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

UNIVERSITY OF LIVERPOOL

Second Semester 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

A. State the two main types of data mining models. **(2 marks)**

B. You are required to classify a given set of 100 flowers into three classes based on four attributes: sepal length, sepal width, petal length, and petal width. Answer the following questions about this experiment.

(a) State two algorithms that you can use to learn a three-class classifier for this purpose. **(2 marks)**

(b) Assume that 20 out of the 100 flowers in your dataset do not have their petal length measured. John suggests that you use ignore petal length as an attribute and use the remaining three attributes during training. State a problem that you might encounter if you follow John's suggestion. **(3 marks)**

(c) Instead of dropping petal length as an attribute, David suggests that you ignore the 20 flowers for which you do not have petal length measurements, and use the remaining 80 for training the classifier. State a problem that you might encounter if you follow David's suggestion. **(4 marks)**

(d) Instead of following John's or David's suggestions, Mary suggests that you set the missing petal length attribute values to zero, and use the entire 100 flowers for training the classifier. State a problem that you might encounter if you follow Mary's suggestion. **(4 marks)**

(e) Instead of replacing the missing petal lengths by zero as suggested by Mary, propose a better value. **(4 marks)**

(f) Assuming that there is a high positive correlation between sepal length and petal length, how can you exploit this information to overcome the missing value problem of petal lengths? **(4 marks)**

(g) Without having access to a separate test dataset, how can we evaluate the accuracy of our three-class classifier? **(2 marks)**

Question 2 Assume that we are required to cluster a given set of 100 news articles according to the news topics mentioned in those articles. We tokenise each document into a set of unigrams and remove stop words using a pre-defined stop words list. We then count the frequency of occurrence of a word in a document, and represent the document using a feature vector where each dimension corresponds to a particular word. The feature values are set to the frequency of occurrence of the corresponding word in the document. We ℓ_2 normalise each feature vector. We then measure the distance between two documents using the Euclidean distance between the respective feature vectors. Finally, we use k-means clustering to generate the document clusters. Answer the following questions related to this document clustering task.

A. Explain what is meant by a unigram as opposed to a bigram. **(2 marks)**

B. What is meant by *stop words* in text mining? **(2 marks)**

C. Given a vector $\boldsymbol{x} = (1, -1, 0)$, ℓ_2 normalise $\boldsymbol{x}$. **(3 marks)**

D. Why should we normalise the feature vectors representing documents before we compute Euclidean distance?. **(3 marks)**

E. State a problem of using frequency of occurrence of a word in a document as its feature value? **(4 marks)**

F. Propose a solution to overcome the problem you described in part **E** above. **(3 marks)**

G. Let us assume that we wanted to cluster the news articles into three clusters corresponding to political news, sports news, and foreign news. For this purpose let us assume that we ran k-means clustering with $k = 3$ but could not obtain three clusters covering the three categories as we wished. Propose a method to improve our chances of discovering clusters for the required categories using k-means. **(4 marks)**

H. Assuming that you are given a manually labeled set of news articles for the three categories mentioned in part **G**, explain a method to determine the optimal k value for the k-means clustering. **(4 marks)**

Question 3 Consider a training dataset consisting of four instances $(\boldsymbol{x}_1, 1)$, $(\boldsymbol{x}_2, -1)$, $(\boldsymbol{x}_3, 1)$ $(\boldsymbol{x}_4, -1)$ where $\boldsymbol{x}_1 = (1, 0)^\top$, $\boldsymbol{x}_2 = (0, 1)^\top$, $\boldsymbol{x}_3 = (-1, 0)^\top$, and $\boldsymbol{x}_4 = (0, -1)^\top$. Here, $\boldsymbol{x}^\top$ denotes the transpose of vector $\boldsymbol{x}$. We would like to train a binary Perceptron to classify the four instances in this dataset. For this question ignore the bias term b in the Perceptron and answer the following.

A. Write the perceptron update rule for a training instance $(\boldsymbol{x}, y)$ which is misclassified by the current weight vector $\boldsymbol{w}^{(k)}$. **(2 marks)**

B. Plot the four data points in the x-y plane. Is this dataset linearly separable? Justify your answer. **(3 marks)**

C. Let us predict an instance $\boldsymbol{x}$ to be positive if $\boldsymbol{w}^\top \boldsymbol{x} \geq 0$, and negative otherwise. Let us initialize $\boldsymbol{w} = (0, 0)^\top$. Compute the final value of the weight vector after presenting the training instances in the order $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_3$, and $\boldsymbol{x}_4$. **(6 marks)**

D. If we continue training the perceptron in the same order as in part **C** for multiple iterations over the dataset, will the weight vector ever converge to a fixed solution? Explain your answer. **(4 marks)**

E. If we present the four instances in the reverse order $\boldsymbol{x}_4$, $\boldsymbol{x}_3$, $\boldsymbol{x}_2$, $\boldsymbol{x}_1$, to the Perceptron, what would be the final value of weight vector at the end of the first iteration? **(6 marks)**

F. Now, let us re-assign the target labels for this dataset as follows $(\boldsymbol{x}_1, 1)$, $(\boldsymbol{x}_2, 1)$, $(\boldsymbol{x}_3, -1)$ $(\boldsymbol{x}_4, -1)$. Can we use Perceptron algorithm to linearly classify this revised dataset? Justify your answer. **(4 marks)**

Question 4 Assume that whether John would go to watch a football match depends on several factors such as the weather (being rainy, cloudy or sunny), temperature (being hot or cold), and traffic (high or less). Six past cases related to John's visits to football matches are shown in Table 1. Answer the following questions.

Table 1: John's past visits to football matches.

Weather	Temperature	Traffic	Go?
sunny	hot	less	yes
sunny	cold	high	no
cloudy	hot	high	no
rainy	cold	less	no
rainy	hot	less	yes
cloudy	cold	less	yes
sunny	hot	high	no
rainy	hot	high	no
cloudy	cold	high	no
sunny	cold	less	yes

A. State three problems that are frequently observed in rule-based classifiers. **(6 marks)**

B. Using the dataset shown in Table 1, compute the coverage and the accuracy of the rule,

IF Weather = sunny THEN Go = Yes

(6 marks)

C. Using Table 1 compute the conditional probabilities $P(Go = yes|Weather = sunny)$, $P(Go = yes|Weather = cloudy)$, and $P(Go = yes|Weather = rainy)$. **(6 marks)**

D. Use the Bayes' rule to compute $P(Weather = sunny|Go = yes)$. **(4 marks)**

E. Describe a method to overcome zero-probabilities when computing the likelihood of an event that can be decomposed into the product of a series of multiple independent events. **(3 marks)**

Question 5

A. Let us assume that we used some clustering algorithm to cluster a set 9 balls containing 2 red balls, 4 blue balls, and 3 green balls into 3 clusters as follows:
Cluster 1 = (red, red, blue)
Cluster 2 = (blue, blue, green)
Cluster 3 = (green, green, blue).
Answer the following questions about these clusters.

- **(a)** Following the majority labeling method determine the cluster labels. **(3 marks)**
- **(b)** Using the labels assigned in **(a)**, compute the precision of red, blue and green classes. **(3 marks)**
- **(c)** Using the labels assigned in **(a)**, compute the recall of red, blue and green classes. **(3 marks)**
- **(d)** Using the labels assigned in **(a)**, compute the macro-averaged precision and macro-averaged recall. **(2 marks)**
- **(e)** Using the labels assigned in **(a)**, compute the micro-averaged precision and micro-averaged recall. **(4 marks)**

B. Let us assume that we used a naive Bayes classifier for predicting whether a given email message is spam (positive class indicated by +1) or not (negative class indicated by -1). Table 2 shows the predicted probabilities $p(t = 1|\boldsymbol{x})$ for the positive class $t = 1$ for a given instance $\boldsymbol{x}$, and the correct labels when evaluated on a test dataset containing 10 email messages. Answer the following questions about this classifier.

Table 2: Predicted class probabilities and actual labels for 10 test instances.

$p(t = 1\|\boldsymbol{x})$	actual label
0.79	1
0.83	-1
0.63	1
0.43	-1
0.32	1
0.23	-1
0.43	1
0.93	-1
0.83	1
0.75	-1

- **(a)** If we set the classification threshold to be 0.5 (i.e. if $p(t = 1|\boldsymbol{x} > 0.5))$ then we predict $\boldsymbol{x}$ to be spam), compute the confusion matrix for this classifier. **(3 marks)**
- **(b)** Compute the classification accuracy under the threshold 0.5. **(2 marks)**
- **(c)** What would be the accuracy if we increase the threshold to 0.7? **(3 marks)**
- **(d)** Among the two spam classifiers obtained by setting the classification threshold to 0.5 and 0.7, which classifier would you prefer. Justify your answer. **(2 marks)**