HAND IN
Answers recorded
on question paper

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

Fall 2011

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. In what time period is it estimated that the total amount of data stored electronically doubles?

2. What is a natural experiment?

3. What is a categorical attribute?

4. Give an example of an application where the false positive rate might be much more important than the false negative rate or *vice versa*.

5. What is an opaque predictor?

6. What is meant by missing data, and why is it significant?

7. Explain briefly how to choose an out-of-bag sample.

8. What is the advantage of using axis-parallel boundaries in a predictor?

9. What is the denominator of Bayes Rule and why is it usually ignored?

10. What is the Naive Bayes Assumption?

11. What is the definition of entropy in the context of prediction?

12. What is the advantage of information gain ratio over information gain?

13. What is a brittle predictor?

14. Describe the use of the sigmoid function in neural networks.

15. What is the hidden layer in a neural network.

16. Describe any one of the three key features of Support Vector Machines.

17. What is the support of a rule?

18. What is a “bag of words”?

19. What is an outlier in a dataset? What is the difference between a global outlier and a local outlier?

20. What is L_∞ distance?

21. How can a minimal spanning tree be used to build a clustering?

22. What is latent semantic indexing?

23. What is an online data mining algorithm?

24. What is the name of the algorithm Google uses to rank web pages, and what mathematical technique does it depend upon?

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. Some clustering algorithms have a way to compare one cluster to another using the same similarity measure that is used to compare one record to another; other clustering algorithms require a new inter-cluster similarity measure to be defined. Provide some examples of each, from the clustering algorithms we have seen this term. For those of the second kind, explain what the new inter-cluster similarity measure is, or could be.

26. A matrix decomposition expresses a data matrix, A, as the product of two other matrices, $A = C.F$. There are three different ways of interpreting such a decomposition that differentiate aspects that are often useful to an analyst. Describe these three different interpretations.

27. In recommender systems, the key is to find records that resemble those of a new individual so that properties of these records can be used to make recommendations. Give some examples of how "resemble" can be made algorithmic, and discuss the advantages and disadvantages of each.

28. Suppose that I start surfing the web, and for every page I visit, I choose a random outgoing link as my next step. I will visit some pages more than once and, in fact, the more often I visit a page the better-connected and therefore somehow important it must be.

    Explain how this intuition can be turned into practical computation, mentioning particularly the problem that, if I enter a region of the web with no outgoing edges, I will be trapped in that region, and the visits = importance idea will no longer work.

29. Density-based clustering tries to find clusters that do not necessarily have obvious geometric properties or shapes. Explain how a typical density-based clustering algorithm works. Do such algorithms continue to work well even if clusters are more conventional, for example roughly elliptical?

30. Rule-based systems choose rules by using some kind of covering strategy for training-set data records. Even when the rule set is unordered, the order in which this covering happens can affect the quality of the resulting rule set. Explain this issue fully, using some example of the potential problems.

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)

You have almost an hour to spend on this question – thinking beats writing.

31. Public companies in the United States have to file quarterly reports with the Securities and Exchange Commission. Suppose that we wanted to predict which companies were committing fraud using the language of the textual sections of such reports, rather than more conventionally looking at the financial accounts.

    Class labels are available, because the SEC identifies, after the fact, quarters that it deems fraudulent.

    The goal is to build a predictor that will predict fraud for new quarterly filings; much earlier and more cheaply than the SEC manages.

    Describe the steps you would take, using the following headings as a guide:

    - Any data preprocessing that you think is necessary;
    - The way you would handle the uncertainty in labels, since the SEC has to be very conservative in which filings it labels as fraudulent; and sometimes frauds are not discovered until long afterwards.
    - The prediction technique(s) you would use;
    - 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.

32. The problem of email spam has been almost completely solved, but it's a useful surrogate for other interesting problems. In this question, the goal is to build a spam prediction system that could be integrated into your favourite email environment.

    Explain the design of your complete system. You should include discussion of the following points:

    - The cold start problem: you rely on the user labelling some emails as spam and, by default, labelling the rest of the emails as normal. However, at the beginning you have very few examples of spam.
    - The need to update the predictor regularly, in principle after every email.
    - How you might make use of the structured form of the data in an email.
    - How many classes you would use for prediction and why.
    - Any role for attribute selection that you envisage.
    - Any role for clustering that you envisage.
    - The prediction 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.