Queen Mary
University of London

Main Examination period 2023 – January – Semester A

MTH781P: Data Analytics

Duration: 4 hours

The exam is available for a period of **4 hours**, within which you must complete the assessment and submit your work. **Only one attempt is allowed – once you have submitted your work, it is final**.

All work should be **handwritten** and should **include your student number**.

You should attempt ALL questions. Marks available are shown next to the questions.

In completing this assessment:

- **You may use books and notes.**
- **You may use calculators and computers, but you must show your working for any calculations you do.**
- **You may use the Internet as a resource, but not to ask for the solution to an exam question or to copy any solution you find.**
- **You must not seek or obtain help from anyone else.**

When you have finished:

- scan your work, convert it to a **single PDF file**, and submit this file using the tool below the link to the exam;
- e-mail a copy to **maths@qmul.ac.uk** with your student number and the module code in the subject line;
- with your e-mail, include a photograph of the first page of your work together with either yourself or your student ID card.

Examiners: C. Beck, A. Baule

 Continue to next page

Question 1 [25 marks].

(a) State and explain the definition of statistical independence of two events A, B. [**3**]

(b) Suppose we throw a die. Define the events A to be "the outcome is an even number", B to be "the outcome is 3", C to be "the outcome is either a 5 or a 6".

(i) What is the sample space Ω? [**3**]

(ii) Are the events A and B independent? Explain your reasoning. [**3**]

(iii) Are the events A and C independent? Explain your reasoning. [**3**]

(iv) Are the events A,B and C independent? Explain your reasoning. [**3**]

(c) State and explain the Law of Total Probability. [**3**]

(d) Suppose that there are two urns. The first contains 5 red balls, 3 green balls, and 2 blue balls. The second contains 2 red balls and 4 green balls.

We pick a ball at random from the first urn and place it in the second urn. We then pick a ball at random from the second urn (which might be the ball we have just placed there). What is the probability that the ball is red? [**7**]

Question 2 [25 marks]. You are interested to buy a terraced house in East London. You perform a statistical analysis of house prices in a particular area of East London. From last year's data, the mean sales price for terraced houses has been determined as $\mu = 600450$ GBP. However, a recent sample of 37 new sales showed a sample mean of 655500 GBP with a sample standard deviation of $\sigma = 142000$ GBP.

(a) Explain the notion of a sampling distribution in this context. [**5**]

(b) What is the test statistic and the associated sampling distribution in the above situation? [**3**]

Assume now that we want to conduct a hypothesis test concerning our suspicion that $\mu \neq 600450$ GBP.

(c) What are appropriate null and alternative hypotheses in this situation? Is the test one-tailed or two-tailed? [**3**]

(d) Explain the notion of a p-value and determine the p-value of this hypothesis test. You can find a Standard Normal Z-Table in the appendix. [**6**]

(e) For a significance level of $\alpha = 0.05$, which conclusion can be drawn from the test? Explain your reasoning. [**3**]

(f) Explain the meaning of type 1 and type 2 errors. What are the business implications of making these errors in the context given here? [**5**]

 Continue to next page

Question 3 [27 marks]. A company wants to predict sales as a function of various other quantities including advertising spend, and whether it is summer (coded as 1 for summer, 0 for the rest of the year). These are named `sales`, `ad_spend`, `summer` in the dataset in R, and the dataset is called `sales_data`.
The following R commands are run:

```
model1 = lm(sales~ad_spend+summer, data=sales_data)
summary(model1)
```

Part of the output is:

```
Coefficients:
             Estimate  Std.  Error  t value  Pr(>|t|)
(Intercept)  579.33464     19.26756   30.068    <2e-16  ***
ad_spend       0.73191      0.03334   21.956    <2e-16  ***
summer       -53.05770     21.41483   -2.478     0.015    *
```

(a) What are the interpretations of the coefficients for `ad_spend` and `summer`? [**2**]

(b) From this model, do sales appear to be significantly related to advertising spend? [**1**]

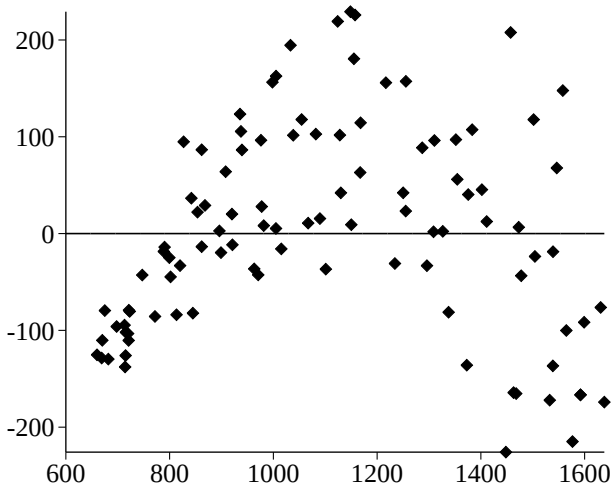
(c) If the following R command is run:

```
predict(model1, newdata=data.frame(ad_spend=200,summer=1))
```

what will the numerical output be? What is the interpretation of this quantity? [**4**]

(d) The following R commands are run to produce the plot below:

```
plot(x=model1$fitted.values, y=model1$residuals)
abline(h=0)
```



Which assumptions of the model appear not to hold here? [**4**]

Consider now the following R code:

```
n = nrow(sales_data)
pr = vector(length=n)
for(i in 1:n){
      modeli = lm(sales~ad_spend + summer, sales_data[-i,])
      pr[i] = predict(modeli, newdata=sales_data[i,])
}
r = sales_data$sales - pr
S = sum(r^2)
```

(e) What is the name for the statistical procedure that the code is carrying out? [**2**]

(f) Explain what each of the two lines of code inside the loop is doing. [**6**]

(g) When the code has run, what will `r` contain? What will `S` contain? [**3**]

(h) What aspect of the model is this method trying to assess? [**3**]

(i) How would the method compare two or more models fitted to the same dataset? [**2**]

Question 4 [17 marks]. Four classification methods are denoted by M_1, M_2, M_3 and M_4. These are tested using a validation dataset Y_{TRUE}. The table below lists the output from the validation exercise for the first two methods.

Y_{TRUE}	M_1	M_2
0	0	1
1	1	0
0	0	0
0	1	0
1	1	0
1	1	0
1	1	1
1	1	1
0	0	0
1	0	1

(a) Calculate the confusion matrix for each of methods M_1 and M_2. [**6**]

(b) Calculate the false positive rate (FPR) and true positive rate (TPR) for M_1 and M_2. [**4**]

(c) For methods M_3 and M_4, the results are summarized as follows.

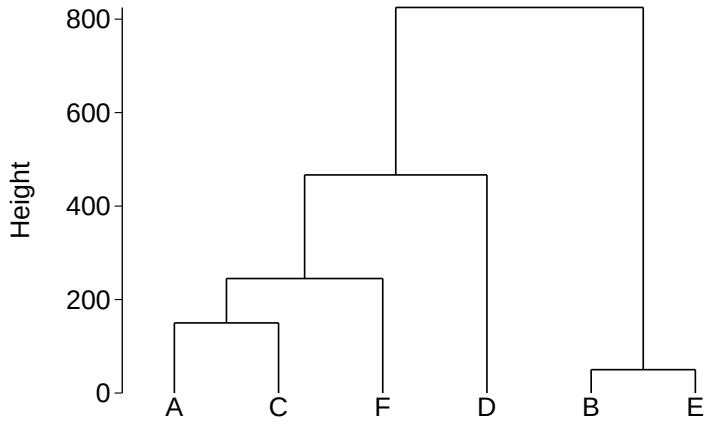
Method	FPR	TPR
M_3	0.5	0.833
M_4	0.5	0.333

Draw a ROC plot (by hand) using the data for all four methods, clearly labelling the quantities that are plotted on each axis. [**3**]

(d) Based on the ROC plot, which method is the best classifier? Comment on the performance of method M_4. [**4**]

 Continue to next page

Question 5 [6 marks]. The following is a dendrogram that shows the results of hierarchical clustering between six data-points, using complete linkage.

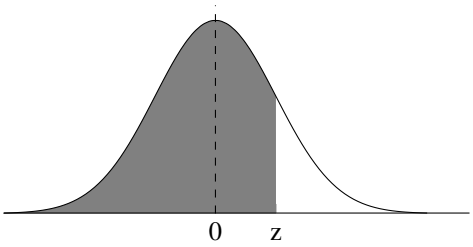


(a) If the dendrogram is cut at height 400, how many clusters would result, and which data-points would be in each cluster? [**3**]

(b) Explain why the first two points to be joined would be the same if complete linkage or single linkage was used. [**3**]

End of Paper – An appendix of 1 page follows.

Standard Normal Z-Table



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

End of Appendix.