HAND IN
Answers recorded
on question paper

Faculty of Arts and Science
School of Computing
CISC333 Examination
Instructor: D.B. Skillicorn

Winter 2007

Instructions:

1. You have three hours. This exam is worth 60 marks. You should spend about 3 minutes per mark.
2. Answer ALL questions in Section A. Answer any 2 questions in Section B. Answer 1 question in Section C.
3. You may bring in one 8.5x11 sheet of paper containing notes, and use it during this exam.
4. Answer questions on the exam paper. Nothing else will be marked. Use the amount of space under each question as a guide to the detail of your answer. You may use scrap paper and the back of the examination paper pages for rough working.
5. Proctors are unable to respond to queries about the interpretation of exam questions. Do your best to answer exam questions as written.
6. Make sure your student number is on every page of the exam paper.

STUDENT NUMBER: _____

STUDENT NUMBER written in words: _____

Section A: Short Answer Questions – answer all (1 mark each)

Answers should be brief enough to fit the space after the question.

1. What is the name of the algorithm Google uses to order web pages that it returns as the result of search queries?

2. Most of the datasets we considered in the course were *attributed*, that is their rows described objects, and their columns described attributes of these objects. What are two other kinds of data that might be used in data mining?

3. What is a *natural experiment* and how does it relate to data mining?

4. What is an ordinal attribute?

5. Explain how to derive a prediction accuracy from a confusion matrix.

6. In a dataset there are 127 objects labelled as Class A and 23 objects labelled as Class B. A predictor trained on this dataset achieves prediction accuracy of 84%. Is this a good result or not? Explain.

7. Explain briefly what a Self-Organising Map is, and how it works.

8. Explain briefly how to build a Random Forest classifier.

9. In the construction of a decision tree, how often does each record get counted as part of the computation of split points?

10. What is the "kernel trick" used in building support vector machines?

11. What are parallel coordinates, and what are they used for?

12. What is SemiDiscrete Decomposition. Give a small example.

13. What is hierarchical clustering?

14. What are Chernoff faces, and what are they used for?

15. Briefly explain what collaborative filtering is.

16. Briefly describe the DBSCAN clustering algorithm.

17. What is the difference between an opaque and a transparent predictor.

18. What is the difference between classification and regression?

19. What is a maximal margin separator?

20. Briefly explain the CN2 rule-discovery algorithm.

21. What is independent component analysis?

22. Given a large, sparse word-document matrix, explain how Latent Semantic Indexing can be used to improve retrieval of documents based on sets of search words.

23. What are the learning rate and momentum in backpropagation training of a neural network?

24. Categorical attributes with many values or numeric attributes discretized into many intervals can cause difficulties when building decision trees. What are these difficulties and how can they be avoided?

Section B: Medium Answer Questions – answer two of six (9 marks each)

Marks will be given for quality rather than quantity.

Space for answers can be found on pages 12 and 13.

25. We considered four different approaches to clustering, each using a different idea of similarity. These are: *distance-based* clustering, *density-based* clustering, *distribution-based* clustering, and *decomposition-based* clustering. Give a brief description of an algorithm of each kind, and an example of the kind of dataset for which it would be most appropriate.

26. Explain carefully why attribute scaling is a critical part of all clustering algorithms, and how such scaling should be done for at least two such algorithms.

27. Biclustering is an important special case of clustering. Explain what biclustering is, and describe two algorithms that are usually effective in discovering biclusters.

28. In settings where some people are manipulating the data that is collected about them to conceal themselves or their activities, many data-mining algorithms will not perform well. In particular, if the attributes of only one record are altered, it may be very hard to detect such manipulation. However, if a *group* of people are manipulating their data, it may be easier to detect this. Explain, giving some examples of algorithms that might address this problem well.

29. Given a set of labelled data, there are two main ways to divide the set into training and test data and use it to assess the performance of a classifier. One is to use cross validation, and the other is to use a specially chosen training subset and an out-of-bag test set. Explain in details how each of these approaches works, including reasonable choices for the parameters each uses, and explain which one is better in general and why.

30. The following matrix

$$\begin{bmatrix} 1.00 & 0.85 & 0.81 & 0.86 & 0.47 & 0.40 & 0.30 & 0.38 \\ 0.85 & 1.00 & 0.88 & 0.83 & 0.38 & 0.33 & 0.28 & 0.41 \\ 0.81 & 0.88 & 1.00 & 0.80 & 0.38 & 0.32 & 0.24 & 0.34 \\ 0.86 & 0.83 & 0.80 & 1.00 & 0.44 & 0.33 & 0.33 & 0.36 \\ 0.47 & 0.38 & 0.38 & 0.44 & 1.00 & 0.76 & 0.73 & 0.63 \\ 0.40 & 0.33 & 0.32 & 0.33 & 0.76 & 1.00 & 0.58 & 0.58 \\ 0.30 & 0.28 & 0.24 & 0.33 & 0.73 & 0.58 & 1.00 & 0.54 \\ 0.38 & 0.41 & 0.34 & 0.36 & 0.63 & 0.58 & 0.54 & 1.00 \end{bmatrix}$$

can be decomposed into this product

$$\begin{bmatrix} .40 & .28 \\ .39 & .33 \\ .38 & .34 \\ .39 & .30 \\ .35 & -.39 \\ .31 & -.40 \\ .29 & -.44 \\ .31 & -.31 \end{bmatrix} \begin{bmatrix} 4.67 & 0.0 \\ 0.0 & 1.77 \end{bmatrix} \begin{bmatrix} .40 & .39 & .38 & .39 & .35 & .31 & .29 & .31 \\ .28 & .33 & .34 & .30 & -.39 & -.40 & -.44 & -.31 \end{bmatrix}$$

Explain the factor interpretation, the geometric interpretation, and the component interpretation for this particular set of matrices. What can you say about the original data because of what is visible in the decomposition matrices?

Answer to first question from Section B:

Answer to second question from Section B:

Section C: Long Answer Questions – answer one of two (18 marks)

31. According to the the U.S. Centers for Disease Control, "people suffering from Chronic Fatigue Syndrome experience severe, all-encompassing mental and physical fatigue that is not relieved by rest and that has lasted longer than six months". It is characterised by: impaired memory or concentration, postexertional malaise, unrefreshing sleep, muscle pain, multijoint pain without swelling or redness, headaches of a new type or severity, sore throat that is frequent or recurring; and tender cervical or axillary lymph nodes

    There is still considerable uncertainty about whether this syndrome is caused by several different underlying diseases. Suppose that you were going to investigate this question, and you had access to clinical data about a set of 1000 healthy patients and a set of 1000 patients who have been diagnosed with chronic fatigue syndrome. For example, the clinical data might contain estimates of severity and/or frequency of the symptoms described in the previous paragraph.

    Explain how you would approach this problem from a data-mining perspective, including at least:

    - What, specifically, you would be looking for;
    - The overall process you would follow;
    - The algorithms you would try, in order of their expected usefulness;
    - The kind of results you would expect from each algorithm, depending on what is true about this syndrome;
    - Any particular pitfalls you would be watching out for;
    - How you would assess how well you have succeeded at the task.

32. An insurance company has data about 3 million of its customers who have bought car insurance, house insurance, life insurance, and any combination of these. For each customer, there might be data about each different kind of insurance; for example, length of time the customer has purchased this insurance, total amount paid, number of claims, and total value of claims. As well, there is demographic data for each customer such as age, type of job, type of education, household income, and postal code.

    The company wants to analyse this data with three goals:

    (a) Find the customers who have provided the greatest profit, that is who have paid the most in premiums, and claimed the least, so that new customers like them can receive high-quality service from the start.

    (b) Find and model the difference between customers who have bought all three kinds of insurance from the company and those who have bought only one or two, so that future sales can be focused on those who might buy all of their insurance from the company.

(c) Find and model customers who are poor risks so that new potential customers like them can be turned away.

Explain how you would approach these problems, including at least:

- Any difficulties that might be caused by the form of the data;
- The overall process you would follow;
- The algorithms you would try, in order of their expected usefulness;
- The kind of results you would expect from each algorithm;
- Any particular pitfalls you would be watching out for;
- How you would assess how well you have succeeded at the task.

Answer to long question:

Answer to long question (continued):

Answer to long question (continued):

Extra space (if needed) – make sure there's a clear pointer from earlier in the exam.