UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR11073 MACHINE LEARNING AND PATTERN RECOGNITION (LEVEL 11)

Monday 5 th May 2014

09:30 to 11:30

INSTRUCTIONS TO CANDIDATES

Answer QUESTION 1 and ONE other question.

Question 1 is COMPULSORY.

All questions carry equal weight.

CALCULATORS MAY BE USED IN THIS EXAMINATION

Year 4 Courses

Convener: I. Stark
External Examiners: A. Cohn, T. Field

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

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

(a) Consider a binary classification problem. For each training instance i, where $i \in \{1, 2, \dots N\}$, the data contains two real-valued features, $x_{i1} \in \mathbb{R}$, and $x_{i2} \in \mathbb{R}$, and a binary class label $y_i \in \{0, 1\}$. You have reason to believe that given the class label, each of the feature values is distributed approximately according to a Gaussian distribution. *[4 marks]*

Describe a naive Bayes model for the situation above. Clearly indicate which are the parameters of the model. Give the equation for the log likelihood of this model. Your answer should be specific to this model.

(b) What assumption does the naive Bayes classifier make about the classification problem? If you thought that this assumption was unrealistic, how would you relax this assumption? *[3 marks]*

(c) You are working for an Internet advertising agency. One particular advertisement has been displayed to three different users, and each time you have recorded a binary value $x_i \in \{0, 1\}$, for $i \in \{1, 2, 3\}$, indicating whether the user clicked on the advertisement. Let π indicate the probability that a user, sampled at random from the entire population of users, would click on the ad.

i. What is the maximum likelihood estimator of π? *[2 marks]*

ii. Before you showed the advertisement to any users, you believed based on your experience that this advertisement is either highly effective (π very near 0.7) or highly ineffective (π very near 0.2). Describe how you could incorporate this information into your analysis. *[4 marks]*

iii. Compare the methods that you suggested in Question 1(c)i and Question 1(c)ii. In what situations might the method from Question 1(c)i be a better choice? In what situations might the method from Question 1(c)ii be a better choice? Are there situations in which both methods will perform similarly? Explain your answers. *[3 marks]*

(d) Briefly compare and contrast logistic regression and neural networks. Describe one aspect in which the two methods are similar. Describe one aspect in which the two methods are different. *[3 marks]*

(e) You play a game with your friend Bob, in which you bet on the outcome of a coin flip. The coin has been provided by Bob. You think that there is a 50% chance that Bob would have provided an unfair coin. If the coin is unfair, you have no knowledge of the probability that the coin will turn up heads, so if asked, you would model the distribution over this probability as uniform. You flip the coin once, and it comes up heads. What is the probability that the coin is fair? You flip the coin a second time, and it comes up heads again. What is the probability now that the coin is fair? Justify your answer. *[6 marks]*

2. **Neural Networks.** Consider the following data set. Here each row represents a data item in the training set, x_1 and x_2 are real-valued features, and $y \in \{0, 1\}$ is a binary class label.

x_1	x_2	y
0.2	1.1	1
0.8	0.8	1
1.0	0.2	1
0.4	0.3	1
0.5	0.7	0
0.7	0.2	0

(a) Consider the following neural network for this problem (with no hidden units). [*2 marks*]

$$y = \sigma(w_1 x_1 + w_2 x_2 + b),$$

where $w_1 = 1$, $w_2 = 0$, and $b = -0.45$, and $\sigma(z)$ represents the sigmoid function. What is the accuracy of the network on the classification problem above? Explain your answer.

(b) Is it possible to design a network without hidden units that has better accuracy? If so, describe a network that achieves this, including weights. If not, explain why not. [*2 marks*]

(c) Describe a two-layer network for this classification problem that has 100% accuracy on the data set above. *Hint:* It is possible to accomplish this with two hidden units. [*5 marks*]

(d) A colleague suggests that you might want to train a three-layer neural network on this data. Is this a good idea? Why or why not? [*3 marks*]

Variational Methods

(e) How does a variational approximation use the KL divergence to approximate a posterior distribution? Explain why a variational approximation leads to a lower bound on the likelihood. [*4 marks*]

(f) Let $p(x)$ the distribution function for the uniform distribution over the set $\{0, 1\}$. What distribution q minimizes $KL(q\|p)$? What distribution q maximises $KL(q\|p)$? Explain your answers. [*2 marks*]

(g) Give an example of a model where it would be reasonable to consider the use of a variational approximation. Why might it be helpful? [*4 marks*]

(h) What is an advantage of variational methods over Markov Chain Monte Carlo (MCMC)? What is an advantage of MCMC? [*3 marks*]

3. **Markov Chain Monte Carlo**

(a) What does "burn in" refer to in Markov Chain Monte Carlo methods? Why is it used? [*3 marks*]

(b) Consider a probability distribution over a parameter θ that is defined by the following probability density function [*5 marks*]

$$p(\theta) = \begin{cases} 2\theta & \text{if } \theta \in [0,1] \\ 0 & \text{otherwise.} \end{cases}$$

Describe how a slice sampler could be used to sample from this distribution. Your answer must be specific to this particular distribution. Include any equations that are required to apply the slice sampler in this context.

(c) Consider a Markov chain $x_1, x_2, \ldots$, where each $x_t \in \mathbb{R}$. The state of the chain x_{t+1} is sampled from the previous state x_t as follows. First, a value z is sampled from a standard normal distribution. Then the value a is calculated as [*5 marks*]

$$a = \min\left\{1, \exp\left\{\frac{z^2}{4} - \frac{x_t^2}{4}\right\}\right\}.$$

Then finally, with probability a, x_{t+1} is set to z. Otherwise, x_{t+1} is set to x_t. What is the stationary distribution of this chain? Justify your answer.

Optimization

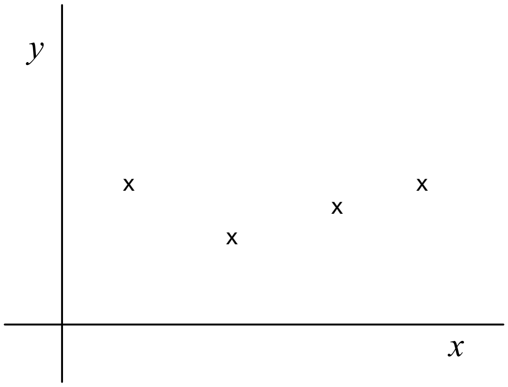
(d) Give a specific example of an optimization problem on which the gradient descent algorithm will perform well. Give an example of a problem on which gradient descent will perform poorly. Explain your answers. [*3 marks*]

(e) For what types of problems is the online gradient descent algorithm used? Why? [*2 marks*]

Regression

For the next questions, consider the data set below. Here the goal is to use regression to predict y from x. Imagine that you have available an existing code base that supports multivariate linear regression (and *only* linear regression).

QUESTION CONTINUES ON NEXT PAGE



(f) Describe in equations the linear regression model for this data set. Describe a function that, when minimized, could be used to fit the regression line. [*2 marks*]

(g) You have trained a linear regression model on this data using the existing code base that you have available (see the introduction to this set of questions). A colleague suggests that y might be a quadratic function of x. Is it possible to re-use the existing code in order to fit a quadratic function to this data set? If so, explain how. If not, explain why not. [*2 marks*]

(h) Which function would you prefer, the linear function, the quadratic function, or a different functional form entirely? Why? What more could you do to be sure? [*3 marks*]