

PAPER CODE NO.  
**COMP527**

EXAMINER : Dr. Danushka Bollegala  
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  $S_1 = I \text{ like data mining}$  and  $S_2 = I \text{ do not like data science}$ . Answer the following questions about  $S_1$  and  $S_2$

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

*I, like, data, mining*

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

*I+do, do+not, not+like, like+data, data+science*

**C.** 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.*

**D.** Assuming unigrams to be the feature space, represent  $S_1$  and  $S_2$  respectively by two vectors  $s_1$  and  $s_2$ , where the elements corresponds to the number times the corresponding unigram feature occurs in the sentence. **(4 marks)**

*Let us assign indexes to the features as follows: I=0, like=1, data=2, mining=3, not=4, science=5, do=6. Then  $s_1 = (1, 1, 1, 1, 0, 0, 0)^T$  and  $s_2 = (1, 1, 1, 0, 1, 1, 1)^T$ .*

**E.** Compute the inner-product between  $s_1$  and  $s_2$ . **(3 marks)**

$$s_1^T s_2 = 3$$

**F.** Compute the  $\ell_2$  norms of  $s_1$  and  $s_2$ . **(2 marks)**

$$\|s_1\|_2 = 2, \|s_2\|_2 = \sqrt{6}$$

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

$$\cos(S_1, S_2) = \frac{s_1^T s_2}{\|s_1\|_2 \|s_2\|_2} = 3/2\sqrt{6}$$

**H.** Compute the Manhattan distance between  $s_1$  and  $s_2$ . **(2 marks)**

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

**I.** Compute the Euclidean distance between  $s_1$  and  $s_2$ . **(2 marks)**

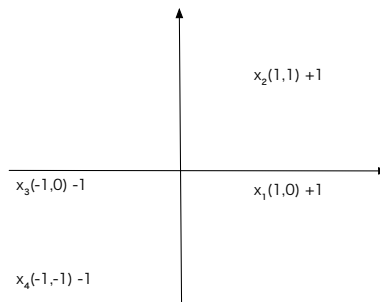
$$\sqrt{(1 - 1)^2 + (1 - 1)^2 + (1 - 1)^2 + (1 - 0)^2 + (0 - 1)^2 + (0 - 1)^2 + (0 - 1)^2} = 2$$

**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)**

*The issue here is that the negation indicator *not* matching (or not matching) between the sentences has equal contribution towards the cosine similarity as with any other feature. We can emphasise negation by assigning higher weights (larger than 1) to negation related features such as *not* to overcome this problem.*

**Question 2** Consider a training dataset  $\{(\mathbf{x}_n, t_n)\}_{n=1}^4$ , where  $\mathbf{x}_n \in \mathbb{R}^2$  and  $t_n \in \{-1, 1\}$ . Here,  $\mathbf{x}_1 = (1, 0)^\top$ ,  $\mathbf{x}_2 = (1, 1)^\top$ ,  $\mathbf{x}_3 = (-1, 0)^\top$ , and  $\mathbf{x}_4 = (-1, -1)^\top$ . 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  $\mathbf{w} = (\alpha, \beta)^\top$  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  $t_n(\mathbf{w}^\top \mathbf{x}_n + b) \geq 1$  with respect to each training data point. **(2 marks)**

This inequality means that each train data point must be classified correctly by the decision hyperplane  $\mathbf{w}$  with a margin of at least 1.

**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)**

$$\begin{aligned} \alpha + b &\geq 1 \\ \alpha + \beta + b &\geq 1 \\ \alpha - b &\geq 1 \\ \alpha + \beta - b &\geq 1 \end{aligned}$$

One mark is awarded for each inequality.

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

Maximisation of the margin corresponds to the minimisation of the  $\ell_2$  norm of the vector  $w$ , which is  $\alpha^2 + \beta^2$ .

**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)**

$$L(\alpha, \beta, b, \boldsymbol{\lambda}) = \frac{1}{2}(\alpha^2 + \beta^2) - \lambda_1(\alpha + b - 1) - \lambda_2(\alpha + \beta + b - 1) - \lambda_3(\alpha - b - 1) - \lambda_4(\alpha + \beta - b - 1)$$

**F.** Find the values of  $\alpha, \beta, b$  by minimising the objective function defined in **E**. **(7 marks)**

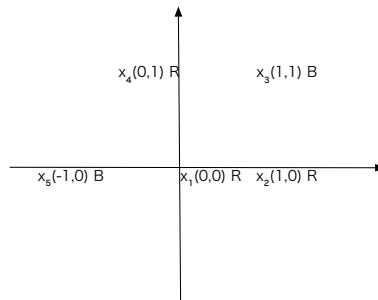
By setting the gradients of  $L$  w.r.t.  $\alpha, \beta, b$  to zero, we obtain three equations and w.r.t. each Lagrangian multiplier another four equations. By solving these equations we obtain  $\alpha = 1, \beta = 0, b = 0, \lambda_1 = \lambda_3 = 1/2, \lambda_2 = \lambda_4 = 0$ . One mark is awarded for each variable.

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

The weight vector is  $(1, 0)$ . Therefore, we have  $y_1 = y_2 = 1 > 0$  and  $y_3 = y_4 = -1 < 0$ . One mark is awarded for each data point.

**Question 3** Consider five data points in  $\mathbb{R}^2$  given by  $\mathbf{x}_1 = (0, 0)^\top$ ,  $\mathbf{x}_2 = (1, 0)^\top$ ,  $\mathbf{x}_3 = (1, 1)^\top$ ,  $\mathbf{x}_4 = (0, 1)^\top$ , and  $\mathbf{x}_5 = (-1, 0)^\top$ . Here, data points  $\mathbf{x}_1$ ,  $\mathbf{x}_2$ , and  $\mathbf{x}_4$  are labelled with the colour red, whereas data points  $\mathbf{x}_3$  and  $\mathbf{x}_5$  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  $\mathbf{x}_2$  and  $\mathbf{x}_3$ . Compute the clusters after the first assignment. **(4 marks)**

$$c_1 = \{3, 4\}, c_2 = \{1, 2, 5\}$$

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

0 further iterations. The  $k$ -means clustering has converged because the points 4 and 3 are the closest to  $c_1$  and points 5, 1, and 2 are the closest to  $c_2$ . Therefore, the cluster centroids do not get updated any further. Answers without the justification will receive only 2 marks.

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

$P(x_4) = P(x_3) = 1/2, P(x_1) = P(x_2) = 2/3, P(x_5) = 1/3$ . Therefore, the overall precision will be the summation of individual precisions divided by the total number of instances (5) in the dataset, which is  $8/15$ .

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

$R(x_4) = 1/3, R(x_3) = 1/2, R(x_1) = R(x_2) = 2/3, R(x_5) = 1/2$ . Therefore, the overall recall will be the summation of individual precisions divided by the total number of instances (5) in the dataset, which is  $8/15$ .

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

$$F = \frac{2PR}{P+R} = 8/15$$

- 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)**

$\{5\}, \{1, 4\}, \{2, 3\}$

**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)**

$$L = p^k(1 - p)^{n-k}$$

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

$\frac{\partial \log(L)}{\partial p} = \frac{\partial}{\partial p} k \log(p) + (n - k) \log(1 - p) = 0$ . This gives,  $k/p - (n - k)/(1 - p) = 0$ , from which we obtain  $p = k/n$ .

**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)**

$$p(a_1) = p(a_2) = p(a_3) = 1/3$$

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

$$p(a_1|t = 1) = 3/5, p(a_2|t = 1) = 2/5, \text{ and } p(a_3|t = 1) = 0$$

- (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)**

$$p(t = 1|x_3) = \frac{p(x_3|t=1)}{p(x_3)}$$

However, note that

$$p(x_3|t = 1) = p(a_1|t = 1)^1 \times p(a_2|t = 1)^2 \times p(a_3|t = 1)^3 = 0$$

because  $p(a_3|t = 1) = 0$ . Therefore,  $p(t = -1|x_3) = 1 - p(t = 1|x_3) = 1$ .

Review	$a_1$	$a_2$	$a_3$	label (t)
$x_1$	2	2	1	1
$x_2$	3	2	1	1
$x_3$	2	3	4	-1
$x_4$	1	2	2	-1
$x_5$	2	1	2	-1

Table 2: Smoothed counts.

- (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)**

Smoothed counts are shown in Table 2.  $p(a_1) = p(a_2) = p(a_3) = 10/30 = 1/3$ ,  $p(a_1|t = 1) = 5/11$ ,  $p(a_2|t = 1) = 4/11$ ,  $p(a_3|t = 1) = 2/11$ . Therefore, we have,

$$\begin{aligned}
 p(t = 1|x_3) &= \frac{p(x_3|t = 1)p(t = 1)}{p(x_3)} \\
 &= \frac{p(a_1|t = 1)^2 p(a_2|t = 1)^3 p(a_3|t = 1)^4 p(t = 1)}{p(a_1)^2 p(a_2)^3 p(a_3)^4} \\
 &= \frac{(5/11)^2 (4/11)^3 (2/11)^4 (1/2)}{(1/3)^2 (1/3)^3 (1/3)^4}.
 \end{aligned}$$

2 marks are awarded for computing the smoothed counts and another 2 marks for correctly computing  $p(t = 1|x_3)$ . No penalties for not simplifying the answers.

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

Likelihood of the dataset is given by  $p(x_1)p(x_2)p(x_3)p(x_4)p(x_5)$ . By substituting  $p(a_1), p(a_2), p(a_3)$  and aggregating the powers for each attributes we have,  $p(a_1)^{10} p(a_2)^{10} p(a_3)^{10} = \frac{1}{3^{30}}$ .



**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.

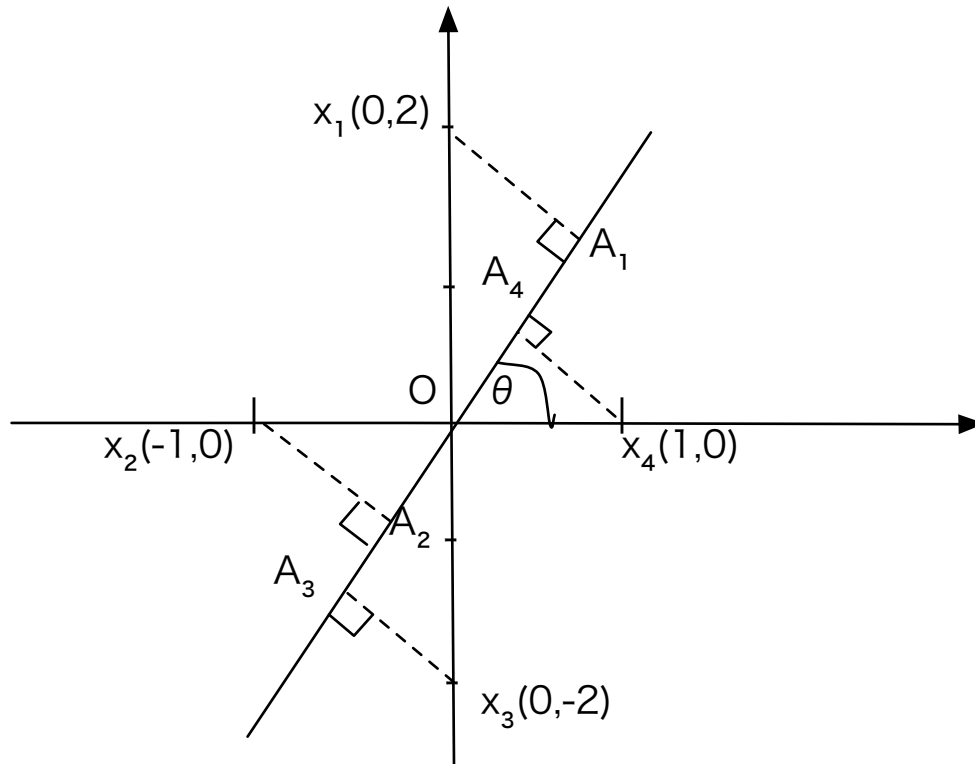


Figure 1: Four data points projected onto a straight line.

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

Principle Component Analysis, Singular Value Decomposition

- 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)**

Let  $d_1 = x_1A_1, d_2 = x_2A_2, d_3 = x_3A_3, d_4 = x_4A_4$ . Then we have,  $d_1 = 2 \cos(\theta), d_2 = \sin(\theta), d_3 = 2 \cos(\theta), d_4 = \sin(\theta)$

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

The sum of squared projection error,  $d_{sqd} = d_1^2 + d_2^2 + d_3^2 + d_4^2 = 8 \cos^2(\theta) + 2 \sin^2(\theta)$ . To find the minimum value of this error, we differentiate w.r.t.  $\theta$  and set the result to zero.

$\frac{\partial e}{\partial \theta} = -16 \cos(\theta) \sin(\theta) + 4 \sin(\theta) \cos(\theta) = 0$ , which gives  $\cos(\theta) \sin(\theta) = 0$ . Therefore,  $\theta = 0, \pi/2$ . This corresponds to  $x$  and  $y$  axes respectively.

- 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)**

$$e_1 = 2 \sin(\theta), e_2 = \cos(\theta), e_3 = 2 \sin(\theta), e_4 = \cos(\theta)$$

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

Mean of the four points is  $((0 - 1 + 0 + 1)/4, (2 + 0 - 2 + 0)/4) = (0, 0)$ . Note that  $(0, 0)$  gets projected to itself by  $y = mx$ . Therefore, the projected mean of the four data points is  $(0, 0)$ .

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

Because the mean is  $(0, 0)$  and we have measured distances on the line from  $(0, 0)$  already the variance is simply the sum of the squared distances (which is also the mean of the squared distances  $\mathbb{E}[e^2]$  in this case) and can be computed as follows:

$$\mathbb{V}[e] = \mathbb{E}[e^2] = e_1^2 + e_2^2 + e_3^2 + e_4^2 = 8 \sin^2(\theta) + 2 \cos^2(\theta).$$

- G.** Compute the value of  $\theta$  that maximises the variance of the projected points. **(5 marks)**

To find the optimal value of  $\theta$  we set  $\frac{\partial \mathbb{V}[e]}{\partial \theta}$  to zero, which gives,

$$\begin{aligned} 16 \sin(\theta) \cos(\theta) - 4 \cos(\theta) \sin(\theta) &= 0 \\ \sin(\theta) \cos(\theta) &= 0 \end{aligned}$$

The solutions are,  $\theta = 0, \pi/2$ . Therefore, maximising the sum of squared projection errors and minimising the variance of the projected points lead to the same optimal solution.