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

UNIVERSITY OF LIVERPOOL

RESIT EXAMINATIONS 2017/18

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, the marks awarded for the excess questions answered will be discarded (starting with your lowest mark).

Question 1 Consider a dataset $\mathcal{D}$ of N instances, where each instance $x_i \in \mathcal{D}$ is represented by a three dimensional real-valued vector $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3})^\top$. Moreover, a label $t_i \in \{-1, 1\}$ is assigned to $\boldsymbol{x}_i$. We would like to learn a binary classifier using $\mathcal{D}$. However, for some instances, we do not have x_{i3} values measured. Answer the following questions.

A. Explain what is meant by the *missing value problem* in data mining. **(2 marks)**

If some features are missing (not measured, unobserved) for some data points in a dataset, then this is called the missing value problem.

B. Compute the ℓ_2 norm of $\boldsymbol{x}_i$. **(2 marks)**

$||\boldsymbol{x}_i||_2 = \sqrt{x_{i1}^2 + x_{i2}^2 + x_{i3}^2}$

C. Write the ℓ_2 normalised version of $\boldsymbol{x}_i$. **(2 marks)**

$\frac{\boldsymbol{x}_i}{||\boldsymbol{x}_i||_2}$

D. Compute the means μ_1, μ_2, μ_3 and standard deviations $\sigma_1, \sigma_2, \sigma_3$ for the three features in $\mathcal{D}$. **(6 marks)**

$$\mu_1 = \frac{1}{N}\sum_{n=1}^{N} x_{n1}$$

$$\sigma_1 = \sqrt{\frac{\sum_{n=1}^{N}(x_{n1} - \mu_1)^2}{N-1}}$$

$$\mu_2 = \frac{1}{N}\sum_{n=1}^{N} x_{n2}$$

$$\sigma_2 = \sqrt{\frac{\sum_{n=1}^{N}(x_{n2} - \mu_1)^2}{N-1}}$$

$$\mu_3 = \frac{1}{N}\sum_{n=1}^{N} x_{n3}$$

$$\sigma_3 = \sqrt{\frac{\sum_{n=1}^{N}(x_{n3} - \mu_1)^2}{N-1}}$$

E. Apply Gaussian scaling on $\boldsymbol{x}_i$. **(2 marks)**

$\left(\frac{x_{i1}-\mu_1}{\sigma_1}, \frac{x_{i2}-\mu_2}{\sigma_2}, \frac{x_{i3}-\mu_3}{\sigma_3}\right)$

F. Given that $\mu_3 = 0$ would it be problematic to replace missing values of x_{i3} to zero? Explain your answer. **(2 marks)**

Yes, this would be problematic because $x_{i3} \in \mathbb{R}$, if we replace missing x_{i3} values by zero we would not be able to distinguish among the instances for which x_{i3} was measured but turned out to be zero vs. instances where x_{i3} is missing.

G. As a solution to the missing value problem, we would like to predict x_{i3} using x_{i1} and x_{i2} using the linear relationship $\hat{x_{i3}} = ax_{i1} + bx_{i2} + c$, where $a, b, c \in \mathbb{R}$ are parameters that must be estimated from $\mathcal{D}$ and $\hat{x_{i3}}$ is the predicted value for x_{i3}. Write the squared loss for this prediction problem. **(3 marks)**

$E(\mathcal{D}) = \sum_{i=1}^{N}(ax_{i1} + bx_{i2} + c - x_{i3})^2$

H. Compute the gradient of the squared loss function w.r.t. a, b and c. **(3 marks)**

$$\frac{\partial E}{\partial a} = 2\sum_{i=1}^{N}(ax_{i1} + bx_{i2} + c)x_{i1}$$
$$\frac{\partial E}{\partial b} = 2\sum_{i=1}^{N}(ax_{i1} + bx_{i2} + c)x_{i2}$$
$$\frac{\partial E}{\partial c} = 2\sum_{i=1}^{N}(ax_{i1} + bx_{i2} + c)$$

I. Write the update rules for a, b and c using stochastic gradient descent. **(3 marks)**

$$a^{(k+1)} = a^{(k)} - 2\eta\sum_{i=1}^{N}(ax_{i1} + bx_{i2} + c)x_{i1}$$
$$b^{(k+1)} = b^{(k)} - 2\eta\sum_{i=1}^{N}(ax_{i1} + bx_{i2} + c)x_{i2}$$
$$c^{(k+1)} = c^{(k)} - 2\eta\sum_{i=1}^{N}(ax_{i1} + bx_{i2} + c)$$

Question 2 We would like to use the Perceptron algorithm to learn a linear classifier $y = \boldsymbol{w}^\top \boldsymbol{x} + b$, defined by a weight vector $\boldsymbol{w} \in \mathbb{R}^d$ and a bias $b \in \mathbb{R}$ from a training dataset consisting of three instances, $\{(t_n, \boldsymbol{x}_n)\}_{n=1}^3$. Here, $\boldsymbol{x}_1 = (0,0)^\top$, $\boldsymbol{x}_2 = (1,1)^\top$ and $\boldsymbol{x}_3 = (-1,1)^\top$, and the labels are $t_1 = 1$, $t_2 = -1$ and $t_3 = 1$. We predict an instance $\boldsymbol{x}$ as positive if $\boldsymbol{w}^\top \boldsymbol{x} + b \geq 0$, and negative otherwise. The initial values of the weight vector and the bias are set respectively to $\boldsymbol{w}^{(0)} = (0,0)^\top$ and $b = 0$. We visit the training instances in the order $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$. Answer the following questions.

A. Plot the dataset in the two-dimensional space. **(2 marks)**

The three points form a triangle with x_1 at the origin and x_2 and x_3 mirroring each other on the y-axis.

B. Write the perceptron update rule for a misclassified instance $(t, \boldsymbol{x})$. **(3 marks)**

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

C. What will be the values of the weight vector and the bias after observing the instance $\boldsymbol{x}_1$. **(3 marks)**

$y_1 = \boldsymbol{w}^{(0)\top} \boldsymbol{x}_1 + b^{(0)} = 0$. *Therefore, this instance is classified correctly as positive. The weight vector and bias are not updated.* $w^{(1)} = (0,0)^\top$, $b^{(1)} = 0$.

D. What will be values of the weight vector and the bias after observing $\boldsymbol{x}_2$. **(4 marks)**

$y_2 = 0$. *Therefore, $\boldsymbol{x}_2$ will be incorrectly classified as positive. The weight vector and the bias will be update to* $\boldsymbol{w}^{(2)} = (-1,-1)$, $b^{(2)} = -1$

E. What will be the values of the weight vector and the bias after observing $\boldsymbol{x}_3$. **(4 marks)**

$y_3 = (-1,1)^\top(-1,-1) - 1 = -1 < 0$. *Therefore, $\boldsymbol{x}_3$ will be incorrectly classified as negative. The updated values will be* $\boldsymbol{w}^{(3)} = (-2,2)^\top$, $b^{(3)} = 0$.

F. Is the dataset consisting of x_1, x_2, x_3 linearly separable?. Justify your answer. **(2 marks)**

Yes. We have already found a Perceptron with a weight vector and a bias that would correctly classify all three instances in this dataset.

G. Is it the case that a dataset consisting of three points is always linearly separable? If yes, explain your answer. If no, provide a counter example. **(4 marks)**

No. For example, if the three datapoints are on a straight line and the middle point has the opposite label than the other two points, then this dataset cannot be linearly separable.

H. Explain a method that you can use to learn a Perceptron from a non-linearly separable dataset. **(3 marks)**

Apply a kernel method to project the dataset into a high dimensional feature space and learn a Perceptron in this high dimensional feature space.

Question 3 Consider the two sentences S_1 and S_2 given by:

$$S_1 = \text{I love cake with tea}$$
$$S_2 = \text{I drink beer with cake}$$

Answer the following questions.

A. Represent S_1 and S_2 respectively by feature vectors $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$, where elements correspond to the frequency of unigrams. **(4 marks)**

Let the unigram features be indexed as follows: I=0, love=1, cake=2, with=3, tea=4, drink=5, beer=6. Then we have $\boldsymbol{s}_1 = (1,1,1,1,1,0,0)^\top$ *and* $\boldsymbol{s}_2 = (1,0,1,1,0,1,1)^\top$.

B. Compute the ℓ_2 norms of $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$. **(4 marks)**

$||\boldsymbol{s}_1||_2 = \sqrt{5}$, $||\boldsymbol{s}_2||_2 = \sqrt{5}$

C. Compute the ℓ_1 norms of $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$. **(4 marks)**

$||\boldsymbol{s}_1||_1 = 5$, $||\boldsymbol{s}_2||_1 = 5$

D. Compute the cosine similarity between $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$. **(2 marks)**

$\frac{\boldsymbol{s}_1{}^\top \boldsymbol{s}_2}{||\boldsymbol{s}_1||_2 ||\boldsymbol{s}_2||_2} = 3/5$

E. Compute the Manhattan distance between $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$. **(2 marks)**

$|1-1| + |1-0| + |1-1| + |1-1| + |1-0| + |0-1| + |0-1| = 4$

F. Assume that for all the unigrams u_i and bigrams $u_i u_{i+1}$ that appear in S_1 and S_2 we are given the marginal probabilities respectively $p(u_i)$ and $p(u_i u_{i+1})$. Compute the conditional probability of observing u_{i+1} given u_i. **(2 marks)**

$$p(u_{i+1}|u_i) = \frac{p(u_{i+1}, u_i)}{p(u_i)}$$
$$p(u_{i+1}, u_i) = p(u_{i+1} u_i) + p(u_i u_{i+1})$$
$$p(u_{i+1}|u_i) = \frac{p(u_{i+1} u_i) + p(u_i u_{i+1})}{p(u_i)}$$

G. Using the Markov assumption, compute the likelihood $p(S_1)$ and $p(S_2)$. **(4 marks)**

$$p(S_1) = p(I)p(love|I)p(cake|love)p(with|love)p(tea|with)$$
$$p(S_2) = p(I)p(drink|I)p(beer|drink)p(with|drink)p(tea|with)$$

H. Explain how you can use the computation done in part (F) to evaluate whether S_2 is less common than S_1 in English texts written by native speakers. **(3 marks)**

Compare $p(S_1$ *and* $p(S_2)$. *If the likelihood of a sentence is small, then it is unlikely to be produced by a native speaker.*

Question 4 Table 1 shows how four users u_1, u_2, u_3, u_4 purchased four items I_1, I_2, I_3, I_4 in an on-line shopping site over a period of one year. A cell value of 1 indicates that the user corresponding to the row has purchased the item corresponding to the column, and 0 otherwise. Answer the following questions.

	I_1	I_2	I_3	I_4
u_1	1	0	1	1
u_2	1	1	0	0
u_3	0	0	1	1
u_4	0	1	0	0

Table 1: A table showing four users u_1, u_2, u_3, u_4 who have purchased four items I_1, I_2, I_3, I_4 in an online shopping site over a period of one year.

A. Given that the users have been initially clustered into two clusters $\mathcal{S}_1 = \{u_1, u_2\}$ and $\mathcal{S}_2 = \{u_3, u_4\}$, compute the centroids for the two clusters respectively denoted by $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. For this purpose, consider a user is represented by a vector over the items he or she has purchased in the past. **(2 marks)**

$\boldsymbol{\mu}_1 = (1, 0.5, 0.5, 0.5)^\top$ and $\boldsymbol{\mu}_2 = (0, 0.5, 0.5, 0.5)^\top$

B. Compute Euclidean distances between μ_1 and each of the four users. **(4 marks)**

$$d(\boldsymbol{u}_1, \boldsymbol{\mu}_1) = \sqrt{0.75}$$
$$d(\boldsymbol{u}_2, \boldsymbol{\mu}_1) = \sqrt{0.75}$$
$$d(\boldsymbol{u}_3, \boldsymbol{\mu}_1) = \sqrt{1.75}$$
$$d(\boldsymbol{u}_4, \boldsymbol{\mu}_1) = \sqrt{0.75}$$

C. Compute Euclidean distances between μ_2 and each of the four users. **(4 marks)**

$$d(\boldsymbol{u}_1, \boldsymbol{\mu}_2) = \sqrt{1.75}$$
$$d(\boldsymbol{u}_2, \boldsymbol{\mu}_2) = \sqrt{1.75}$$
$$d(\boldsymbol{u}_3, \boldsymbol{\mu}_2) = \sqrt{0.75}$$
$$d(\boldsymbol{u}_4, \boldsymbol{\mu}_2) = \sqrt{0.75}$$

D. Based on the distances computed in parts (B) and (C), determine the assignment of users to clusters for the next iteration. **(2 marks)**

Two possible assignments exist. $\mathcal{S}_1 = \{u_1, u_2, u_4\}, \mathcal{S}_2 = \{u_3\}$ *or*
$\mathcal{S}_1 = \{u_1, u_2\}, \mathcal{S}_2 = \{u_3, u_4\}$

E. Let us denote the probability of a user purchasing an item I_j when he or she has purchased I_i by $p(I_j|I_i)$. From Table 1, compute $p(I_1|I_4), p(I_2|I_4)$ and $p(I_3|I_4)$. **(3 marks)**

$p(I_1|I_4) = 0.5$, $p(I_2|I_4) = 0$ and $p(I_3|I_4) = 1$

F. Based on your calculations in part (E), explain what is the best item to recommend to a user who has just purchased I_4. **(2 marks)**

I_3 because $p(I_3|I_4) = 1$ and the user is likely to buy I_3 too, given that he/she has already purchased I_4

G. Represent the information shown in Table 1 by a bi-partite graph where the users and items are represented by vertices, and an undirected edge is formed between the vertices corresponding to u_i and I_j if and only if u_i has purchased I_j. **(4 marks)**

H. Consider a random walker moving along the edges of the graph you created in part (G), where the probability of moving from u_i to I_j is given by $\frac{1}{d(u_i)}$ and the probability of moving from I_j to u_i is given by $\frac{1}{d(I_j)}$. Here, $d(x)$ is the out-degree of the vertex x. Given that the random walker started from u_1, compute the probability that the random walker will be in u_3 after two time steps. **(4 marks)**

$p(u_1 \rightarrow I_3)p(I_3 \rightarrow u_3) + p(u_1 \rightarrow I_4)p(I_4 \rightarrow u_3) = 1/3$

UNIVERSITY OF LIVERPOOL

Question 5 Consider the three points $x_1 = (0, 1)$, $x_2 = (-1, 0)$ and $x_3 = (1, 0)$. We would like to project these three points onto a straight line using principle component analysis. Answer the following questions.

A. Compute the total projection error if we project the three points onto the y-axis. **(3 marks)**

1+1 = 2

B. Compute the total projection error if we project the three points onto the x-axis. **(3 marks)**

1

C. Compute the mean $\bar{\boldsymbol{x}}$ of the three points. **(2 marks)**

(0, 1/3)

D. Compute the covariance matrix for the three points. **(3 marks)**

Compute $\boldsymbol{x}_i - \bar{\boldsymbol{x}}$ *and adding the outer product matrices gives [[2, 0], [0, 2/3]].*

E. Compute the eigenvalues of the covariance computed in part (D). **(4 marks)**

Because the covariance matrix is diagonal we have $\lambda_1 = 2, \lambda_2 = 2/3$

F. Compute the first principle component of the projection. **(3 marks)**

The eigenvector corresponding to λ_1 *(the larger eigenvalue) is (1, 0). This is the* x*-axis.*

G. Compute the second principle component of the projection. **(3 marks)**

The eigenvector corresponding to λ_2 *(the smaller eigenvalue) is (0,1). This is the* y*-axis.*

H. Compute the total variance if we had projected the three points on to the first principle component. **(2 marks)**

$\frac{(1-0)^2+(-1-0)^2+(0-0)^2}{3} = 2/3$

I. Compute the total variance if we had projected the three points on to the second principle component. **(2 marks)**

$\frac{(1-0.5)^2+(0-0.5)^2+(0-0.5)^2}{3} = 0.25$