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

UNIVERSITY OF LIVERPOOL

Second Semester Examinations 2014/15

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)**

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

B. Consider that you measured the height and weight of 100 students for a health survey. For 20 students in your sample you could only measure either their height or weight, but not both values. Assume that we would like to train a binary classifier to predict whether a student is overweight compared to the students in this dataset. Answer the following questions about this experiment.

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

logistic regression, SVM, perceptron, etc.

(b) What is meant by the *missing-value* problem in data mining? **(3 marks)**

Some of the feature values (attributes) in the data might be missing because either the measurements were not taken and/or the data is corrupted.

(c) State two disadvantages we will encounter if we ignore the 20 instances that we have incomplete measurements for and use the remaining 80 instances to train the classifier. **(4 marks)**

The dataset size will be too small and we might overfit to it. The dataset size might be too small to learn anything useful (under fitting). The missing data points might contain useful information about the target task.

(d) The average height of the students in this dataset is 169cm. Provide a reason for and a reason against using the average to fill the missing values. **(4 marks)**

For: It is a typical value for the height of the students. Against: The 20 *students for which we do not have height measurements could be outliers.*

(e) Assume that we would like to check whether there is any correlation between the height and the weight of the students in this dataset. How do we check this? **(4 marks)**

We could measure the Pearson correlation coefficient between the height and the weight, and if it is high we could conclude that there is a high correlation between the two variables.

(f) Given that there is a high correlation between the height and the weight of a student, how can we use this information to overcome the missing-value problem? **(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 binary 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 binary 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 trying to learn a binary sentiment classifier from Amazon product reviews. Each review is assigned a rating (1-5 stars) by a user. We have 1000 such reviews for training purposes and a separate collection of 1000 reviews for testing. Answer the following questions about this experiment.

A. Define what is meant by unigrams and bigrams. **(2 marks)**

A unigram would be a single word, whereas a bigram would be two consecutive words.

B. Why would it be a good idea to use bigrams as well as unigrams to represent reviews in this task? **(3 marks)**

*Negations such as **not like** can only be captured using bigrams.*

C. Propose a method to assign binary target labels to this dataset such that we could train a binary sentiment classifier from it. **(3 marks)**

For example, we could assign positive labels to reviews that have 4 *or* 5 *ratings and negative labels to reviews that have* 1 *or* 2 *ratings. We could ignore reviews that a have rating value of* 3.

D. Assume that we trained a logistic regression classifier from this binary labeled dataset. How can we find out what features are most useful when predicting positive sentiment in Amazon reviews? **(4 marks)**

Sort the features in the descending order of their weights in the final weight vector. The top positive features are the ones that are most useful when predicting positive sentiment.

E. 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.

F. What effect would it have if we were to remove stop words in our sentiment classification task **(3 marks)**

It will reduce the dimensionality of the feature space thereby speeding up both the train and test stages.

G. Assume that our test dataset turns out to have 700 positive instances and 300 negative instances. What would be the classification accuracy of a random guessing algorithm on our test dataset? Explain your answer. **(4 marks)**

A random guesser will predict positive and negative classes with 0.5 probability. Therefore, it will predict 350 *out of the positive instances as positive and* 150 *out of the negative instances as negative. Therefore, the total number of correctly classified instances will be* $150 + 350 = 500$, *giving a classification accuracy of* $500/1000 = 50\%$.

H. For the unbalanced test dataset described in part **G**, what would be the accuracy obtained by a prediction algorithm that always predicts an instance to be positive? Explain your answer. **(4 marks)**

Because there are 700 *positive instances in the test dataset and all of those instances will be correctly classified by this predictor, we will have* $700/1000 = 70\%$ *accuracy.*

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 $\boldsymbol{x}_1 = (1, 1)^\top$, $\boldsymbol{x}_2 = (-1, 1)^\top$, $\boldsymbol{x}_3 = (-1, -1)^\top$, and $\boldsymbol{x}_4 = (1, -1)^\top$. 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. Let us predict an instance $\boldsymbol{x}$ to be positive if $\boldsymbol{w}^\top \boldsymbol{x} \geq 0$, and negative otherwise. Initializing $\boldsymbol{w} = (0, 0)^\top$, show that after observing $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_3$, and $\boldsymbol{x}_4$ in that order the weight vector will be $-\boldsymbol{x}_3 - \boldsymbol{x}_4$. **(6 marks)**

When $\boldsymbol{w} = \boldsymbol{0}$, we have $\boldsymbol{w}^\top \boldsymbol{x}_1 = 0$. Hence, $\boldsymbol{x}_1$ is correctly predicted as positive. Same applies for $\boldsymbol{x}_2$ as well. However, $\boldsymbol{x}_3$ will be misclassified and the weight vector will be updated to $\boldsymbol{w} = \boldsymbol{0} - \boldsymbol{x}_3 = \boldsymbol{x}_3$. Next, $-{\boldsymbol{x}_3}^\top \boldsymbol{x}_4 = 0$ and $\boldsymbol{x}_3$ will be classified incorrectly as positive. Therefore, $\boldsymbol{w} = -\boldsymbol{x}_3 - \boldsymbol{x}_4$.

B. If we present the four instances in the reverse order $(\boldsymbol{x}_4, -1)$, $(\boldsymbol{x}_3, -1)$, $(\boldsymbol{x}_2, 1)$, $(\boldsymbol{x}_1, 1)$, to the Perceptron, what would be the final value of weight vector at the end of the first iteration? **(4 marks)**

$-\boldsymbol{x}_4 - \boldsymbol{x}_3 + \boldsymbol{x}_2 + \boldsymbol{x}_1$

C. Normalize each of the four instances $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_3$, and $\boldsymbol{x}_4$ into unit L2 length. **(4 marks)**

All the normalized vectors will have a factor $\frac{1}{\sqrt{2}}$ in front.

D. What would be the final weight vector after observing the four instances if you used the L2 normalized training instances instead of the original (unnormalized) instances to train the Perceptron as you did in the part (A) of above? **(4 marks)**

$-\frac{1}{\sqrt{2}}(\boldsymbol{x}_3 + \boldsymbol{x}_4)$.

E. 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)**

No. The dataset is no longer linearly separable. Answers that either plots the data points in the 2D space or use some other method to show this will receive full marks. If no justification is given, then such answers will receive 2 marks.

F. Describe a method to learn a binary linear classifier for the revised dataset described in part (E) above. **(3 marks)**

Kernalized versions such as using the product of the two features as a third feature will receive full marks.

Question 4 Consider the dataset shown in Table 1 from which we would like to learn a classifier that could predict whether Play=yes using the four features *outlook*, *temperature*, *humidity*, and *windy*. Answer the following questions about this dataset.

Table 1: Weather dataset for decision tree learning.

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

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 Outlook = Sunny THEN Play = Yes

(6 marks)

The rule covers 5 *out of the* 14 *cases. Therefore, its coverage is* $5/14$*. Out of those* 5 *matches,* 2 *cases have PLAY = YES. Therefore, the accuracy of the rule is* $2/5$*. 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 $P(play = yes|outlook = sunny)$, $P(play = yes|outlook = overcast)$, and $P(play = yes|outlook = rainy)$. **(6 marks)**

$P(play = yes|outlook = sunny) = 2/5$, $P(play = yes|outlook = overcast) = 4/4$, *and* $P(play = yes|outlook = rainy) = 3/5$

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

$P(outlook = sunny|play = yes) = P(play = yes|outlook = sunny)P(outlook = sunny)/P(play = yes) = (2/5) \times (5/14) \times (14/9) = 2/9$

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 Big data sets and the availability of high performance computing resources such as GPUs, have given birth to the so called *Big Data Mining* era. By combining different datasets and performing pattern analysis across datasets, we can discover trends that were not previously possible to detect using small scale individual datasets. Big Data Mining has received much attention not only from the academia but also from the industry. Answer the following questions about Big Data Mining.

A. Explain three challenges we face when performing data mining on large datasets. **(12 marks)**

B. Propose a separate solution to each of the challenges that you described in the previous question (part A) **(13 marks)**

Some of the important challenges and their solutions are

(a) Resolving disambiguates when merging datasets (named entity resolution, word sense disambiguation)

(b) Privacy issues (Privacy Preserving Data Mining)

(c) Difficulties in loading large datasets in to memory to train classification/clustering algorithms (online learning, distributed ML)

(d) Ethical issues in data collection (anonymized data)

(e) Reliability issues (statistical confidence tests) Answers that elaborate on these lines will receive full marks.