HAND IN
Answers recorded
on question paper

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

Fall 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 24 (1 mark each)

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

1. What is a 'natural experiment'?

2. What is the difference between a categorical and an ordinal attribute?

3. What is the difference between classification and regression?

4. What is the bias of a model?

5. Explain how a prediction model can underfit a set of training data.

6. When would it be appropriate to use information gain ratio in preference to information gain in building a decision tree?

7. What is the function computed by an internal node of a neural network used as a predictor?

8. Give examples of two prediction techniques that are able to output both a prediction and a confidence in that prediction.

9. What is the kernel trick for support vector machines and why is it important?

10. What are association rules?

11. What is the main difference between the PRISM and CN2 rule construction algorithms?

12. What are ensemble techniques for prediction?

13. What is maximal likelihood estimation?

14. What is the structure produced by a hierarchical clustering called?

15. What is latent semantic indexing?

16. What is independent component analysis?

17. What are parallel coordinates?

18. What is biclustering?

19. Give an example of a property that could be predicted from clickstream data.

20. What is collaborative filtering?

21. What is meant by adversarial data mining?

22. What are some properties of a document's author that can be predicted from its textual content?

23. What would be the preferred prediction technique for a dataset with 1 million records, 100 attributes and 7 classes, and why?

24. What would be the preferred clustering technique for a dataset with 1 million records, 100 attributes, and probably 5-10 clusters, and why?

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

Marks will be given for quality rather than quantity.
Space for answers can be found on pages 11 and 12.

25. The performance of a predictor depends, to some extent, on the particular training data from which it was built. To understand how sensitive prediction performance is to this choice it is usual to build multiple predictors using cross validation. Explain what this means, and what can be learned by doing it.

    It is also possible to use out-of-bag testing. Explain what this means, and how it compares to cross validation.

26. The classification function for a support vector machine is:

    $$f(x) \ = \ \sum \alpha_i y_i < x_i, x > +b$$

    where the sum is over the training records. Explain the meaning of each of the terms in this expression, and explain, in an intuitive way, how this calculates on which side of the maximal margin separator each new record lies.

27. Clustering records requires some idea of similarity between each pair of records. What decisions must be made to define such a similarity, and what is their effect on the resulting clustering. Explain also why normalization is important and how it can be done.

28. One way to address the difficult issues of false-positive and false-negative rates for a predictor is to build one that uses regression rather than classification, and then impose a boundary on the regression values. Explain how this can help to produce better predictors for setttings where false-positive and false-negative rates are very important.

29. What is the Naive Bayes assumption? Why is it often made, and what are the implications for prediction?

30. Explain the procedure for building a predictor using boosting.

31. Most clustering techniques require the number of expected clusters to be given as a parameter but this number is hardly ever known in practice. Describe some techniques that can be used to handle this problem when the number of clusters to expect is not known.

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)

32. Fredmart, a large multinational retailer, wants to understand its customers better. To this end they want to collect details of each purchase (time of day, date, number of each item purchased on this visit) and then cluster the resulting data. If there are well-defined clusters, some of their staff will try to label each with some descriptive property of the kind of customers it contains. Fredmart sells roughly 10,000 products in each of their stores.

    They have retained you to do the data mining involved. Describe the steps you would take, under the following headings:

    - The number and type of records that you would request (you can't get other attributes, but you can control which records are saved for you);
    - Any data preprocessing that you think is necessary;
    - The clustering technique(s) you would use, and the exact process you would apply to use it.
    - The kind of results that you would anticipate;
    - How you would assess how well your approach has worked;
    - and, any other issues that you think are worth mentioning.

33. A large bookstore chain asks its customers to rate any books that they might have read. This rating is done online, and customers must identify themselves before they can rate, so ratings that come from the same individual are known. The chain sells 30,000 different books and, so far, 5,000 customers have signed up and given some ratings.

    Such a dataset could be used for collaborative filtering and recommendation. However, the bookstore chain wants to use this data in a different way, to try and determine which books are *sleepers*. A sleeper is a book that is not heavily publicised by its publisher, but becomes a bestseller slowly because of word of mouth recommendations. The chain thinks that it can do well by knowing which books are sleepers before they become bestsellers, so that it can order lots of copies, and gain reputation by knowing in advance which books its customers will like.

    You want to build a prediction system for this chain that will predict sleepers for them, based on the rating dataset. Describe your process in detail. It may be helpful to think about answers to these questions:

    - What does a sleeper look like? Does this change over time from when it first comes out, as it becomes a cult favourite, and then becomes a bestseller?
    - Are there characteristics of customers who are especially good at seeing sleepers early, and can they be identified and exploited?

- What is the effect of books that are heavily publicised by their publishers on the problem?
- Would it help to get the ratings data in a form that includes temporal information explicitly?

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.