UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR11007 DATA MINING AND EXPLORATION

Friday 22$^{\underline{\text{nd}}}$ May 2015

14:30 to 16:30

INSTRUCTIONS TO CANDIDATES

Answer QUESTION 1 and ONE other question.

Question 1 is COMPULSORY.

All questions carry equal weight.

CALCULATORS MAY NOT BE USED IN THIS EXAMINATION

MSc Courses

Convener: F. Keller
External Examiners: A. Burns, S. Denham, P. Healey, T. Norman

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

1. **You MUST answer this question.**

   (a) You are at the start of a project to assess the benefits of each individual marketing method for a national furniture retailer involved in multiple parallel advertising streams (e.g. television, local radio, direct marketing). Briefly discuss what issues you would initially choose to consider and what initial questions you would choose to ask. Give reasons for your choices. [*7 marks*]

   (b) A train company is interested in assessing the predictive factors for train overcrowding on a particular line during the rush hour. To do this they have historical passenger count data for each train, information on ticket sales, timetables, and data on delays that occurred. They also have access to similar information from other trains which could connect people to the train line in question. You have been employed to do this analysis. Describe, using illustrations from this example, the four main tasks you would need to do in preprocessing this data. [*7 marks*]

   (c) In a classification problem, what visualisation procedure would you use to help in choosing a good attribute set or feature set? How does this procedure aid the choice? [*5 marks*]

   (d) What is the basic principle of projection pursuit? Outline a projection pursuit algorithm, commenting on the reasons for each part. [*6 marks*]

2. (a) In the paper "Meme-tracking and Dynamics of the News Cycle", the authors track phrase useage in news reports. Describe how a phrase is represented, and why an optimal representation of this form is not feasible. [*6 marks*]

(b) Data is given for a two class classification problem. The class 1 data is

$$x_1 = (2, 1),\ x_2 = (1, 0),\ x_3 = (0, 0),\ x_4 = (0, -1),\ x_5 = (-1, -2)$$

and the class 2 data is

$$x_6 = (-1, 3),\ x_7 = (-2, 5),\ x_8 = (-3, 6)$$

You wish to build a classifier using a linear support vector machine. Plot the labelled data and draw a line indicating the decision boundary of the maximum margin classifier for this data. Draw circles around the support vectors, and indicate on the figure the critical distances that define the decision boundary. [*6 marks*]

(c) A probabilistic classifier trained on other training data gives the following classification probabilities for the data in Question 2b:

$$P(x_1 = 1) = 0.6,\ P(x_2 = 1) = 0.8,\ P(x_3 = 1) = 0.7,\ P(x_4 = 1) = 0.6,$$

$$P(x_5 = 1) = 0.4,\ P(x_6 = 1) = 0.7,\ P(x_7 = 1) = 0.2,\ P(x_8 = 1) = 0.4$$

Give the confusion matrix for these results using a decision boundary of probability 0.5. Plot the ROC curve for this data. [*6 marks*]

QUESTION CONTINUES ON NEXT PAGE

(d) There are two highly specialist shops that only sell handbags and watches.

Shop A sells, on average, 10 handbags a day at 30 pounds each and 40 watches a day at 50 pounds each.

Shop B sells, on average, 40 handbags a day at 40 pounds each and 10 watches a day at 60 pounds each.

Using only these results, and assuming all averages are true and exact, how would you rate the truth or falsity of the following statements:

i. The average cost of items sold at shop A is higher than shop B.

ii. The average cost of handbags sold at shop A is higher than shop B.

iii. The average cost of watches sold at shop A is higher than shop B.

Which of the following conclusions are valid, which are invalid, and why?

i. You offer to pay for a friend to buy themselves a present from one of the two shops, but would rather they spent less. You conclude that you hope they go into shop B because of the difference in the average cost of items sold.

ii. Another friend comes to you and says they have decided what to buy for your birthday (it is a secret). They could buy it from shop A or shop B, but would like advice regarding which shop would work out cheaper. You conclude that you should send them to shop B because of the difference in the average cost of items sold.

iii. The second friend follows your advice, and goes to shop B. You conclude that as shop B sells more handbags, it is more likely that friend B has bought you a handbag for your birthday.

iv. You find that yet another friend has bought you a present from shop A. You conclude that as shop A sells more watches, it is more likely that this friend has bought you a watch for your birthday. [*7 marks*]

3. (a) Give pseudocode for the bagging algorithm for a binary classification problem. Describe how the classifier is constructed, and how the classifier predicts the labels for new instances. [*5 marks*]

(b) Suggest a potential base learner for the bagging algorithm. Why might bagging improve this learner? [*2 marks*]

(c) Consider a binary classification problem. What is the confusion matrix? Define accuracy, precision, and recall in terms of the confusion matrix. Give a situation in which precision and recall are to be preferred to accuracy. [*7 marks*]

(d) Describe the approach in the paper "Suggesting Friends using the Implicit Social Graph" by Roth *et*. al at a high level. How do they give more weight to more recent instances? [*5 marks*]

(e) In the paper "Connecting the Dots between News Articles" by Shahaf et al, what is the main computational formalism used? How do the authors evaluate their system? [*6 marks*]