HAND IN
Answers recorded
on question paper

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

Winter 2006

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 mean by *customer relationship management*. Give two examples of how data mining can be applied to it.

2. Describe one way in which a categorical attribute can be converted to a numerical one.

3. In prediction, explain why some of the labelled data are withheld as a *test set*.

4. Explain what is meant by a false positive.

5. What is regression?

6. If you are given a dataset with missing values, what is the most important question to ask about these missing values.

7. Explain why variance is an important property of a predictor.

8. What is the Naive Bayes assumption, and why is it important?

9. Explain one of the criteria used for selecting the best attribute to use at each stage of decision tree construction.

10. Explain the structure of the function used as the output of a single neuron in a neural network used for prediction.

11. Briefly describe the levelwise algorithm used to compute association rules.

12. What is the bias-variance decomposition, and what do we learn from it?

13. What is bagging?

14. Briefly explain the k-means clustering algorithm.

15. What is maximum likelihood estimation?

16. Why are bottom-up hierarchical clustering algorithms more popular than top-down hierarchical clustering algorithms.

17. What is the single-link measure for hierarchical clustering, and what is its biggest weakness?

18. What is the component interpretation of a matrix decomposition of the form $A = CF$?

19. Briefly explain the BIRCH clustering algorithm for large datasets.

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

21. Explain what is meant by biclustering.

22. Google's basic PageRank algorithm fails when a region of the web has no outgoing links. Explain how they adapt the basic eigenvector algorithm to solve this problem.

23. What is different about mining data in settings where humans are involved in covert or criminal activity? is happening?

24. We've discussed the problem of overfitting a model. Explain how it is possible to *underfit* a model, giving an example based on one prediction technique.

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. Support vector machines are good predictors, because of three key ideas:

    maximum margin separators;
    linear separation in a higher-dimensional space;
    the kernel trick.

    Explain carefully what each of these means, and how they work together. What is the biggest weakness of support vector machines?

26. Explain what random forests are, and how they work. What are the main advantages of random forests over other predictors?

27. Explain what content-based and collaborative-filtering recommender systems are, emphasising the significant differences between them.

28. Explain the way in which Google collects data about the content and connections of the web, and how it processes this data using the PageRank algorithm. How are the results of this analysis used to respond to search queries?

29. Explain the Expectation-Maximisation algorithm in detail, emphasising its strengths and weaknesses.

30. Create your own question that is not asked elsewhere on this exam and answer it. You will receive marks for the quality of the question and the answer.

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. NASA tracks many near-earth objects, partly to watch for any that might be on a collision course with the Earth. Suppose that NASA hired you to determine if any of these objects was actually an alien spacecraft masquerading as an inert object. Explain the way you would approach this problem, and the techniques you would try.

    Remember to discuss what kinds of data you assume would be available to you (keep it plausible).

32. A supermarket knows which items were bought in the same 'market basket', that is by one person in one pass through the checkout. Suppose you are a supermarket manager and your chain is planning to introduce a customer loyalty card, which would be used by an individual each time he or she went through the checkout.

    Discuss some of the issues that will arise in getting customers to adopt the loyalty card, for example by describing reasons they might want to, and reasons they might not want to.

    Describe the data-mining techniques that you would probably be using now, and then explain how these would change with the introduction of the loyalty card. Be sure to explain the new analysis that would be made possible, including clustering and predictive techniques, as well as the pros and cons of the new environment.

Answer to long question:

Answer to long question (continued):

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