HAND IN
Answers recorded
on question paper

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

Winter 2008

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 meant by Manhattan distance?

2. Why is prepruning less common than postpruning in decision tree construction, even though it requires less work?

3. What is meant by bootstrap or out-of-bag sampling?

4. What is temporal data mining? Give an example.

5. What is meant by agglomerative hierarchical clustering?

6. Give two examples of similarity measures used in hierarchical clustering, and the implications of choosing them.

7. What is meant by overfitting?

8. What is meant by underfitting?

9. Explain what information gain is, and how it is used.

10. Explain how you would set up a neural network to predict a two-class problem with associated estimates of the confidence of the predictions.

11. Explain what boosting is.

12. Explain, briefly, what SemiDiscrete Decomposition is.

13. What is the Naive Bayes assumption? Why is it important?

14. What are some ways of dealing with missing values?

15. What is the biggest weakness of the PageRank algorithm?

16. What are parallel coordinates and what can be learned from them?

17. What is an appropriate similarity measure for categorical data?

18. Explain the ‘kernel trick’ for SVMs.

19. The k-means clustering algorithm can produce poor clusterings because of poor choices of the initial centroid positions. What are two techniques that have been used to avoid this?

20. In intrusion detection and other outlier-detection settings, there are two broad approaches: anomaly detection (anything I haven't seen before is bad), and malice detection (I have a list of known bad things). Explain the weaknesses of both of these approaches.

21. What is the component interpretation of a matrix decomposition?

22. Explain why discretization of attribute values is sometimes needed.

23. Give two examples of the difficulties of data mining using English textual data.

24. What is the main advantage of density-based clustering?

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 11 and 12.

25. One of the prediction techniques we didn't talk (much) about is nearest-neighbour prediction. Given a new record, find its k most similar records in the training data, and predict its class to be the plurality of the classes of these neighbours, or the average of their class labels for regression.

    What benefits and disadvantages can you see for nearest-neighbour prediction?

26. Explain carefully what it means to say that attributes are chosen in random forests in a doubly contextualized way.

27. For partitional clustering algorithms, choosing the 'right' number of clusters is one of the most difficult problems. Describe several ways of addressing this problem.

28. What are some of the implications for data mining of an adversarial setting (that is where some of the people involved are actively trying to mislead the analysis)?

29. Explain why normalization of a dataset is often required before it is analysed, and describe several ways in which this could be done.

30. Explain carefully what a confusion matrix for a two-class prediction problem looks like, and what its entries mean. Explain what is meant by the false-positive and false-negative rates. Give examples of situations where each of these rates is more important than the other.

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. Suppose that a model (predictive or clustering) has been built from a set of data; and that new data is constantly arriving, and being assessed using the existing model.

    Define an *outlying* newly-arrived record to be one that is different from any of the data from which the model was built. Such a record might be far from any of the training data, but its 'meaning' is not necessarily unclear – a predictor might still be confident about predicting its class, and a clustering algorithm sure about which cluster it belongs to.

    Define an *interesting* newly-arrived record to be one that suggests that the existing model should be changed. Such a record casts doubt on the structure of the model – it might be close to the boundary of a predictor, or indicate that two of the original clusters are actually one.

    Some outlying records might also be interesting, but perhaps not many of them.

    If the real problem is to detect interesting new records, what forms of data mining would be appropriate for the original model? How should it be extended to concentrate attention on new interesting records?

    To maximise your marks, you should make sure to include some discussion of:

    - The role of prediction;
    - The role of clustering;
    - Any particular pitfalls you would be watching out for;
    - How you would assess how well you have succeeded at the task.

32. Telephone companies have often analysed their data, but they face challenges because there is so much of it: perhaps 100 billion calls per day worldwide. Any sizable telephone company is responsible for tens of millions of calls each day.

    Keeping data about calls for a reasonable amount of time, so that it can be analysed in a context, is difficult. One solution that has been used is to keep, for each telephone number, a list of the top m numbers called, and an approximation to the number of times each has been called.

    Such a list is built as follows:

    - On the first day, count the m most frequently called numbers, together with their frequencies.
    - On the second, and subsequent, days, compute a similar list, and merge the two lists, giving today's list a weight of α and the historical list a weight of $1 - \alpha$. The new list may need to be truncated if it has more than m numbers called in total.

The result is a list of popular numbers called, with estimates of how often they are called. Different choices of α allow different amounts of emphasis on recent versus older call patterns.

Telephone businesses routinely charge businesses more for a phone line than residential customers; and businesses sometimes try to pass themselves off as residential to save money.

Explain how you would build a predictor that would take call lists, as described above, and predict which ones belong to businesses. These can then be checked to see whether they are paying the business rate.

You should include discussion of the following points:

- 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.