Queen Mary
University of London

BSc/MSci EXAMINATION

PHY-328 Statistical Data Analysis: SOLUTIONS

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.

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}, \tag{1}$$

where $A = data$, $B = theory/hypothesis$, and

$$P(A) = \sum_i P(A|B_i)P(B_i). \tag{2}$$

The terms have the following meanings: $P(B|A)$ is the posterior probability of the theory given the data, $P(A|B_i)$ is the probability of the data given the theory B_i, $P(B)$ is the prior, and $P(A)$ is the total probability of the data based on the theory. **[4 marks in total for this information].**

(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.

$$\begin{aligned} P(rain|rain\,forecast) &= 0.7, & (3)\\ P(rain|no\,rain\,forecast) &= 0.05, & (4)\\ P(rain) &= 100/365 = 0.274, & (5)\\ P(no\,rain) &= 265/365 = 0.726, & (6) \end{aligned}$$

so

$$\begin{aligned} P(rain\,forecast|rain) &= \frac{0.7 \times 0.274}{0.7 \times 0.274 + 0.05 \times 0.726}, & (7)\\ &= 0.841 & (8) \end{aligned}$$

[4 marks in total for this information].

(c) Compare you result with one assuming that the priors for it to rain or not are equal. As above, but with $P(rain) = P(no\,rain) = 0.5$, hence

$$\begin{aligned} P(rain\,forecast|rain) &= \frac{0.7 \times 0.5}{0.7 \times 0.5 + 0.05 \times 0.5}, & (9)\\ &= 0.933 & (10) \end{aligned}$$

so either way round, we expect it to rain.
[2 marks in total for this information].

[10]

A2

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

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{11}$$

[2 marks in total for this information].

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

$$\sigma_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}. \tag{12}$$

The Bessel correction factor $\frac{N}{N-1}$ used to get this form of the standard deviation is introduced in order for σ_x to be unbiased for small N.
[2 marks in total for this information].

(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.

$$\gamma = \frac{1}{N\sigma^3}\sum_{i=1}^{N}(x_i - \overline{x})^3. \tag{13}$$

Figure 1 shows an example of a distribution that is skewed to the left.

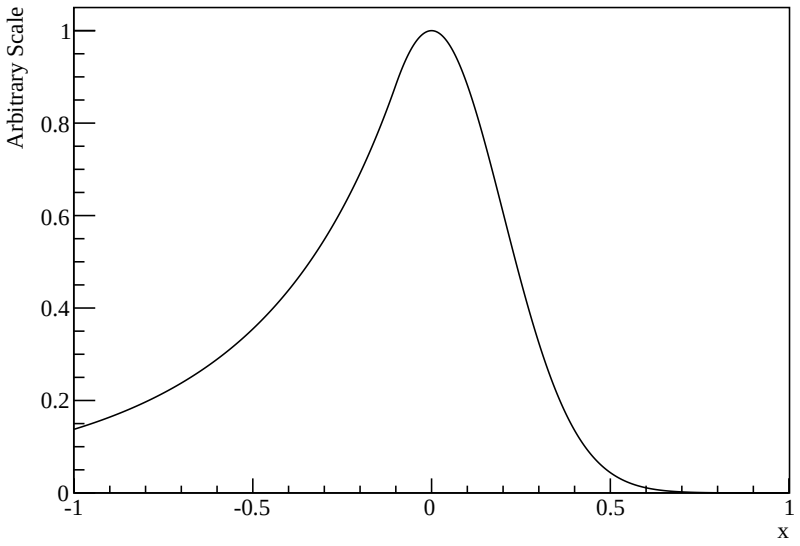

Figure 1: A distribution that is skewed to the left.

[2 marks in total for this information].

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

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}), \tag{14}$$

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}. \tag{15}$$

[2 marks in total for this information].

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

x	$(x - \overline{x})$	$(x - \overline{x})^2$	$(x - \overline{x})^3$
1.0	-2.5	6.25	-15.625
2.5	-1.0	1.00	-1.000
3.0	-0.5	0.25	-0.125
4.0	0.5	0.25	0.125
4.5	1.0	1.00	1.000
6.0	2.5	6.25	15.625

$$\overline{x} = 3.5 \tag{16}$$

$$\sigma_x^2 = 3.0 \tag{17}$$

$$\sigma_x = 1.73 \tag{18}$$

$$\gamma = 0.0 \tag{19}$$

If no bessel correction factor was used for the variance, then deduct 1/2 a mark (in that case one would have calculated $\sigma_x^2 = 2.5$ and $\sigma_x = 1.58$.
[2 marks in total for this information].

[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?
One would expect 99.73% of results to lie within 3σ.
[1 mark in total for this information].

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

$$\sigma_\epsilon = \sqrt{\frac{\epsilon(1-\epsilon)}{N}}. \tag{20}$$

where ϵ is the efficiency (number of interesting events divided by N), and N is the total number of events.
[3 marks in total for this information].

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

- The analysis will provide an objective result, i.e. there should be no experimeter bias.

[1 mark in total for this information].

(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.
The normal form of σ_f^2 is

$$\sigma_f^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2 + 2\frac{\partial f}{\partial x}\frac{\partial f}{\partial y}\sigma_{xy}. \tag{21}$$

In matrix form this is

$$\sigma_f^2 = ([x-\overline{x}],[y-\overline{y}]) \begin{pmatrix} \left(\frac{\partial f}{\partial x}\right)^2 & \frac{\partial f}{\partial x}\frac{\partial f}{\partial y} \\ \frac{\partial f}{\partial x}\frac{\partial f}{\partial y} & \left(\frac{\partial f}{\partial y}\right)^2 \end{pmatrix} \begin{pmatrix} x-\overline{x} \\ y-\overline{y} \end{pmatrix} \tag{22}$$

and the off-diagonal terms are the covariance terms between x and y.
[5 marks in total for this information: 2 each for the variance on f in matrix and normal form, and one mark for an explanation of the off-diagonal terms].

[10]

A4

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

- Useful to fit data in order to extract the values of one or more parameters (with uncertainties).
- Useful to obtain a multi-dimensional confidence interval.
- Useful to average results that are correlated

[**2 marks in total for this information: 1 for each of the above**]

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

$$\chi^2 = \sum_{i=1}^{N} \frac{[\underline{x}_i - \theta(\underline{x}_i)]^2}{\sigma_i^2}. \tag{23}$$

where σ_i is the error on the datum $\underline{x}_i$.
[**2 marks in total for this information**]

(c) Explain the term "number of degrees of freedom" and how this relates to the number of data N in a data set.
The number of degrees of freedom is the total number of data minus the total number of constraints on the data. If there are no external constraints imposed, then the total number of data constitutes the only constraint on the problem, thus $\nu = N - 1$.
[**2 marks in total for this information**]

(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 $\mathcal{P}_{bg}$.

$$\mathcal{L} = \prod_{i=1}^{N} \mathcal{L}_i, \tag{24}$$

$$\mathcal{L}_i = f_{sig}\mathcal{P}_{sig} + f_{bg}\mathcal{P}_{bg}. \tag{25}$$

Here f_{sig} and f_{bg} are the fractions of signal and background components, respectively, and we require $f_{sig} + f_{bg} = 1$. [**2 marks in total for this information**]

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

$$-\ln \mathcal{L} = -\sum_{i=1}^{N} \ln \mathcal{L}_i, \tag{26}$$

$$\tag{27}$$

[**2 marks in total for this information**]

[**10**]

A5

(a) Describe a Bayesian classifier.
Given that $P(B|A) \propto P(A|B)P(B)$, for some hypothesis B and data A, one can compute $P(B_i|A)$ or ratios of $P(A|B_i)P(B_i)$. The event is then assigned classification of the most probable hypothesis based on these ratios or posterior probabilities.
[**2 marks in total for this information**]

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

$$F = \sum_{i=1}^{m} \alpha_i x_i + \beta, \tag{28}$$

$$\alpha \propto W^{-1}(\mu_A - \mu_B), \tag{29}$$

where the α_i are weight coefficients, β is an arbitrary offset, W is the sum of covariances for samples A and B, and μ_i is the vector of mean values of sample A or B.
[**2 marks in total for this information**]

(c) Describe a decision tree classifier to distinguish between two categories of event.
The decision tree is a collection of linked nodes, where at each node one performs

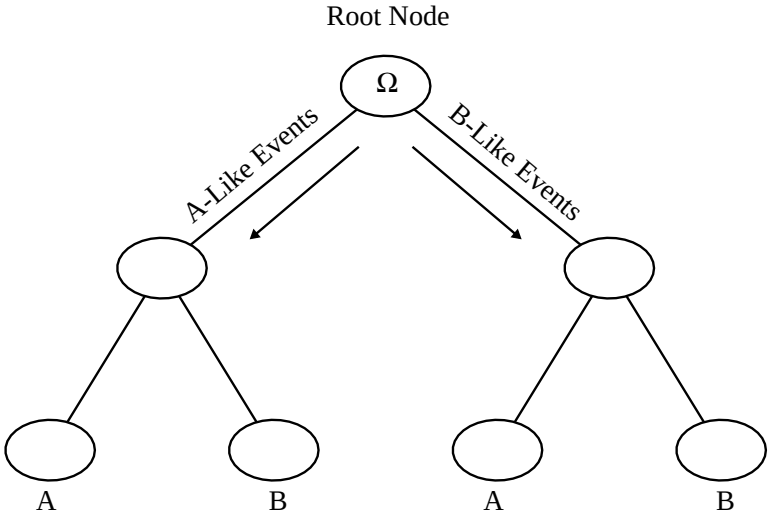

Figure 2: Schematic of a decision tree to classify two types of event A and B.

a cut based optimisation to separate different classifications of events. Some or all of the discriminating variables may be used at each node of the tree. Starting from the ROOT node, one can work down through several layers of categorisation steps (See Figure 2).
[**2 marks in total for this information**]

(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.
See Figure 3. The best classifier is the one with the efficiency curve that passes closest to the bottom right (signal efficiency high, background efficiency low).
[**2 marks in total for this information**]

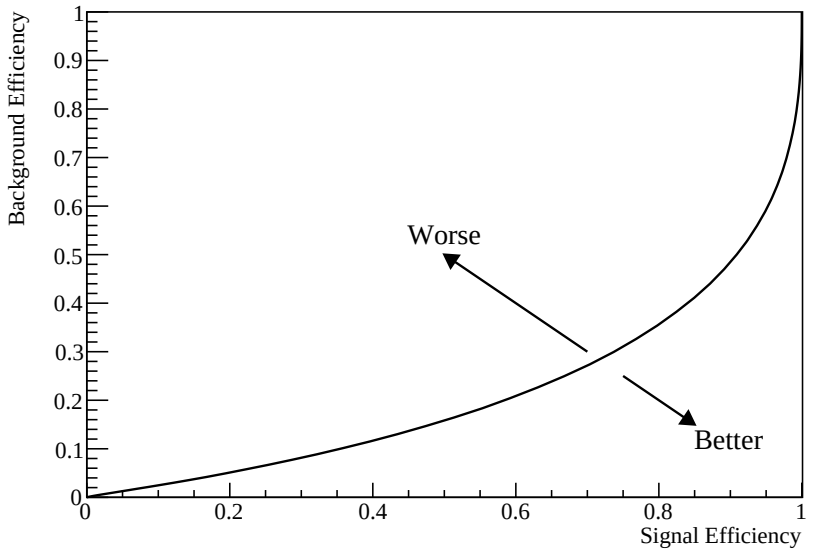

Figure 3: Deciding which classifier is better... where $A = signal$, $B = background$.

(e) When might one decide not to use the most performant classifier as selected by the method illustrated in part **(d)**.
Might not use the most performant classifier identified in the previous part of this question when that is

- Difficult to use in subsequent analysis (e.g. hard to define a sensible PDF to use in a fit).
- not understood
- not validated

[2 marks in total for this information, 1 mark for each of the points]

[10]

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

B1

a) Describe the procees of testing a null hypothesis, and subsequently placing a confidence level on the conclusions drawn.
In order to test a null hypothesis H_0 against some data, one has to

- Clearly define H_0.
- Thus define $H_1 = \overline{H_0}$.
- Define the confidence level that we will test the hypothesis against, i.e. what level of false negatives and false positives is acceptible for the problem.
- Make the comparison.
- If the data are found to agree with the hypothesis, then the CL corresponding to that level of agreement can be quoted. Otherwise the level at which H_0 disagrees with the data can be reported.

[5 marks in total for this information, 1 mark for each of the points] **[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?

$$\begin{aligned} P(positive\,test\,result|infected) &= 0.98, & (30)\\ P(positive\,test\,result|healthy) &= 0.0005, & (31)\\ P(infected) &= 0.0001, & (32)\\ P(healthy) &= 0.9999, & (33)\\ & & (34) \end{aligned}$$

so

$$\begin{aligned} P(infected|positive\,test\,result) &= \frac{0.98 \times 0.0001}{0.98 \times 0.0001 + 0.0005 \times 0.9999}, & (35)\\ &= 0.164. & (36) \end{aligned}$$

Therefore this particular test is not particularly good.
[5 marks in total for this information **[5]**

c) Describe how you would uses Bayes theorem to compare two theoretical hypotheses and how one would interpret the results obtained.
Note that $P(B_i|A) \propto P(A|B_i)P(B_i)$ for a given hypothesis B_i and data A. Therefore one can compare the posterior probabilities $P(B_i|A)$ or the ratios of $P(A|B_i)P(B_i)$ if one can not determine the normalisation constants. For example, given two scenarios H_0 and H_1, the ratio is

$$R = \frac{P(H_0|data)}{P(H_1|data)} \tag{37}$$

$$= \frac{P(data|H_0)P(H_0)}{P(data|H_1)P(H_1)}. \tag{38}$$

Thus if

- $R > 1$, H_0 is preferred.
- $R < 1$, H_1 is preferred.
- $R \simeq 1$, insufficient information exists to decide which hypothesis is preferred.

[5 marks in total for this information **[5]**

d) THIS PART OF THE QUESTION IS UNSEEN
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 the both the most probable value of the location and the confidence interval associated with a 1σ statistical uncertainty.
Starting from Bayes theorem, where

$$P(H_0|data) = \frac{P(data|H_0)P(H_0)}{P(data)} \tag{39}$$

for some given point in space $H_0 = (x, y)$ (see Figure 4), we can compute a posterior probability for α using

$$P(\hat{\alpha}|\alpha) = \frac{P(\alpha|\hat{\alpha})P(\hat{\alpha})}{P(\alpha)}, \tag{40}$$

where

$$\hat{\alpha} = \arctan(y/x), \tag{41}$$

and similarly for β we can obtain the posterior probability

$$P(\hat{\beta}|\beta) = \frac{P(\beta|\hat{\beta})P(\hat{\beta})}{P(\beta)}, \tag{42}$$

where

$$\hat{\beta} = \arctan(y/[2-x]), \tag{43}$$

where the probabilities $P(\alpha|\hat{\alpha})$ and $P(\beta|\hat{\beta})$ are Gaussian distributions with means and uncertainties corresponding to the measured values and errors of α and β, respectively. As we are otherwise ignorant, we can assume that we have uniform priors.

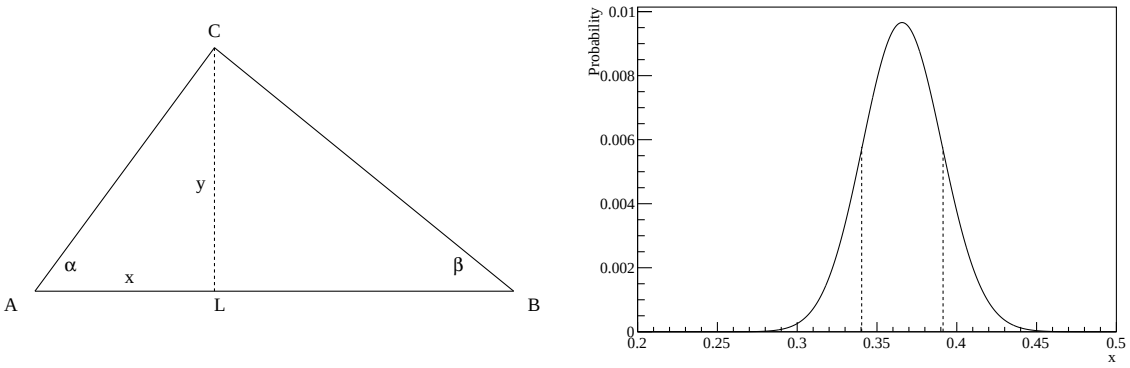

Figure 4: (left) a schematic of the problem indicating the values of x and y. Here $L = 2\,km$. (right) an illustration of the posterior probability obtained, and the corresponding 1σ interval.

The total probability for any given location (x, y) for the apex of the triangle is the product of the two posteriror probabilities: $P_C = P(\hat{\alpha}|\alpha)P(\hat{\beta}|\beta)$. Given this two dimensional probability (in terms of $(\hat{\alpha}, \hat{\beta})$ or (x, y) space, we can integrate out the posterior probability for x or y.

The most probable value of the coordinate is the seed with which we construct the interval about. Starting from this point, one integrates out (integrating over contours of equal probability), until one obtains the desired coverage. For example see Figure 4.

[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.

$$\mathcal{L}_i = P_i, \tag{44}$$
$$\mathcal{L} = \prod_i \mathcal{L}_i, \tag{45}$$

where $P_i = 1/(\sigma\sqrt{2\pi})\exp[-(x-\mu)^2/2\sigma^2]$. The $-\ln\mathcal{L}$ is minimised, so

$$-\ln\mathcal{L} = \sum_i -\ln\left[\frac{1}{\sigma\sqrt{2\pi}}\right] + \frac{(x-\mu)^2}{2\sigma^2}, \tag{46}$$

- The first term corresponds to the value of $-\ln\mathcal{L}$ at the minimum, $\mathcal{L}_0$, when $x = \mu$.
- When $x - \mu$ corresponds to $\pm 1\sigma$, then the second term is 1/2.

Thus a change of 1/2 from the best fit value $\mathcal{L}_0$ in a maximum-likelihood fit 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.

x	χ^2_1	χ^2_2	χ^2_{TOT}
1.0	0.44	7.11	7.56
1.2	0.00	4.00	4.00
1.4	0.44	1.78	2.22
1.6	1.78	0.44	2.22
1.8	4.00	0.00	4.00
2.0	7.11	0.44	7.56
(1.5)	(1.00)	(1.00)	(2.00)

So the mean value of these two measurements is $x = 1.5$. If one sketches the points shown, then a reasonable estiamte of the error obtained from that curve, is when the χ^2 changes by one from the minimum. This is about ± 0.2.

If one performs a crosscheck by computing a weighted average, where this is given by

$$\overline{x} \pm \sigma_x = \frac{x_1/\sigma_1^2 + x_2/\sigma_2^2}{1/\sigma_2^2 + 1/\sigma_2^2} \pm \left(1/\sigma_2^2 + 1/\sigma_2^2\right)^{-1/2}, \tag{47}$$
$$= \frac{\sigma_2^2 x_1 + \sigma_1^2 x_2}{\sigma_1^2 + \sigma_2^2} \pm \left(\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\right)^{1/2}. \tag{48}$$

one finds that $x = 1.5 \pm 0.21$.

The estimate from the χ^2 scan is in good agreement with this computation of weighted average.

[10]

c) THIS PART OF THE QUESTION IS UNSEEN

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.

$$\chi^2 = \sum_{i=1}^{N} \left[\frac{y_i - ax_i^2 - b}{\sigma_i} \right]^2 \tag{49}$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{N} \left[y_i - ax_i^2 - b \right]^2 \tag{50}$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{N} \left[y_i^2 + a^2 x_i^4 + b^2 - 2ay_i x_i^2 - 2by_i + 2abx_i^2 \right] \tag{51}$$

$$= \frac{N}{\sigma^2} \left[\overline{y^2} + a^2 \overline{x^4} + b^2 - 2a\overline{yx^2} - 2b\overline{y} + 2ab\overline{x^2} \right] \tag{52}$$

$$\tag{53}$$

The optimal solution exists for

$$\frac{\partial \chi^2}{\partial a} = 0, \tag{54}$$

$$\frac{\partial \chi^2}{\partial b} = 0, \tag{55}$$

so

$$\frac{\partial \chi^2}{\partial a} = \frac{N}{\sigma^2} \left[2a\overline{x^4} - 2\overline{yx^2} + 2b\overline{x^2} \right], \tag{56}$$

$$\frac{\partial \chi^2}{\partial b} = \frac{N}{\sigma^2} \left[2b - 2\overline{y} + 2a\overline{x^2} \right], \tag{57}$$

hence

$$a = \frac{\overline{yx^2} - \overline{y}\,\overline{x^2}}{\overline{x^4} - (\overline{x^2})^2} \tag{58}$$

$$b = \overline{y} - a\overline{x^2}. \tag{59}$$

[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.

$$\mu_A = \begin{pmatrix} 1.0 \\ 2.0 \end{pmatrix}, \tag{60}$$

and

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

for sample A, and:

$$\mu_B = \begin{pmatrix} 0.0 \\ 1.0 \end{pmatrix}, \tag{61}$$

and

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

for sample B.

$$\Delta\mu = \mu_A - \mu_B \tag{62}$$

$$= \begin{pmatrix} 1.0 \\ 1.0 \end{pmatrix} \tag{63}$$

$\det W = -1$ as.

$$W = \begin{pmatrix} 2 & 3 \\ 3 & 4 \end{pmatrix},$$

and

$$W^{-1} = \begin{pmatrix} -4 & 3 \\ 3 & -2 \end{pmatrix},$$

As $\alpha \propto W^{-1}\Delta\mu$, we find

$$\alpha \propto \begin{pmatrix} -1.0 \\ 1.0 \end{pmatrix} \tag{64}$$

[10]

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

- Figure 5 illustrates a perceptron, which is an analogy of a neuron. Taking n inputs, processing these inputs via an activation function, and providing some output.

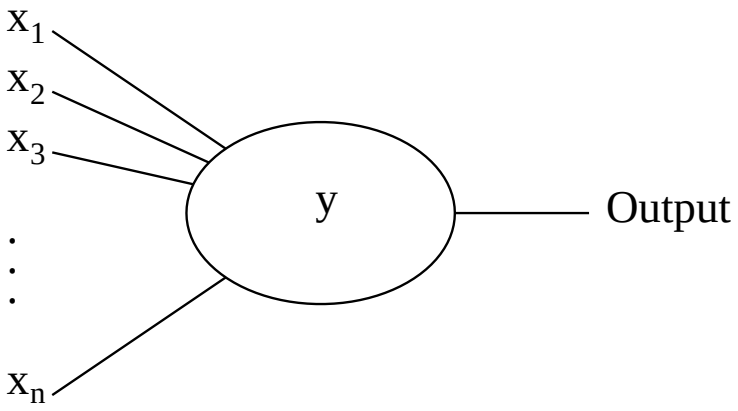

Figure 5: A perceptron.

Possible activation functions include

- Binary threshold: $y_i = 1$ if $\underline{w}.\underline{x} > 0$, else $y_i = 0$.
- Sigmoid (or Logistic): $y_i = 1/(e^{\underline{w}.\underline{x}+\beta} + 1)$
- Radial: $y_i = e^{-\underline{w}.\underline{x}}$
- hyperbolic tan: $y_i = \tanh(\underline{w}.\underline{x})$

[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.
A multi-layer perceptron is a neural network with one or more hidden layers of perceptrons acting as nodes within the net. This is shown schematically in Fig 6.

- A training sample is required with $\underline{x}$ variables for each event e_i to be used in the determination of the weights $\underline{w}$. In addition the true target value t_i for each event must be known as this is a supervised learning algorithm.
- Then we can define a mis-classification error ϵ_i for an event

$$\epsilon_i = \frac{1}{2}(y_i - t_i)^2 \tag{65}$$

where the y_i is the activation function output (assumes this varies between 0 and 1). The total error on the training sample is therefore given by

$$E = \sum_i \epsilon_i = \sum_i \frac{1}{2}(y_i - t_i)^2 \tag{66}$$

where the sum is over all data.

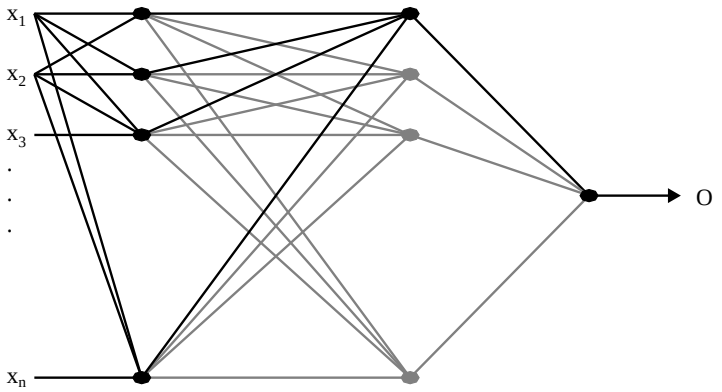

Figure 6: An MLP with one hidden layer.

- Now we can guess an intial set of weights, and based on this, use the error to determine the next iteration of weights via

$$\underline{w}_{j+1} = \underline{w}_j + \Delta\underline{w}, \tag{67}$$

$$\Delta\underline{w} = -\alpha\frac{dE}{d\underline{w}}. \tag{68}$$

  where α is a small positive learning rate. The logic behind this is that in using this definition of $\Delta\underline{w}$, the $j + i^{th}$ set of weights will have an error less than the j^{th} set as:

$$\Delta E = \Delta\underline{w}\frac{dE}{d\underline{w}}, \tag{69}$$

$$= -\alpha\left(\frac{dE}{d\underline{w}}\right)^2 \tag{70}$$

$$\tag{71}$$

- The total error for the network is the sum of errors for the input, hidden and output layers of notes, and this is minimised through training.
- Having determined the error rate on a training sample, one should process a validation sample of events with the same weights as the training sample, in order to determine a differnece in errors between training and validation sets. This should be small if the network configuration is not overtrainined, and the total error difference can be used as a criterion to stop the training process.

2 points for each of these issues **[10]**

[TOTAL FOR B3 = 25]