\{\boldsymbol{x}_1, \boldsymbol{x}_4\}$. The first set of neighbours would predict positve whereas the second will be a split. A possible strategy will be to consider only the sets with a unique prediction and ignore splits (predicts positive). Alternative solutions would be to guess randomly.

F. What is the set of positively predicted data points by a 2-nearest neighbour classifier for this dataset? **(3 marks)**

This will be the set of data points in the first quadrant. $\{(\alpha, \beta) | \alpha > 0, \beta > 0\}$. Marks will not be penalised for not specifying the boundary conditions.

G. Why would it be unwise to predict labels for test data using $k = 1$? **(2 marks)**

If the dataset is noisy and/or have outliers then the nearest neighbour only predictions are unreliable.

H. If you were given a large labeled train dataset, propose a method to determine the best value of k for the k-nearest neighbour classifier. **(4 marks)**

Methods that splits the train dataset into held-out vs. train portions and perform parameter tuning, or cross-validation will receive full marks.

Question 3 In a typical information retrieval system the following steps are conducted to produce an inverted index. A document is first tokenised using space character as the delimiter. Next, using a pre-defined list, stop words are removed. Next, unigram tokens are selected from the remaining tokens in a document for creating an inverted index. The inverted index stores the list of ids of documents (posting list) in which a particular unigram occurs.

When a user enter a query, it is tokenised using the same procedure, stop words removed, and unigrams are extracted. Next, each unigram is searched against the created inverted index to obtain the corresponding posting lists. Finally, the intersection of all posting lists are returned to the user as the search result. Answer the following questions about this search engine.

A. Explain what is meant by *stop word removal* in the context of text mining? **(2 marks)**

Remove non-content words such as articles using a pre-defined list of words. Correct answers will receive 2 marks.

B. Explain what is meant by a *unigram*. **(2 marks)**

A unigram is an individual token, e.g. cat.

C. Let us assume that a user entered the query *data mining*. The posting list for the token *data* consists of two documents (denoted by their ids) $\{d_1, d_{10}\}$, and the token *mining* consists of three documents (denoted by their ids) $\{d_{10}, d_{12}, d_{15}\}$. What would be the result for the query *data mining*? Explain your answer. **(4 marks)**

This will be d_{10} because we are considering the intersection of the posting lists. Answers without explanations will receive only 2 marks.

D. Using an example, explain how skip pointers can be used to speed up the computation of conjunctive (AND) queries. **(5 marks)**

We can put a pointer that indicates the value of the document id that we will encounter after a fixed number of skips. This pointer is called a skip pointer. For example, consider the two posting lists show below,

$$\textit{Brutus} \rightarrow 2[16] \rightarrow 4 \rightarrow 8 \rightarrow 16[28] \rightarrow 19 \rightarrow 23 \rightarrow 28$$

$$\textit{Caesar} \rightarrow 1[5] \rightarrow 2 \rightarrow 3 \rightarrow 5[51] \rightarrow 8 \rightarrow 41 \rightarrow 51$$

Here, the skip pointers are shown within square brackets. For example, when we are performing a linear search over the posting list for Caesar *when we have matched up to 8 and must match 41 next, we know that we can skip the four documents after 16 and move straight up to 28 because we will not find 41 before we reach 28.*

E. State an advantage of performing stop word removal in an information retrieval system. **(3 marks)**

The posting lists for stop words such as the, an, at, *etc. can be very large because those words occur virtually in all documents. Computing the intersections of long posting lists can be time consuming. By removing stop words prior to indexing the documents we can avoid this problem.*

F. The search engine described in this question does not take the word order in to consideration. Propose a solution to overcome this problem. **(3 marks)**

We can use bigrams in addition to unigrams to represent both documents and queries. For example, data mining *and* mining data *will be considered as different bigrams, hence different searchable units.*

G. Using examples describe what is meant by a *part-of-speech* in text mining? **(3 marks)**

Morphological categories of words such as nouns, verbs, adjectives, and adverbs are called part-of-speeches in text mining. An example of a noun would be cat. *Answers that do not provide examples will receive only 2 marks.*

H. Explain the difference between static and dynamic ranking as used in information retrieval. **(3 marks)**

Static ranking methods such as the PageRank does not take into account the query when ranking the documents. On the other hand, dynamic ranking methods will consider both the query and the documents when ranking the documents.

Question 4 Consider the six transactions shown in Table 2 related to four items a, b, c, and d. We would like to find frequent itemsets from the database shown in Table 2. Answer the following questions.

Table 2: Itemsets for six transactions in a database.

Transaction ID	Itemset
T1	b,d
T2	a,b,c,d
T3	a,c
T4	c,d
T5	b,c,d
T6	a,b

A. Using examples explain the difference between a substring and a subsequence. **(4 marks)**

Let is consider the string notebook. *Substrings are continuous sequences of characters of a string. In this example, we have* note, ote, tebo *etc. as substrings. On the other hand, a subsequence does not have to be continuous and can have missing characters as long as the ordering (as given in the original string) is preserved. For example, we have* n,t,k, n,o,b,k, n,o,o,k *as subsequnces. Answers that define the terms without examples will get a maximum of 2 marks.*

B. Why would it be unwise to generate all subsequences of strings in a database to find the most frequent subsequences? **(3 marks)**

The number of unique subsequences of a string with n unique letters is 2^n. Therefore, the number of candidate subsequences can grow very quickly, making it impossible to store the candidates in memory for finding the most frequent ones.

C. What is meant by the apriori property in sequential pattern mining? **(3 marks)**

If S is a large itemset (with minimum support t), then any subset of S is also a large itemset (with the same minimum support t).

D. Give that we are required to find itemsets with minimum frequency of 2, compute the minimum support for the database shown in Table 2. **(2 marks)**

Minimum support = minimum frequency / total transactions. Therefore, Minimum support = 2/6 = 0.33

E. Under the minimum frequency of 2, state all large itemsets of length 1 for the database shown in Table 2. **(4 marks)**

There are four itemsets with frequency 2 and length 1. They are ((a), (b), (c), (d)). Each itemset will receive one mark.

F. Under the minimum frequency of 2, state all large itemsets of length 2 for the database shown in Table 2. **(5 marks)**

There are 5 such itemsets as follows: ((a,b), (a,c), (b,c), (b,d), (c,d)). Note that (a,d) is small. Each itemset will receive 1 mark.

G. Under the minimum frequency of 2, state all large itemsets of length 3 for the database shown in Table 2. **(2 marks)**

There is only one such itemset which is (b,c,d).

H. Given a database containing a finite number of transactions, should the apriori algorithm always terminate? Explain your answer. **(2 marks)**

It will always terminate. During each iteration, the apriori algorithm grows the itemsets by adding a new item. In other words, the length of itemsets grow by 1 in each iteration. However, the total number of unique items will be finite given a finite set of transactions. Therefore, the apriori algorithm must terminate after a finite number of iterations. Correct answers without any justification will receive 1 mark. Additional 1 mark will be awarded for the justification.

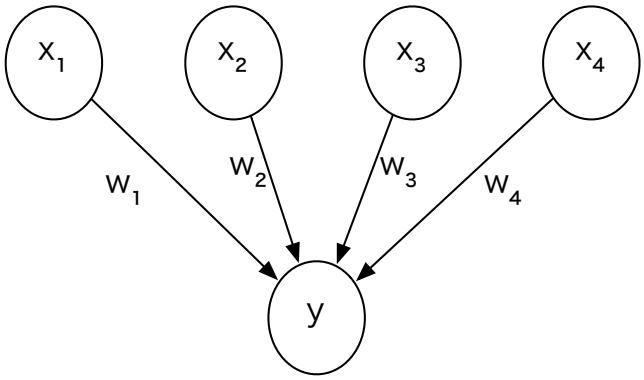

Figure 1: A single layer neural network.

Question 5 Figure 1 shows a neural network that receives 4 inputs x_1, x_2, x_3, and x_4, and multiplies inputs respectively by weights w_1, w_2, w_3, and w_4. The activation at the output node is y. Answer the following questions about this neural network.

A. State one advantage and one disadvantage of using a single layer neural network vs. multi-layer neural network. **(2 marks)**

A single layer neural network can only classify linearly separable data, whereas a multi-layer neural network has the capability to handle non-linear data. However, the number of edge weights grow rapidly with the number of layers used in the neural network making it (a) time consuming to train, (b) require large labeled datasets, and (c) likely to overfit to the train data.

B. Using the symbols specified in Figure 1, compute the activation at the output node y. **(4 marks)**

$y = x_1w_1 + x_2w_2 + x_3w_3 + x_4w_4$

C. Let us assume that the activation function at the output node is the logistic-sigmoid $\sigma(y)$ given by,

$$\sigma(y) = \frac{1}{1 + \exp(-y)}.$$

Compute the output value of the neural network. **(3 marks)**

$$\sigma(y) = \sigma(x_1w_1+x_2w_2+x_3w_3+x_4w_4) = \frac{1}{1 + \exp(-x_1w_1 - x_2w_2 - x_3w_3 - x_4w_4)}$$

D. Let us assume the target label for this input to be t. Assuming squared loss, compute the loss associated with the prediction made by the neural network for the given input. **(2 marks)**

$(t - \sigma(y))^2$

E. Using the inputs x_1, x_2, x_3, x_4 and the target label t, we would like to learn the optimal values for the weights w_1, w_2, w_3, and w_4. Show that the partial derivative of the loss ℓ with

respect to w_1 is given by,

$$\frac{\partial \ell}{\partial w_1} = -2(t - \sigma(y))\sigma(y)(1 - \sigma(y))x_1.$$

If required, you may use the fact that,

$$\frac{\partial \sigma(y)}{\partial y} = \sigma(y)(1 - \sigma(y)).$$

(8 marks)

$$\begin{aligned}
\frac{\partial \ell}{\partial w_1} &= \frac{\partial}{\partial w_1}(t - \sigma(y))^2 \\
&= -2(t - \sigma(y))\frac{\partial \sigma(y)}{\partial w_1} \\
&= -2(t - \sigma(y))\frac{\partial \sigma(y)}{\partial y}\frac{\partial y}{\partial w_1} \\
&= -2(t - \sigma(y))\sigma(y)(1 - \sigma(y))x_1
\end{aligned}$$

F. We would like to use stochastic gradient descent with a fixed learning rate η for the optimisation. If the current value of the weight w_1 is denoted by $w_1^{(k)}$, then show that updated value $w_1^{(k+1)}$ of the weight w_1 is given by:

$$w_1^{(k+1)} = w_1^{(k)} + 2\eta(t - \sigma(y))\sigma(y)(1 - \sigma(y))x_1$$

(4 marks)

Substitute the loss gradient in the stochastic gradient descent update rule to obtain the update equation for w_1.

G. Using the update equation given in part **F**, explain why we should scale the initial values of the weights such that the activation at the output node does not fall in the saturated regions of the logistic sigmoid function. **(2 marks)**

At the saturated regions of the logistic sigmoid we have either $\sigma(y) \to 0$ or $\sigma(y) \to 1$, for which the update becomes small (or zero). Therefore, the weight will get stuck in the initial values and will not get updated.