PAPER CODE NO.	EXAMINER : Dr. Danushka Bollegala
COMP527	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).

UNIVERSITY OF LIVERPOOL

Question 1

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

Predictive models and descriptive models. Each point will be assigned 1 mark.

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)**

logistic regression, SVM, perceptron, etc.

(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)**

Petal length might be an important feature for this classification task. You might potentially loose accuracy of the classifier trained if you drop this attribute.

(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)**

You loose training instances if you follow this suggestion. We could overfit to the smaller training dataset if we were to fit many attributes on them, thereby potentially having poor test accuracy.

(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)**

A flower cannot have a zero length petal. Therefore, you have effectively introduced noise into your train data set by replacing the missing values to zero. This might reduce the accuracy of the classifier because the replacement value of zero is inconsistent with the distribution of values for the petal length.

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

We could replace the missing petal lengths to the average value for the 80 flowers for which we have petal lengths.

(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)**

We could learn a linear relationship between the two variables using a technique such as the linear regression and then use the learnt predictor to predict the missing values. We can then train a three-class classifier using this predicted data points as well as the original data points.

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

We can set aside a portion of the train data as held out data, and evaluate using that portion.

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)**

A unigram is a single word, whereas a bigram consists of consecutive two words.

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

Stop words are non-content features such as prepositions and articles. For example, the, an, what, etc.

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

$(1/\sqrt{2}, -1/\sqrt{2}, 0)$

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

Without normalising longer documents we will have vectors with large feature values compared to shorter documents under the representation method described in the question.

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

Common words such as functional words are likely to have higher co-occurrence frequencies, although they do not convey much information about the topics discussed in the document.

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

Instead of using term frequency use tfidf as the feature value.

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)**

Instead of randomly selecting the initial three documents (means), we can manually select three documents representing the three categories that we want to discover clusters for. We could further improve the probability of discovering the required categories by using centroids computed using multiple documents for each category as the initial cluster means.

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)**

We can use the manually labeled news articles as the gold standard and compute the B-cubed F-score for a given set of clusters. During B-cubed evaluation we will consider only the labeled data and ignore the unlabeled data in a cluster. We will vary the value of k and select the value that produces the highest F-score.

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)**

$\boldsymbol{w}^{(k+1)} = \boldsymbol{w}^{(k)} + y\boldsymbol{x}$

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

No. it is not. Answers that draw lines on the 2D plane for the extreme cases will receive full 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)**

When $\boldsymbol{w} = \mathbf{0}$, we have $\boldsymbol{w}^\top \boldsymbol{x}_1 = 0$. Hence, $\boldsymbol{x}_1$ is correctly predicted as positive. However, $\boldsymbol{w}^\top \boldsymbol{x}_2 = 0$, and $\boldsymbol{x}_2$ is misclassified. Therefore, using the update rule we will update the weight vector to $(0, 0) + (-1)(0, 1) = (0, -1)$. Next, for $\boldsymbol{x}_3$ we have $(0, -1)(-1, 0)^\top = 0$. Therefore, $\boldsymbol{x}_3$ is correctly classified. However, $\boldsymbol{x}_4$ will be misclassified because $(0, -1)^\top(0, -1) = 1 > 0$. Therefore, the final weight vector will be $(0, -1) + (-1)(0, -1) = (0, 0)$.

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)**

*No it will not. As seen from part **B**, after the first iteration the weight vector returns to its initial value. The dataset is not linearly separable and this is a case where perceptron does not converge on such a dataset.*

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)**

$(0, 0)(0, -1)^\top = 0$. Therefore, x_4 is misclassified. The weight vector will be updated to $(0, 1)$. Next, $(0, 1)(-1, 0)^\top = 0$ and x_3 is correctly classified. Next, $(0, 1)(0, 1)^\top = 1 > 0$ and x_2 is misclassified. The updated weight vector will be $(0, 2)$. Finally, $(0, 2)(1, 0)^\top = 0$ and x_1 is correctly classified. The final weight vector will be $(0, 2)$

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)**

Yes. The dataset is linearly separable now and and for a linearly separable dataset the perceptron will find a hyperplane that separates the two classes. Answers that either plots the data points in the 2D space or use some other method to show this will receive full 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)**

likely to overfit to the train data, can be time consuming when the dataset is large, too sensitive to the noise in the training data, cannot produce confidence scores. Each point will receive 2 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)

The rule covers 4 *out of the* 10 *cases. Therefore, its coverage is* $4/10$*. Out of those* 4 *matches,* 2 *cases have Go = YES. Therefore, the accuracy of the rule is* $2/4$*. Correct answers for coverage will receive 3 marks and the correct answers for accuracy will receive 3 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)**

$P(Go = yes|Weather = sunny) = 2/4$, $P(Go = yes|Weather = cloudy) = 1/3$, *and* $P(Go = yes|Weather = rainy) = 1/3$

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

$$\begin{aligned} & P(Weather = sunny|Go = yes) \\ = \; & P(Go = yes|Weather = sunny)P(Weather = sunny)/P(Go = yes) \\ = \; & (2/4) \times (4/10)/(4/10) = 0.5 \end{aligned}$$

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)**

Answers that describe Laplace smoothing or any other smoothing methods will receive full 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)**

Cluster 1 = red, Cluster 2 = blue, and Cluster 3 = green

(b) Using the labels assigned in **(a)**, compute the precision of red, blue and green classes. **(3 marks)**

Precision(red) = 2/3, Precision(blue) = 2/3, and Precision(green) = 2/3.

(c) Using the labels assigned in **(a)**, compute the recall of red, blue and green classes. **(3 marks)**

Recall(red) = 2/2, Recall(blue) = 2/4, and Recall(green) = 2/3.

(d) Using the labels assigned in **(a)**, compute the macro-averaged precision and macro-averaged recall. **(2 marks)**

Macro-averaged precision = $1/3 \times (2/3 + 2/3 + 2/3) = 2/3$. *Macro-averaged recall* = $1/3 \times (2/2 + 2/4 + 2/3) = 13/18$.

(e) Using the labels assigned in **(a)**, compute the micro-averaged precision and micro-averaged recall. **(4 marks)**

Let us denote the 2x2 contingency table (confusion matrix) for a label type in the following format: [(pred=true, actual=true), (pred=true, actual=false), (pred=false, actual=true), (pred=false, actual=false)]. Then, the contingency tables for the three label types will be, red = [2, 1, 0, 6], blue = [2, 1, 2, 4], green = [2, 1, 1, 5]. Adding the three tables element-wise we obtain, [6, 3, 3, 15]. Therefore, micro-averaged precision = 6 / 9, and micro-average recall = 6 / 9. One mark is awarded for computing the individual confusion matrices, one mark for adding them, one mark for computing precision, and one mark for computing recall.

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.

(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)**

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

The predicted labels under 0.5 threshold will be as shown below.
$p(t=1|\boldsymbol{x})$, actual label predicted label
0.79, 1 1, correct
0.83 , -1 1, incorrect
0.63 , 1 1, correct
0.43 , -1 -1, correct
0.32 , 1 -1, incorrect
0.23 , -1 -1, correct
0.43 , 1 -1, incorrect
0.93 , -1 1, incorrect
0.83 , 1 1, correct
0.75 , -1 1, incorrect
Therefore, the confusion matrix (predicted vs. actual) will be [3, 3, 2, 2]. 1 mark for predicting the correct labels and 2 marks for computing the confusion matrix.

(b) Compute the classification accuracy under the threshold 0.5. **(2 marks)**

accuracy = (3+2) / (3+3+2+2) = 0.5

(c) What would be the accuracy if we increase the threshold to 0.7? **(3 marks)**

The predicted labels under 0.7 threshold will be as shown below.
$p(t=1|\boldsymbol{x})$, actual label predicted label
0.79, 1 1, correct
0.83 , -1 1, incorrect
0.63 , 1 -1, correct
0.43 , -1 -1, incorrect
0.32 , 1 -1, incorrect
0.23 , -1 -1, correct
0.43 , 1 -1, incorrect
0.93 , -1 1, incorrect
0.83 , 1 1, correct
0.75 , -1 1, incorrect
Therefore, the confusion matrix (predicted vs. actual) will be [2, 3, 3, 2].
Classification accuracy is 4/10 = 0.4.

(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)**

Both classifiers have false positive rates of 3/5 = 0.6. However, 0.5 *threshold classifier marks 60% of emails as spam, whereas the* 0.7 *classifier marks only* 50% *of emails as spam. If you would like to avoid the case that accidentally a legitimate email being incorrectly classified as spam, you would prefer the classifier that marks lesser number of spam emails, which is the classifier with* 0.7 *threshold. An alternative answer would be to select the classifier with* 0.5 *threshold because it is more accurate. Both answers are considered correct if the relevant justification is provided. 1 mark for the answer and 1 mark for the justification.*