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

UNIVERSITY OF LIVERPOOL

Second Semester Examinations 2016/17

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 Consider two sentences S_1 = *I like data mining* and S_2 = *I do not like data science*. Answer the following questions about S_1 and S_2

A. Write all unigrams in S_1. **(2 marks)**

B. Write all bigrams in S_2. **(2 marks)**

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

D. Assuming unigrams to be the feature space, represent S_1 and S_2 respectively by two vectors $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$, where the elements corresponds to the number times the corresponding unigram feature occurs in the sentence. **(4 marks)**

E. Compute the inner-product between $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$. **(3 marks)**

F. Compute the ℓ_2 norms of $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$. **(2 marks)**

G. Compute the cosine similarity between the two sentences S_1 and S_2. **(2 marks)**

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

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

J. Despite the two sentences are expressing opposite opinions, the cosine similarity measured in **G** gives a value greater than 0.5 indicating a high-degree of similarity. Suggest a solution to overcome this problem. **(4 marks)**

Question 2 Consider a training dataset $\{(\boldsymbol{x}_n, t_n)\}_{n=1}^4$, where $\boldsymbol{x}_n \in \mathbb{R}^2$ and $t_n \in \{-1, 1\}$. Here, $\boldsymbol{x}_1 = (1, 0)^\top$, $\boldsymbol{x}_2 = (1, 1)^\top$, $\boldsymbol{x}_3 = (-1, 0)^\top$, and $\boldsymbol{x}_4 = (-1, -1)^\top$. Moreover, the labels $t_1 = t_2 = 1$ and $t_3 = t_4 = -1$. Let us consider a support vector machine defined by a weight vector $\boldsymbol{w} = (\alpha, \beta)^\top$ and a bias term b. Here, $\alpha, \beta, b \in \mathbb{R}$.

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

B. Explain what is meant by the inequality $t_n(\boldsymbol{w}^\top \boldsymbol{x}_n + b) \geq 1$ with respect to each training data point. **(2 marks)**

C. Write down the four inequalities that must be satisfied by the four data points in the training dataset if they are to be correctly classified. **(4 marks)**

D. Show that the maximisation of the margin corresponds to the minimisation of $\alpha^2 + \beta^2$. **(4 marks)**

E. Using Lagrange multipliers for the four inequalities, write the objective function $L(\alpha, \beta, b, \boldsymbol{\lambda})$ where $\boldsymbol{\lambda}$ is a vector containing the four Lagrange multipliers given by $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)^\top$ each respectively for the inequalities corresponding to the four data points x_1, x_2, x_3, x_4. **(2 marks)**

F. Find the values of α, β, b by minimising the objective function defined in **E**. **(7 marks)**

G. Show that the decision hyperplane you obtained in **F** correctly classifies the four data points in the training dataset. **(4 marks)**

Question 3 Consider five data points in $\mathbb{R}^2$ given by $\boldsymbol{x}_1 = (0,0)^\top$, $\boldsymbol{x}_2 = (1,0)^\top$, $\boldsymbol{x}_3 = (1,1)^\top$, $\boldsymbol{x}_4 = (0,1)^\top$, and $\boldsymbol{x}_5 = (-1,0)^\top$. Here, data points $\boldsymbol{x}_1, \boldsymbol{x}_2$, and $\boldsymbol{x}_4$ are labelled with the colour red, whereas data points $\boldsymbol{x}_3$ and $\boldsymbol{x}_5$ are labelled with the colour blue. Answer the following questions about this dataset.

A. Plot this dataset in two-dimensional space. **(3 marks)**

B. Assuming that we used k-means clustering to cluster this dataset into two clusters, and we set the initial cluster centres at $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$. Compute the clusters after the first assignment. **(4 marks)**

C. How many further iterations does it take for the k-means algorithms to cluster for this dataset. Justify your answer. **(4 marks)**

D. Using the B-cubed method compute the precision for the final two clusters obtained at convergence. **(4 marks)**

E. Using the B-cubed method compute the recall for the final two clusters obtained at convergence. **(4 marks)**

F. Using the precision and recall values, compute the F-score for the final two clusters obtained at convergence. **(3 marks)**

G. Instead of creating two clusters, we would like to obtain three clusters via k-means where we use the data points 5, 1 and 2 as the centroids of the three initial clusters. Write down the final three clusters at convergence. **(3 marks)**

Question 4

A. Consider a biased coin that returns head with probability p. Let us assume that when this coin was flipped n times, we got $k(< n)$ heads. Answer the following questions about this event.

(a) Compute the likelihood of observing $k(< n)$ heads when this coin was flipped for n times. **(2 marks)**

(b) Using the maximum likelihood estimation, compute the most likely value of p for this probabilistic event. **(4 marks)**

B. Consider the dataset shown in Table 1 consisting of five reviews x_1, x_2, x_3, x_4, x_5 defined over three integer-valued attributes a_1, a_2, a_3 corresponding to the frequency of occurrences of a particular word in the document. The positive or negative sentiments label (t) of each review are denoted respectively by +1 and -1 in Table 1. Assuming that the three attributes to be mutually independent, and the probability of a review to be given by the product of the probabilities of individual attributes raised to their occurrences in the review, answer the following questions.

Review	a_1	a_2	a_3	label (t)
x_1	1	1	0	1
x_2	2	1	0	1
x_3	1	2	3	-1
x_4	0	1	1	-1
x_5	1	0	1	-1

Table 1: A set of five reviews represented using three attributes.

(a) Compute the marginal probabilities $p(a_1), p(a_2)$ and $p(a_3)$. **(3 marks)**

(b) Compute the conditional probabilities $p(a_1|t = 1), p(a_2|t = 1)$ and $p(a_3|t = 1)$. **(3 marks)**

(c) Assuming that the prior probabilities of the sentiment labels to be equal (i.e. $p(t = 1) = p(t = -1) = 0.5$), compute $p(t = -1|x_3)$, the probability of x_3 being negative. **(4 marks)**

(d) Apply Laplace smoothing for the occurrences of attributes in reviews shown in Table 1. Compute $p(t = 1|x_3)$ using the smoothed counts. **(4 marks)**

(e) Assuming the reviews to be independent, compute the likelihood of this dataset using the smoothed counts. **(5 marks)**

Question 5 We would like to project the four data points $x_1 = (0, 2), x_2 = (-1, 0), x_3 = (0, -2), x_4 = (1, 0)$ shown in Figure 1 onto the $y = \tan(\theta)x$ line that passes through the origin $O = (0, 0)$ and has an angle $0 < \theta < \pi/2$ with the positive direction of the x-axis. The four corresponding projected points along the line segment are shown by A_1, A_2, A_3 and A_4. Answer the following questions.

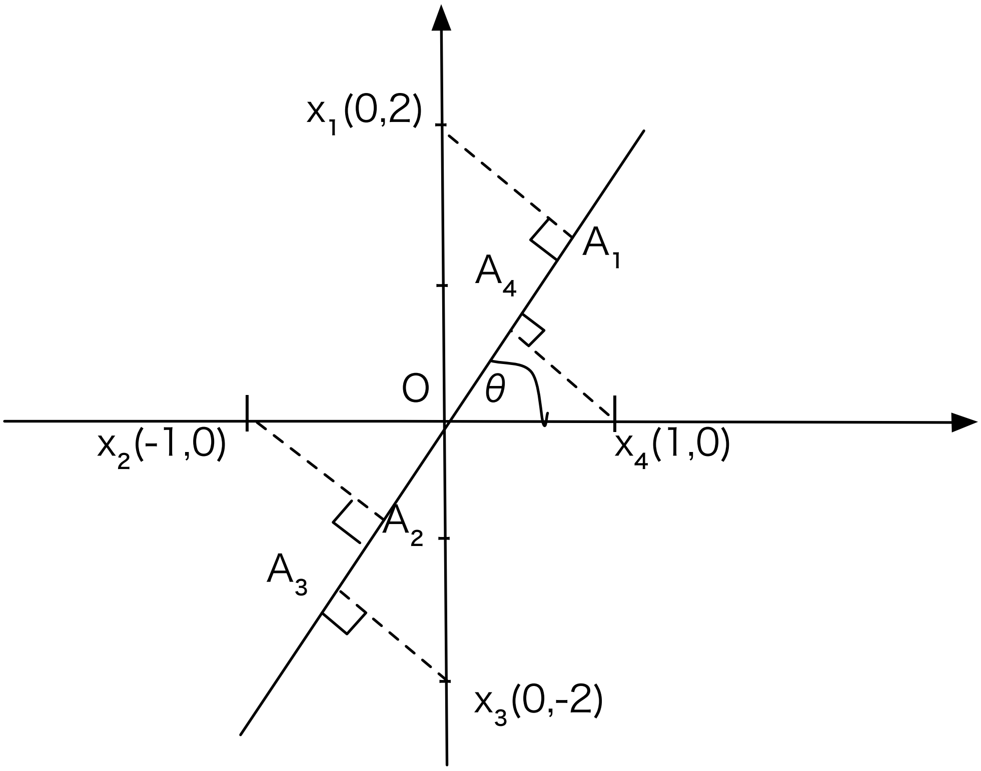

Figure 1: Four data points projected onto a straight line.

A. State two methods that can be used to project high dimensional data onto a lower dimensional space. **(2 marks)**

B. Compute the perpendicular distances x_1A_1, x_2A_2, x_3A_3, and x_4A_4 to the projection line $y = \tan(\theta)x$ respectively from the four points x_1, x_2, x_3, and x_4. **(4 marks)**

C. Find the value of $\tan(\theta)$ for which the sum of squared projection errors is minimised. **(4 marks)**

D. Compute the distances $e_1 = OA_1$, $e_2 = OA_2$, $e_3 = OA_3$ and $e_4 = OA_4$ to each of the projected points A_1, A_2, A_3, A_4 from the origin O. **(4 marks)**

E. Compute the projected mean of the four data points. **(2 marks)**

F. Compute the variance of the four data points on the line. **(4 marks)**

G. Compute the value of θ that maximises the variance of the projected points. **(5 marks)**