Project Guidelines for SCI5020 – Principles of Statistical Inference – Part 1

Lecturers: Dr. Monique Borg Inguanez, Dr. Fiona Sammut and Dr. David Suda

Deadline for the Project (Part 1) is Strictly: noon, Friday 2nd May, 2025.

The project for SCI5020 is made up of two parts. Each part accounts for 50% of the total mark.

You are required to work individually. Any questions related to this part of the project should be sent by Monday 28th April 2025. For queries related to Questions 1 – 3 please contact Dr. Fiona Sammut. For queries related to Question 4 – 5 please contact Dr. David Suda.

Question 1 - [24 marks]

Consider a random variable *X* that follows an exponential distribution with scale parameter λ.

i. Give reference to a publication in which the exponential distribution has been used in practise. Explain the context in which this distribution has been used in this publication. **[2 marks]**
ii. State the mean and the variance of *X*. **[1 mark]**
iii. Derive the moment estimator of λ. **[2 marks]**
iv. Use the second moment to obtain another estimator of λ. **[3 marks]**
v. Comment on the unbiasedness and consistency of the moment estimator for λ derived in Q1iii). State any assumption/s that need to be made to check for unbiasedness and consistency. **[2, 1 marks]**
vi. Use R software to generate 1000 data points from an exponential distributed random variable using any admissible parameter value for λ.
- Write down the log-likelihood function for this exponentially distributed random variable.
- Evaluate the log-likelihood function for the generated data as a function of λ, and plot the resulting log-likelihood function against different values of λ. Present the plot together with the answers.
- Using the plot or otherwise, which estimate for λ is the MLE? Give a reason for your answer.

The R code used for this question should be presented together with the answers. **[3, 3, 5, 2 marks]**

Question 2 - [16 marks]

Suppose that we wish to estimate the parameters of the model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}{}^3 + \varepsilon_i$$

given a sample of size *n.*

i. State the type of estimator that needs to be used in this situation and mention two properties of this estimator. **[2 marks]**
ii. Derive the equations that need to be solved to obtain the required estimates. **[6 marks]**
iii. Generate or source a dataset for which you would use this type of model and fit the model to the data. Present a plot of the fitted model to the data. Comment on the goodness of fit of the model to the data. The R code used for this question should be presented together with the answers. Proper referencing should be provided in the text if the data is sourced. **[4, 2, 2 marks]**

Question 3 - [20 marks]

Let $x_1,...,x_n$ be a random sample selected from a population with a distribution of your choice. [This distribution needs to be different from those used in the lecture notes].

i. Derive the maximum likelihood estimator/s of the parameters of the chosen distribution. **[6 marks]**
ii. Find the Cramer-Rao lower bound for the maximum likelihood estimator/s obtained in Q3i).

[3 marks]

iii. Does/do the ML estimator/s obtained in Q3i) attain the Cramer-Rao lower bound? Give a reason for your answer. **[4 marks]**
iv. Is/are the ML estimator/s obtained in Q3i) a sufficient statistic for the population parameter/s? Give a reason for your answer. **[3 marks]**
v. Why are maximum likelihood estimators considered to be desirable estimators? Mention one situation where a maximum likelihood estimator might not be optimal. **[3, 1 marks]**

Question 4 – Jackknife and Bootstrap

Consider 50 observations of bivariate pair (X, Y) in *resampling.xlsx*. Use the *nls* command in R to estimate the nonlinear regression $Y = \frac{aX}{b+X} + \epsilon$. Furthermore:

a. Construct a computer code in R to find the Jackknife and Bootstrap estimators of a and b. In the case of Jackknife, section randomly the sampling into 5 partitions of size 10. In the case of Bootstrap, generate 1000 samples of size 100 with replacement.

b. For the Jackknife estimator, find a 95% confidence interval using the normal distribution and the t-distribution. For the Bootstrap estimator, find a 95% normal, t and empirical confidence intervals.

[20 marks]

Question 5 – The EM Algorithm

Consider a univariate K - Gaussian mixture model with probability density function:

$$f(x) = \sum_{l=1}^{K} \pi_l \phi(x, \mu_l, \sigma_l)$$

such that $\sum_{l=1}^{K} \pi_l = 1$ and $\pi_l > 0$ for all l, and where $\phi(x, \mu, \sigma)$. The EM algorithm for this works as follows:

1. Initialise $\mu_1^{(0)}, \dots, \mu_k^{(0)}, \sigma_1^{(0)}, \dots, \sigma_k^{(0)}, \pi_1^{(0)}, \dots, \pi_k^{(0)}$.

2. Let $\gamma_n^{(j,k)} = \frac{\pi_k \phi\left(x_n \middle| \mu_k^{(j-1)}, \sigma_k^{(j-1)}\right)}{\sum_{l=1}^{K} \pi_l \phi\left(x_n \middle| \mu_l^{(j-1)}, \sigma_l^{(j-1)}\right)}$.
3. Let $\mu_k^{(j)} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_n^{(j,k)} x_n$, $\sigma_k^{(j)} = \sqrt{\frac{1}{N_k} \sum_{n=1}^{N} \gamma_n^{(j,k)} \left(x_n - \mu_k^{(j)}\right)^2}$ and $\pi_k^{(j)} = \frac{N_{jk}}{N}$ with $N_{jk} = \sum_{i=1}^{N} \gamma_n^{(j,k)}$.

Do the following:

1. Simulate 1000 readings from a mixture Gaussian distribution with 3 or more Gaussians. The following is a sample R code for simulating from a mixture of three Gaussians: N(2,1), N(10,2) and N(-3, 0.5) with weights 0.3, 0.3 and 0.4 respectively.

```
nsim<-1000

u<-numeric(nsim)

x<-numeric(nsim)

pi1<-0.3

pi2<-0.3

pi3<-1-pi1-pi2

for (i in 1:nsim){

        u[i]<-runif(1,0,1)

        if (u[i]>=0&u[i]<pi1) x[i]<-rnorm(1,mean=2,sd=1)

        if (u[i]>=pi1&u[i]<pi1+pi2) x[i]<-rnorm(1,mean=10,sd=2)

        if (u[i]>=pi1+pi2&u[i]<=pi1+pi2+pi3) x[i]<-rnorm(1,mean=-3,sd=0.5)

}
```

   Do not use the same code – create your own unique mixture Gaussian distribution and simulated set of values (do not collude with other participants of the course to use the same set of values). Adapt the above code as needed. Make sure the that Gaussian distributions considered are distinct enough for each other to ensure a multimodal distribution with number of modes equal to the number of Gaussian distributions being considered.

2. Determine initial values $\mu_1^{(0)}, \ldots, \mu_k^{(0)}, \sigma_1^{(0)}, \ldots, \sigma_k^{(0)}, \pi_1^{(0)}, \ldots, \pi_k^{(0)}$ using a K-means clustering approach or otherwise. Note that if you use a clustering approach for to obtain initial values, you need to set the number of clusters equal to the number of Gaussian distributions you are going to use.

3. Run the EM algorithm for a number of iterations.
   Print $\mu_1^{(j)}, \ldots, \mu_k^{(j)}, \sigma_1^{(j)}, \ldots, \sigma_k^{(j)}, \pi_1^{(j)}, \ldots, \pi_k^{(j)}$ for each iteration, and also the loglikelihood which is given by $\sum_{n=1}^{N} \ln\left(\sum_{l=1}^{K} \phi(x_i | \mu_l^{(j)}, \sigma_l^{(j)}\right)$. Determine when to stop the EM algorithm, either via a maximum number of iterations or through a convergence criterion, however ensure that the EM algorithm has converged. Plot the trajectory of the estimates and the likelihood by iteration to illustrate this.

4. Give the final estimated μ_k's, σ_k's and π_k's. How do these compare with the coefficients you chose in the simulation?

[20 marks]

Submission

A soft copy of the assignment should be sent by email to Dr. Fiona Sammut and Dr. David Suda by the given deadline. Hard copies of the assignment will not be accepted. In the correspondence sent, kindly cc our departmental administrator Ms Ann Zammit – administrator-stator.sci@um.edu.mt.