| 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).

UNIVERSITY OF LIVERPOOL

**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)**

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

**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)**

**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)**

**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)**

**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)**

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

**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)**

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

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

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

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

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

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

**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)**

UNIVERSITY OF LIVERPOOL

**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)**

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

**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)**

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

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

**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)**

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

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

UNIVERSITY OF LIVERPOOL

**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)**

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

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

**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)** .

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

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

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

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

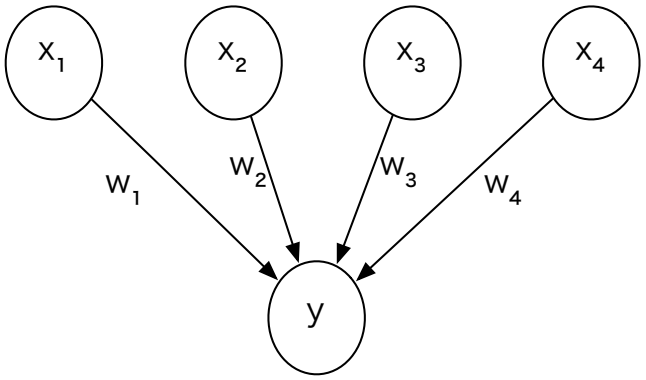


Figure 1: A single layer neural network.

**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)**

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

**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)**

**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)**

**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 $\ell$ 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)**

**F.** We would like to use stochastic gradient descent with a fixed learning rate $\eta$ 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)**

**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)**