Exam Pattern Recognition
Wednesday, February 2, 2022
17.00-20.00 hours

General Instructions

1. Write your name and student number on every sheet.
2. You are allowed to use a (graphical) calculator.
3. You are allowed to consult 1 A4 sheet of paper with notes (written or printed) on both sides.
4. Always show how you arrived at the result of your calculations.
   Otherwise you cannot get partial credit for incorrect final answers.
5. There are five questions for which you can earn 100 points.
6. Questions 1,2, and 3 are graded by Ad Feelders, and questions 4 and 5 by Albert Gatt. Since we would like to work in parallel, please write your answers to Ad's and Albert's questions on separate sheets.

Question 1: Linear Regression (20 points)

We have collected data on the number of COVID-19 related deaths per million inhabitants (as per 29th June 2020), and average male Body Mass Index for 58 countries. The complete data set used can be found on the last page of this exam. The Body Mass Index is defined as the weight (in kg) divided by the square of the height (in meters). We fit a linear regression model using the method of least squares[1], with target variable `Deaths per Million` (number of COVID-19 related deaths per million inhabitants), and predictor variable `BMI` (average male Body Mass Index). This produces the following result:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -581.94     283.71  -2.051   0.0449
BMI            28.08      10.88   2.581   0.0125
---
Multiple R-squared:  0.1063,    Adjusted R-squared:  0.09036
```

(a) (4 pts) Give an interpretation of the estimated coefficient of `BMI` in plain language.

(b) (4 pts) Indicate whether the coefficient of `BMI` is significant at $\alpha = 0.05$. Explain how you determined the answer.

(c) (4 pts) How much of the total variation in `Deaths per Million` is explained by the regression model? Explain how you determined the answer.

(d) (4 pts) How many deaths per million inhabitants does the fitted model predict for Ukraine? (row 35 in the data table). Does the answer make sense? Explain.

[1]Disclaimer: this analysis is not intended as a serious contribution to the subject.

We suspect that the life expectancy (in years) in a country may also influence the number of COVID-19 related deaths. Therefore we add life expectancy (`LIFE`) and the interaction term `BMI * LIFE` as predictor variables to the model. This yields the following result:

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             9154.280   3626.013   2.525  0.01455
BMI                     -408.511    145.951  -2.799  0.00709
LIFE                    -125.768     49.179  -2.557  0.01339
BMI * LIFE                 5.635      1.965   2.867  0.00589
---
Multiple R-squared:  0.3615,    Adjusted R-squared:  0.3261
```

(e) (4 pts) What is the effect of a one unit increase in BMI on the target variable in this model? Explain.

Question 2: Logistic Regression (20 points)

We analyse a data set with 28×28 pixel images of handwritten digits. Each pixel has a grayscale value between 0 (white) and 255 (black), inclusive. In this analysis, we restrict our attention to the digits one and five. We extract two features from each image, called `ink` and `asymmetry` respectively. The feature `ink` is defined as the sum of the pixel values divided by 1000, and `asymmetry` is defined as:

$$\texttt{asymmetry} = \frac{1}{1000} \sum_{j=1}^{28} \sum_{k=1}^{14} |x_{j,k} - x_{j,29-k}|,$$

where $x_{j,k}$ denotes the pixel value in row j and column k of the image. We analyse a data set containing 100 images of each digit using logistic regression, where digit one is coded as 0, and digit five is coded as 1. We fit the model using maximum likelihood estimation, which gives the following result:

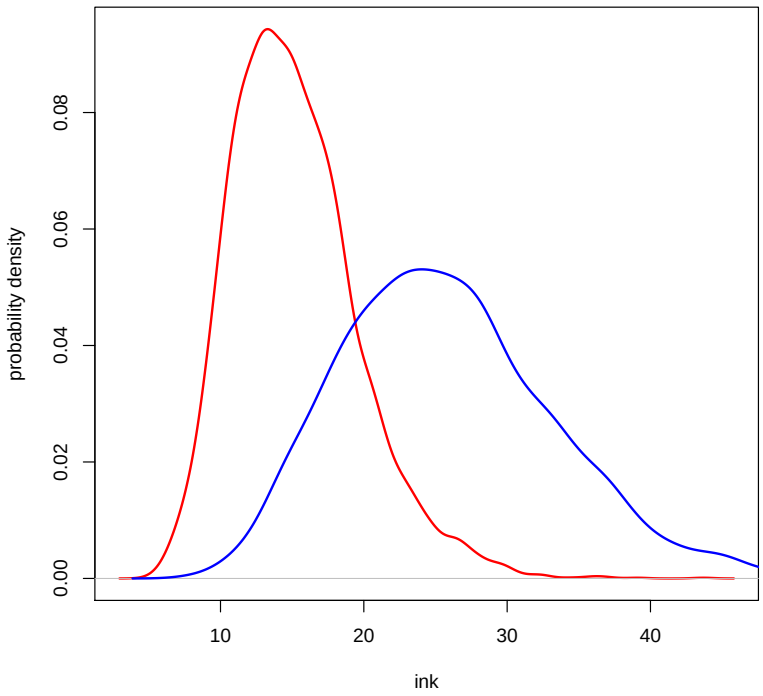
```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.73567    0.67732  -6.992 2.71e-12
ink          0.15896    0.03579   4.441 8.96e-06
asymmetry    0.11784    0.05458   2.159   0.0309
```

(a) (5 pts) Give an interpretation in plain language of the positive sign of the coefficient for `ink`. Does the positive sign make sense? Explain.

(b) (5 pts) Use the fitted model to estimate the probability that an all white image contains the digit one.

(c) (5 pts) Use the fitted model to construct a linear classification rule for the digits.

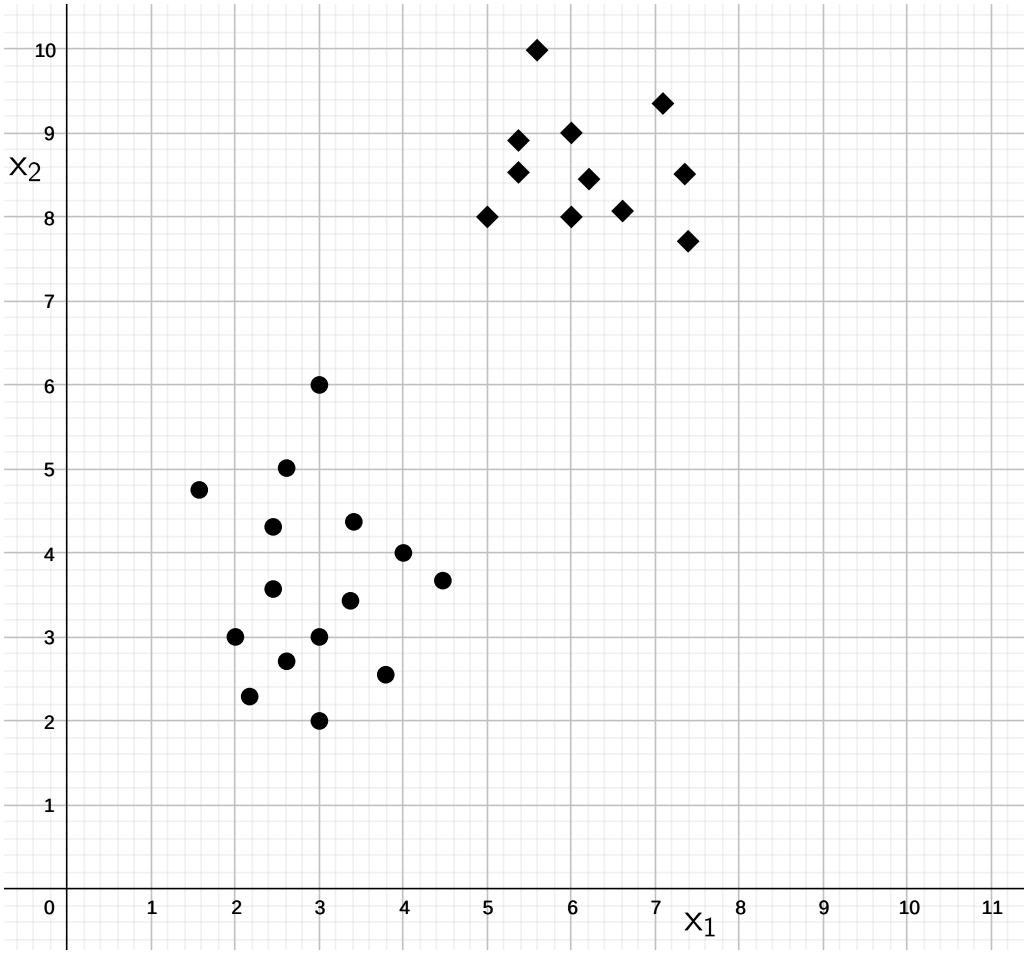
An alternative approach to classification uses density estimation. We used non-parametric density estimation to estimate the probability distribution of `ink` for digit one (the red curve) and for digit five (the blue curve).



(d) (5 pts) Give a classification rule based on the information in the given density plot. The rule need not be exact. Explain how you determined the classification rule from the plot.

Question 3: Support Vector Machines (20 points)

In the figure, circles represent examples of class -1, and diamonds examples of class $+1$.



We want to fit a support vector machine with linear kernel and perfect separation of the training data.

(a) (8 pts) Give the equation of the maximum margin linear decision boundary.
Hint: use the figure and elementary geometrical reasoning.
Scale the coefficients (bias and weights) appropriately.

(b) (4 pts) List the support vectors.

(c) (4 pts) Compute $y(\mathbf{x}_0)$ for $\mathbf{x}_0 = [5 \;\; 7]^\top$ and predict the class of $\mathbf{x}_0$.

You are given the following equations:

$$\mathbf{w} = \sum_{n=1}^{N} a_n t_n \mathbf{x}_n \tag{1}$$

$$\sum_{n=1}^{N} a_n t_n = 0. \tag{2}$$

(d) (4 pts) Determine the value of the Lagrange multiplier for all data points.

PLEASE WRITE YOUR ANSWERS TO QUESTIONS 4 AND 5
ON A SEPARATE SHEET!

Question 4: Feedforward and Convolutional Networks (20 points)

(a) (5 pts) Define the ReLU activation function, using either mathematical notation or plain English. Briefly state the advantages of ReLU, when compared to tanh or sigmoid.

(b) (10 pts) Explain what the convolution operation in a CNN does. Briefly explain the advantages of using convolutions for image data, compared to using a standard feedforward network.

(c) (5 pts) Suppose you have an image of size $32 \times 32 \times 3$. Your network has a convolutional layer consisting of 10 convolution filters, each of size 5×5, applied with stride 1 with no padding. What is the size of the output volume produced by this convolutional layer?

Question 5: Recurrent neural networks (20 points)

(a) (4 pts) Give an example of a classification task involving sequence data. Briefly explain why it is advantageous to use RNNs for this task, rather than feedforward networks.

(b) (8 pts) Explain why the vanishing gradient problem occurs with RNNs (you may do this intuitively, or use mathematical notation). How do GRU units help to address the vanishing gradient problem?

(c) (8 pts) Suppose you want to train a machine translation model. Your training data consists of sentences in one language (e.g. Dutch) paired with their translations in another language (e.g. Urdu). Describe a typical RNN-based architecture that can be used for these types of sequence-to-sequence problems, and explain what the main components of the architecture do. You may use a diagram to illustrate your answer.

	Country	Cases	Deaths	Cases per Million	Deaths per Million	Male BMI	Life Expectancy
1	USA	2637077	128437	7967	388	29.01	78.86
2	Brazil	1345254	57658	6329	271	26.47	75.88
3	Russia	634437	9073	4347	62	25.99	72.58
4	India	549197	16487	398	12	21.82	69.66
5	UK	311151	43550	4584	642	27.48	81.32
6	Spain	295850	28343	6328	606	27.16	83.56
7	Peru	279419	9317	8476	283	26.37	76.74
8	Chile	271982	5509	14229	288	27.99	80.18
9	Italy	240310	34738	3975	575	26.65	83.51
10	Iran	222669	10508	2651	125	25.39	76.68
11	Mexico	216852	26648	1682	207	27.64	75.05
12	Pakistan	206512	4167	935	19	23.42	67.27
13	Turkey	197239	5097	2339	60	27.27	77.69
14	Germany	194864	9029	2326	108	27.48	81.33
15	SaudiArabia	182493	1551	5243	45	28.14	75.13
16	France	162936	29778	2496	456	26.07	82.66
17	SouthAfrica	138134	2456	2329	41	25.11	64.13
18	Bangladesh	137787	1738	837	11	21.40	72.59
19	Canada	103250	8522	2736	226	27.45	82.43
20	Qatar	94413	110	33625	39	28.93	80.23
21	Colombia	91769	3106	1804	61	25.99	77.29
22	China	83512	4634	58	3	24.28	76.91
23	Egypt	65188	2789	637	27	27.93	71.99
24	Sweden	65137	5280	6450	523	26.86	82.80
25	Belarus	61475	383	6506	41	26.63	74.79
26	Belgium	61295	9732	5289	840	26.80	81.63
27	Argentina	59933	1232	1326	27	28.02	76.67
28	Ecuador	55255	4429	3132	251	26.79	77.01
29	Indonesia	54010	2754	197	10	22.59	71.72
30	Netherlands	50147	6105	2927	356	26.14	82.28
31	UAE	47797	313	4833	32	28.19	77.97
32	Iraq	45402	1756	1129	44	27.95	70.60
33	Kuwait	44942	348	10525	82	29.19	75.49
34	Singapore	43459	26	7429	4	24.40	83.62
35	Ukraine	42982	1129	983	26	17.09	72.06
36	Portugal	41646	1564	4084	153	26.27	82.05
37	Oman	38150	163	7474	32	26.52	77.86
38	Philippines	35455	1244	324	11	22.90	71.32
39	Poland	33907	1438	896	38	27.64	78.73
40	Panama	31686	604	7345	140	26.33	78.51
41	Switzerland	31617	1962	3653	227	26.83	83.78
42	Bolivia	31524	1014	2701	87	25.42	71.51
43	DominicanRepublic	31373	726	2892	67	25.69	75.00
44	Afghanistan	31238	733	796	19	22.68	64.83
45	Romania	26313	1612	1368	84	27.33	76.05
46	Bahrain	25705	83	15118	49	24.90	77.29
47	Ireland	25439	1735	5152	351	28.09	82.31
48	Armenia	24645	426	8317	144	25.74	75.09
49	Nigeria	24567	565	119	3	22.65	54.69
50	Israel	23755	318	2583	35	27.83	82.97
51	Kazakhstan	21327	178	1136	9	26.49	73.60
52	Japan	18390	971	145	8	23.68	84.63
53	Honduras	18082	479	1826	48	26.01	75.27
54	Austria	17654	702	1960	78	26.67	81.54
55	Guatemala	16930	727	945	41	25.83	74.30
56	Ghana	16742	112	539	4	22.59	64.07
57	Azerbaijan	16424	198	1620	20	26.25	73.00
58	Moldova	16250	530	4028	131	26.78	71.90

Table 1: Source data for Question 1, taken from: S. Bhaskar et al., Is COVID-19 Another Case of Obesity Paradox? - Results from An International Ecological Study on behalf of the REPROGRAM Consortium Obesity Study Group. Archives of Medical Science, 2021.