Queen Mary
University of London

BSc/MSci EXAMINATION

PHY-328 Statistical Data Analysis: Exam Paper

Time Allowed: 2 hours 30 minutes

Date: 26^{th} May, 2011

Time: 10:00 - 12:30

Instructions: **Answer ALL questions in section A. Answer ONLY TWO questions from section B. Section A carries 50 marks, each question in section B carries 25 marks. An indicative marking-scheme is shown in square brackets [ ] after each part of a question. Course work comprises 20% of the final mark.**

You may wish to use the following information:-

Numeric calculators are permitted in this examination. Please state on your answer book the name and type of machine used. Complete all rough workings in the answer book and cross through any work which is not to be assessed.

Important Note: The academic regulations state that possession of unauthorised material at any time when a student is under examination conditions is an assessment offense and can lead to expulsion from the college. Please check now to ensure that you do not have any notes in your possession. If you have any then please raise your hand and give them to an invigilator immediately. Exam papers cannot be removed from the exam room

You are not permitted to read the contents of this question paper until instructed to do so by an invigilator.

Examiners: Dr. A. Bevan, Dr. J. Wilson

This (reverse cover) page is left blank.

SECTION A. Attempt answers to all questions.

A1

(a) Write down Bayes theorem and explain each term.

(b) Using Bayes theorem determine the probability that it will rain tomorrow given the following: The weather forecast tomorrow is for rain. 70% of the time when it rains, the rain has been correctly forecast. When there is no rain forecast, it rains 5% of the time, and on average it rains 100 days of the year.

(c) Compare your result with one assuming that the priors for it to rain or not are equal.

[10]

A2

(a) Write down the equation for the arithmetic average of a sample of data.

(b) Write down the equation for the standard deviation of a sample of data and explain the significance of the Bessel correction factor.

(c) Write down the equation for the skew γ of a sample of data, and illustrate this with a sketch of a distribution skewed to the left.

(d) Write down the equations for the covariance and Pearsons correlation coefficient for two variables x and y.

(e) Compute the arithmetic average, variance, standard deviation, and skew of the data set $\Omega(x) = \{1.0, 2.5, 3.0, 4.0, 4.5, 6.0\}$.

[10]

A3

(a) What percentage of results would one expect to obtain within $\pm 3\sigma$ of the mean value of an observable when making a set of measurements?

(b) Write down the equation for the binomial error on an efficiency, explaining the meaning of any parameters included.

(c) What benefit is there from performing a blind analysis of data?

(d) Given that $f = f(x, y)$, where x and y are correlated, write down the equation relating the variance on f, σ_f^2, in terms of on σ_x and σ_y and the partial derivatives of f. Also express this in matrix form identifying the significance of the off-diagonal terms.

[10]

A4

(a) When might it be useful to perform a fit to data.

(b) Write down the equation for a χ^2 sum given a data set $\Omega(\underline{x})$ and a model for the data $\theta(\underline{x})$.

(c) Explain the term "number of degrees of freedom" and how this relates to the number of data N in a data set.

(d) Write down the equation for a likelihood used to discriminate between a signal component described by $\mathcal{P}_{sig}$ and a background component described by $\mathcal{P}_{bg}$.

(e) What is usually minimised in a maximum-likelihood fit?

[10]

A5

(a) Describe a Bayesian classifier.

(b) Write down the classification algorithm corresponding to Fisher's linear discriminant.

(c) Describe a decision tree classifier to distinguish between two categories of event.

(d) Given several classifiers of data optimised to distinguish between two categories (A and B), sketch the distribution of the efficiency of A versus the efficiency of B, and note how one would select the most performant classifier using this plot.

(e) When might one decide not to use the most performant classifier as selected by the method illustrated in part **(d)**.

[10]

SECTION B. Answer two of the three questions in this section.

B1

a) Describe the process of testing a null hypothesis, and subsequently placing a confidence level on the conclusions drawn. **[5]**

b) A test for an infection returns a positive result 98% of the time for someone with an infection. The same test reports a positive result 0.05% of the time for patients who are not infected. If, 0.01% of the population are infected, what is the probability that someone with a positive test result is actually infected? Is this a good test? **[5]**

c) Describe how you would uses Bayes theorem to compare two theoretical hypotheses and how one would interpret the results obtained. **[5]**

d) As the result of a survey, you have measured the baseline between two mountain summits A and B to be $2\,km$, and from both of those summits you have determined the angles α and β between the baseline and a third summit C. Outline how you would use Bayes theorem to determine the location of the third summit relative to A, and illustrate how one would determine both the most probable value of the location and the confidence interval associated with a 1σ statistical uncertainty. **[10]**

[TOTAL FOR B1 = 25]

B2

a) Determine the change in $-\ln\mathcal{L}$ from the best fit value that corresponds to a 1σ Gaussian uncertainty. **[5]**

b) Given the measurements $x_1 = 1.2 \pm 0.3$, and $x_2 = 1.8 \pm 0.3$, approximate the mean value and uncertainty on the average value of x obtained by performing a χ^2 scan from 1.0 to 2.0 in steps of 0.2. Compare the results you've obtained with those from a weighted average. **[10]**

c) Assuming that $y = ax^2 + b$, use the method of least squares to derive the values of coefficients a and b. Assume that the uncertainties on each of the data points to be used by the least squares method are all the same. **[10]**

[TOTAL FOR B2 = 25]

B3

a) Compute the coefficients of a Fisher discriminant to optimally separate samples of data A and B given the means and covariance matrices below.

$$\mu_A = \begin{pmatrix} 1.0 \\ 2.0 \end{pmatrix}, \tag{1}$$

and

$$\sigma_A = \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix},$$

for sample A, and:

$$\mu_B = \begin{pmatrix} 0.0 \\ 1.0 \end{pmatrix}, \tag{2}$$

and

$$\sigma_B = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix},$$

for sample B.

[10]

b) Describe a perceptron, and give three possible forms for the activation function. **[5]**

c) Describe how to construct a Multi-layer perceptron (MLP). Note the configuration of a simple MLP, and how the weights can be determined and validated. **[10]**

[TOTAL FOR B3 = 25]