UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR11007 DATA MINING AND EXPLORATION

Friday 17 th May 2013

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: B. Franke
External Examiners: T. Attwood, R. Connor, R. Cooper, S. Denham, T. Norman

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

1. **You MUST complete this question.**

(a) You are working for a site that collects local restaurant reviews from users. You have been asked to develop a system that suggests new restaurants that a user might not be familiar with. Describe how you would approach this as a machine learning problem. You should describe the entire data mining process, from start to finish: what data you would collect and why, which algorithms you would apply, how you would approach the data set, how you could be sure that your algorithms worked, and anything else that you would do as part of the process. Your answer must describe specific features of this problem. [*10 marks*]

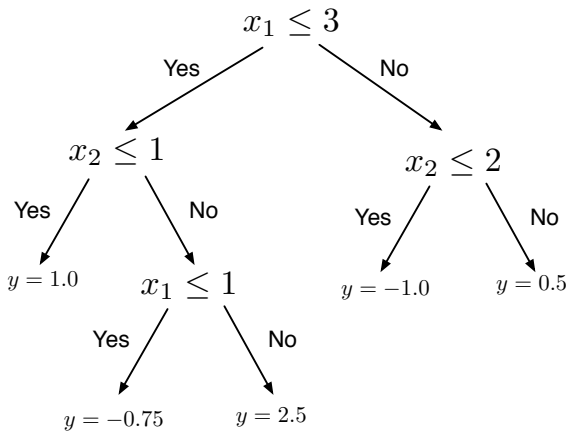
(b) You are working on a classification problem using Weka running on a single PC. A colleague suggests that you try using boosting with a three layer feedforward neural network as a base learner. Do you think this is a good idea? Why or why not? [*3 marks*]

(c) Consider another binary classification problem. You have collected 1000 training instances and 100 test instances. Your colleague has trained a logistic regression classifier with different choices of polynomial features, and got the following results on the test set: [*5 marks*]

Degree of polynomial	Test accuracy
1	75.3
2	76.7
3	77.1

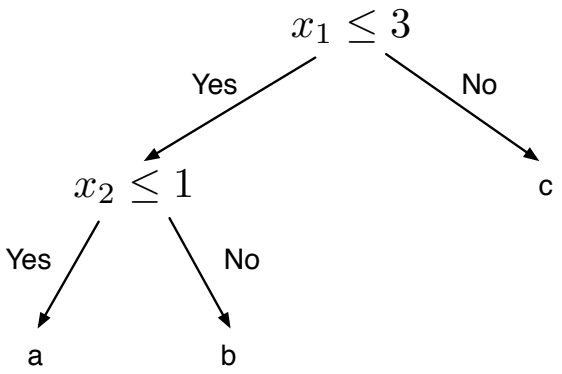
From this, your colleague has concluded that the order 3 polynomial features are best. Do you agree? Is there any other information you would want to collect to be sure?

(d) Consider the following regression tree. The input data are two-dimensional real-numbers $\mathbf{x} = (x_1, x_2)$ and the goal is to predict a real number y. [*3 marks*]



Draw the function that this tree represents. It will be simplest to make a two-dimensional graph with one axis for x_1 and one for x_2, and divide the space into regions according to the predictions of the tree.

(e) Consider a binary classification algorithm with two continuous features. Imagine that the following classification tree has resulted from one of the iterations of the boosting algorithm. [*4 marks*]



In this figure, $a, b, c \in \{-1, 1\}$ are the predictions that the tree will make at the leaves.

The training instances used to construct the tree were

$D_t(i)$	x_1	x_2	y
0.25	-1.5	1.5	+1
0.05	0.75	0.5	-1
0.1	-1.0	0.75	+1
0.4	2.5	1.75	-1
0.2	3.5	-1.0	-1

where $D_t(i)$ is the distribution over training instances that has been used for the current iteration of the boosting algorithm.

Using this training set, what are the correct values for a, b, and c? Give the value of the weighted error that would be used in the boosting algorithm.

2. (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) Give a description of the model used in the paper "User-Level Sentiment Analysis in Social Networks" by Tan et al. What are the main aspects of the problem that the authors intend to incorporate into their model? [*7 marks*]

   (e) In the paper "Connecting the Dots between News Articles" by Shahaf et al, describe what a "story" is. What are some characteristics of a good story? [*4 marks*]

3. (a) You are working on a text classification problem. You decide to use latent Dirichlet allocation (LDA) as a source of additional features. Explain how you would generate features from the output of LDA. (You do not need to explain how LDA works.) How would you choose the number of topics? [*6 marks*]

   (b) The following questions are about k-fold cross validation.

      i. Give a situation in which you might use cross validation. [*2 marks*]

      ii. What does k refer to? Why might you use a large k? A small one? [*3 marks*]

      iii. You have used cross validation to evaluate a classifier that you are about to deploy. You and your colleagues are having a lunchtime discussion in which you talk about how well the classifier will perform when it is deployed in the field. Your colleague Abel says he thinks the classifier will perform better than the cross validation procedure estimates. Another colleague Bluto believes that the classifier performs worse. Who do you agree with? Why? [*3 marks*]

   (c) Consider the paper "Factorization Meets the Neighborhood" by Koren. Why is it difficult to apply singular value decomposition (SVD) in a collaborative filtering context? How is this problem addressed in this paper? [*6 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 interactions? [*5 marks*]