

**NANYANG TECHNOLOGICAL UNIVERSITY****SEMESTER 1 EXAMINATION 2006-2007****CI6124 – Data Mining & Machine Learning**

November 2006

Time Allowed: 3 hours

**INSTRUCTIONS**

1. This paper contains **FOUR (4)** questions and comprises **FIVE (5)** pages.
  2. Answer **ALL** questions.
  3. All questions carry equal marks.
- 

1. (a) Given the following set of 10 sample vectors:

$$\{\mathbf{x}^{(j)}\}_{j=1}^{10} = \left\{ \begin{bmatrix} 1 \\ 50 \end{bmatrix}, \begin{bmatrix} 2 \\ 40 \end{bmatrix}, \begin{bmatrix} 3 \\ 30 \end{bmatrix}, \begin{bmatrix} 4 \\ 20 \end{bmatrix}, \begin{bmatrix} 5 \\ 10 \end{bmatrix}, \begin{bmatrix} 5 \\ 10 \end{bmatrix}, \begin{bmatrix} 4 \\ 20 \end{bmatrix}, \begin{bmatrix} 3 \\ 30 \end{bmatrix}, \begin{bmatrix} 2 \\ 40 \end{bmatrix}, \begin{bmatrix} 1 \\ 50 \end{bmatrix} \right\}$$

- (i) Compute the median vector. (2 marks)
- (ii) Compute the sample mean vector  $\boldsymbol{\mu}$ . (2 marks)
- (iii) Compute the biased sample variance vector  $\boldsymbol{\sigma}^2$ . (2 marks)
- (iv) Is the correlation coefficient between the two dimensions positive or negative? (1 mark)
- (v) The dimensions of  $\mathbf{x}$  are clearly mismatched in scale. How would you normalize  $\mathbf{x}$  so that each dimension has zero mean and unit variance? Express each normalized dimension  $y_i$  as a function of  $x_i$ ,  $\mu_i$ ,  $\sigma_i$  for  $i=1,2$ ; you need **not** do the actual normalization. (2 marks)

Question No. 1 continues on Page 2

- (b) A term-document matrix contains elements  $f_{ij}$  which denotes the raw frequency/count (TF) of the  $i$ -th term/word in the  $j$ -th document. The  $j$ -th document is represented as a column vector  $\mathbf{x}_j = [f_{1j} \ f_{2j} \ \dots \ f_{Dj}]^T$  where  $D$  is the total number of terms/words in the corpus. Consider the following transformation defined by

$$f'_{ij} = f_{ij} \log \frac{N}{n_i}$$

where  $n_i$ , the document frequency (DF), refers to the number of documents containing the  $i$ -th term, and  $N$  is the total number of documents. This transformation is known as TFIDF (Term Frequency Inverse Document Frequency) and is used in many search engines.

- (i) How would this transformation affect the frequency  $f_{ij}$  of common terms/words like “the”, “a”, “an”?  
(2 marks)
- (ii) How would this transformation affect the frequency  $f_{ij}$  of rare terms/words?  
(2 marks)
- (iii) Explain intuitively, without using mathematics/equations, what is typically done to normalize a document vector so that the similarity between two document vectors is not dictated by document length?  
(2 marks)
- (iv) Assuming that normalization has been done for all document vectors, write down the expression for the cosine similarity metric  $s(\mathbf{x}, \mathbf{y})$  between two normalized document vectors  $\mathbf{x}$  and  $\mathbf{y}$ .  
(2 marks)
- (c) Consider the correlation between two attributes  $x_1$  and  $x_2$ .
- (i) Explain the meaning of negative, zero, and positive correlation.  
(4 marks)
- (ii) What is the correlation between  $x_1$  and  $x_2 = \sin(x_1)$ .  
(2 marks)
- (iii) What is the correlation between  $x_1$  and  $x_2 = \cos(x_1)$ .  
(2 marks)

2. Consider the following table of 10 data points with binary attributes A, B, and class labels. Suppose a decision tree is to be trained on this dataset and let  $p(c|t)$  denote the probability of class  $c$  @ node  $t$  of the decision tree.

<b>A</b>	<b>T</b>	<b>T</b>	<b>T</b>	<b>T</b>	<b>T</b>	<b>F</b>	<b>F</b>	<b>F</b>	<b>T</b>	<b>T</b>
<b>B</b>	<b>F</b>	<b>T</b>	<b>T</b>	<b>F</b>	<b>T</b>	<b>F</b>	<b>F</b>	<b>F</b>	<b>T</b>	<b>F</b>
<b>Class</b>	<b>+</b>	<b>+</b>	<b>+</b>	<b>-</b>	<b>+</b>	<b>-</b>	<b>-</b>	<b>-</b>	<b>-</b>	<b>-</b>

- (a) Tally  $2 \times 2$  contingency tables (counts of +/- samples for attribute=T/F) for
- Attribute A if we split on A. (4 marks)
  - Attribute B if we split on B. (4 marks)

- (b) The Entropy at node  $t$  for a two class problem is defined as:

$$\text{Entropy}(t) = - \sum_{c=1}^2 p(c|t) \log p(c|t)$$

- Calculate the overall Entropy before splitting. (1 mark)
- Calculate the overall Entropy after splitting on A. (2 marks)
- Calculate the overall Entropy after splitting on B. (2 marks)
- Which attribute should the decision tree split first? Why? (2 marks)

- (c) The Gini index at node  $t$  for a two class problem is defined as:

$$\text{Gini}(t) = 1 - \sum_{c=1}^2 [p(c|t)]^2$$

- (i)-(iv): Repeat parts (i) to (iv) of Question 2(b) using the Gini index instead of Entropy. (7 marks)
- (d) Both the Gini index and Entropy monotonically increase in  $[0, 0.5]$  and monotonically decrease in  $[0.5, 1]$ . Is it possible that each measure favor different attributes? Explain. (3 marks)

3. (a) Events A and B have probabilities  $P(A)$  and  $P(B)$ , respectively.
- (i) Express the conditional probability  $P(A|B)$  in terms of  $P(A,B)$  and/or  $P(A)$  and/or  $P(B)$ .  
(2 marks)
- (ii) Express conditional probability  $P(B|A)$  in terms of  $P(A,B)$  and/or  $P(A)$  and/or  $P(B)$ .  
(2 marks)
- (iii) **Derive** Bayes Theorem by expressing  $P(A|B)$  in terms of  $P(B|A)$  using your results from parts (i) and (ii).  
(2 marks)
- (iv) Write down the simplified expressions for  $P(A|B)$  and  $P(B|A)$  if events A and B are independent. Explain your results.  
(3 marks)
- (b) Perform single link (inter-cluster distance defined by two closest points) bottom-up agglomerative hierarchical clustering on the following similarity matrix between 5 sample points. Illustrate your results with a dendrogram clearly showing the order in which the points are merged.

	<b>p1</b>	<b>p2</b>	<b>p3</b>	<b>p4</b>	<b>p5</b>
<b>p1</b>	<b>1.00</b>	0.10	0.41	0.55	0.35
<b>p2</b>	0.10	<b>1.00</b>	0.64	0.47	0.98
<b>p3</b>	0.41	0.64	<b>1.00</b>	0.44	0.85
<b>p4</b>	0.55	0.47	0.44	<b>1.00</b>	0.76
<b>p5</b>	0.35	0.98	0.85	0.76	<b>1.00</b>

(10 marks)

- (c) Consider  $N$  points uniformly distributed in the unit hypercube  $[0, 1]^D$  where  $D$  is the dimensionality. Discuss whether the statistical notion of an outlier as an infrequently observed value is meaningful for the following cases:
- (i)  $D = 1$ , on a straight line.  
(3 marks)
- (ii)  $D = 100$ , when curse-of-dimensionality befalls.  
(3 marks)

4. (a) Consider the transaction data set shown below:

<i>TID</i>	<i>Items Bought</i>
1	{a, d, e}
2	{a, b, c, e}
3	{a, b, d, e}
4	{a, c, d, e}
5	{ b, c, e}
6	{ b, d, e}
7	{ c, d }
8	{a, b, c }
9	{a, d, e}
10	{a, b, e}

- (i) Draw the **prefix** tree equivalence class itemset lattice. Show a child itemset if and only if its parent itemset has more than ( $>$ ) 50% support. **No** credit for enumerating all possible itemsets. (6 marks)
- (ii) Write the support count next to every itemset node of your lattice in Question 4(a)(i). (3 marks)
- (iii) From your lattice, extract all association rules with over ( $>$ ) 50% support and ( $>$ ) 50% confidence. Compute and show the support and confidence for each extracted rule. (10 marks)
- (b) Explain with examples the following function approximation terms:
- (i) Regularization (1 mark)
- (ii) Occam's Razor (1 mark)
- (iii) Bias/Variance trade-off. (4 marks)

**End of Paper**

1. Data analysis

(a) Descriptive statistics

- i. [2] median vector =  $[3 \ 30]^T$
- ii. [2] mean vector =  $[3 \ 30]^T$
- iii. [2] variance vector =  $[2 \ 200]^T$       -2 marks if estimate is unbiased
- iv. [1] correlation = -1      (negative)
- v. [2] normalize each dimension ( $i=1,2$ ) via  $y_i = (x_i - \mu_i) / \sigma_i$

(b) TFIDF

- i. [2] It would vastly reduce the  $f_{ij}$  of common terms like “a”, “an”, “the”
- ii. [2] It would boost the  $f_{ij}$  of rare words like “nepotism”.
- iii. [2] The vector is normalized to unit length, by dividing each dimension by the vector length.
- iv. [2]  $s(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$  or  $s(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}$

(c) Correlation

- i. [1] Positive correlation means that when  $x_1$  increases/decreases,  $x_2$  increases/decreases generally
- [1] Negative correlation means that when  $x_1$  increases/decreases,  $x_2$  decreases/increases generally
- [1] Zero correlation means that there is no linear relationship between  $x_1$  and  $x_2$
- [1] Mention that correlation is a linear measure.
- ii. [2] mildly positive
- iii. [2] mildly negative

2. Decision Trees

(a) Contingency tables after splitting on attributes A and B are:

(i) [4]		A=T	A=F
+	4	0	
-	3	3	

(ii) [4]		B=T	B=F
+	3	1	
-	1	5	

(b) Entropy

- i. [1] Before splitting:  

$$E_{orig} = -\frac{4}{10} \log \frac{4}{10} - \frac{6}{10} \log \frac{6}{10} = 0.9710$$
- ii. [2] After splitting @ A  

$$E_{A=T} = -\frac{4}{7} \log \frac{4}{7} - \frac{3}{7} \log \frac{3}{7} = 0.9852$$

$$E_{A=F} = -\frac{3}{3} \log \frac{3}{3} - \frac{0}{3} \log \frac{0}{3} = 0$$

$$E_{overall} = \frac{7}{10} E_{A=T} + \frac{3}{10} E_{A=F} = 0.6896$$
- iii. [2] After splitting @ B  

$$E_{B=T} = -\frac{3}{4} \log \frac{3}{4} - \frac{1}{4} \log \frac{1}{4} = 0.8133$$

$$E_{B=F} = -\frac{1}{6} \log \frac{1}{6} - \frac{5}{6} \log \frac{5}{6} = 0.6500$$

$$E_{overall} = \frac{4}{10} E_{B=T} + \frac{6}{10} E_{B=F} = 0.71532$$
- iv. [2] A as it yields overall lower Entropy!

(c)

(d) [1] Yes.

[2] Any one of the following explanatory sentences:

The plot of measure values has no implication on which point on the Entropy or Gini curve a split ends up with. The difference in scale is not shown. As shown in this example, one index can indeed favor a particular attribute over another.

(c) Gini Index

- i. [1] Before splitting:  

$$E_{orig} = 1 - \left(\frac{4}{10}\right)^2 - \left(\frac{6}{10}\right)^2 = 0.48$$
- ii. [2] After splitting @ A  

$$G_{A=T} = 1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2 = 0.4898$$

$$G_{A=F} = 1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2 = 0$$

$$G_{overall} = \frac{7}{10} G_{A=T} + \frac{3}{10} G_{A=F} = 0.3429$$
- iii. [2] After splitting @ B  

$$G_{B=T} = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = 0.3750$$

$$G_{B=F} = 1 - \left(\frac{1}{6}\right)^2 - \left(\frac{5}{6}\right)^2 = 0.2778$$

$$G_{overall} = \frac{4}{10} G_{B=T} + \frac{6}{10} G_{B=F} = 0.3167$$
- iv. [2] B as it yields overall lower Gini index!  
 See above right

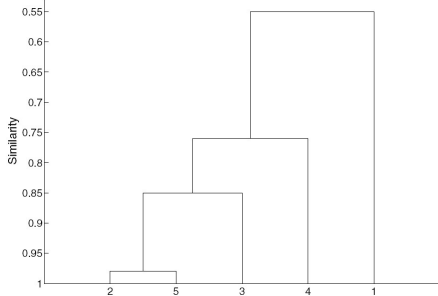
3. Clustering + Naïve Bayes Classifier

(a) Bayes Theorem

- i. [2]  $P(A|B) = P(A, B) / P(B)$
- ii. [2]  $P(B|A) = P(A, B) / P(A)$
- iii. [2]  $P(A|B) = P(B|A) P(A) / P(B)$
- iv. [1] For two independent events,  $P(A, B) = P(A)P(B)$   
 [2] thus  $P(A|B) = P(A)$  and  $P(B|A) = P(B)$

(b) HAC using single Linkage

- [2] for drawing some kind of linkage diagram, however inaccurate
- [4] 1 point each for correct merge order
- [4] 1 point each for correct similarity



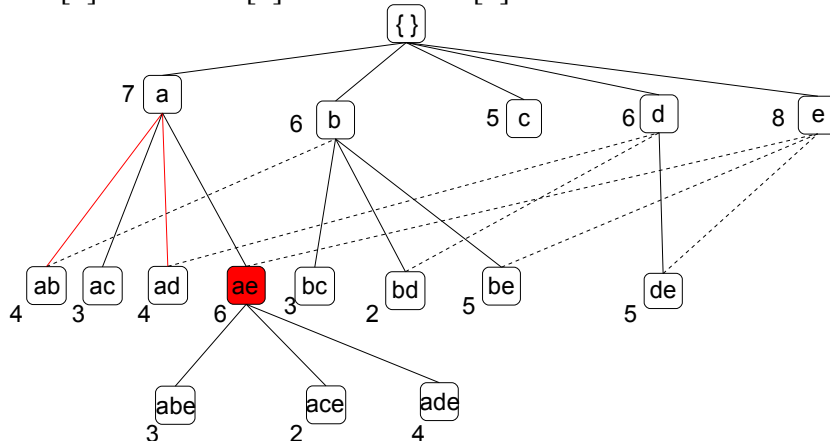
(c) Outlier

- i. [3] No. Because by definition, it has to be an infrequent point, but there are no infrequent points for a uniform distribution, unless for very small N.
- ii. [3] No. Although the majority of the points lie at the peripheral/edge of the space, there are still the same number of points in each linear subspace or sub-cube. Curse-of-dimensionality simply says that most of the points would not be at the center, which is just like any other non-centric subcube. In other words, we cannot say that any point lying in the center  $[0 \text{ to } 0.5]^D$  is considered an outlier, unless N is way smaller than the dimensionality. In other words, the points are still fully populating the hypercube, with no region less probable than another region.

4. Association Rules + Bias Variance + Func. approximation

(a) Association Rules

- i. [2] 1-item      [2] 2-item      [2] 3 item



$a \rightarrow e$	0.60s 0.86c
$e \rightarrow a$	0.60s 0.75c
acceptable:	
$d \rightarrow e$	0.50s 0.83c
$e \rightarrow d$	0.50s 0.63c
$b \rightarrow e$	0.50s 0.83c
$e \rightarrow b$	0.50s 0.63c

- ii. [3] 1 point each for each labeled level, see figures
- iii. [10] Rules with  $>50\%$  support/confidence listed in figure (ok if list rules with  $\geq 50\%$  support)

(b) Functional approximation

- i. [1] Regularization restricts the solution space  $g(x)$  to a small subset of preferred functions.
- ii. [1] Occam's Razor says that if given 2 solutions, one complex and one simple, the easier solution is always preferred.
- iii. [1] Bias is the difference between a model  $g(x)$  and the true function  $f(x)$ .  
 [1] Variance is the variation of  $g(x)$  around the model solution.  
 [1] A simple (e.g. linear) solution has high bias low variance and is less susceptible to noise.  
 [1] A complex solution has lower bias (better fit to data), but high variance.