Machine Learning 1

Ma Data Science

BHT Berlin

Prof. Tim Downie

Example Exam – Summer Semester 2023

Surname, forenames:

Matriculation number:

Allowed material: The provided formula sheets, calculator, two sheets (four sides) of A4-Paper with *hand written* notes, calculator, pen, pencil ruler and blank paper.

Only use pencil for diagrams, all other writing should be done with non-erasable pen. Correction fluid (Tipp-Ex etc.) is not allowed.

Write your name on each page. Clearly label each exam question number/parts on your exam script. Please leave a few lines between your answers to each question.

The duration is 90 minutes. You need 50 marks to pass the course, including your attendance points.

Important: Marks are given for your working. Make sure you hand in all relevant working and calculations to maximise your marks.

Provisional grading scheme

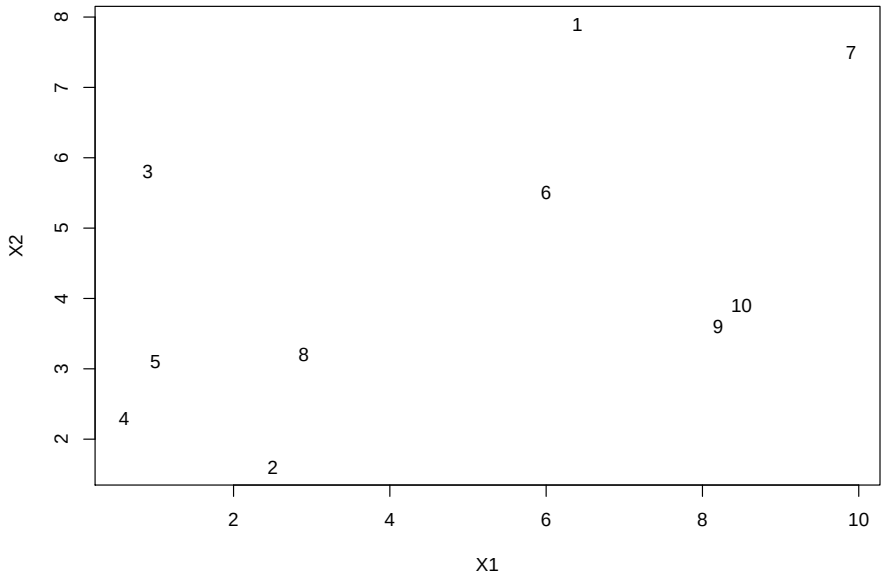
1.0	95	–		2.0	80	–	84	3.0	65	–	69	4.0	50	–	55
1.3	90	–	94	2.3	75	–	79	3.3	60	–	64				
1.7	85	–	89	2.7	70	–	74	3.7	55	–	59	5.0	0	–	49

Question 1 Hierarchical Clustering

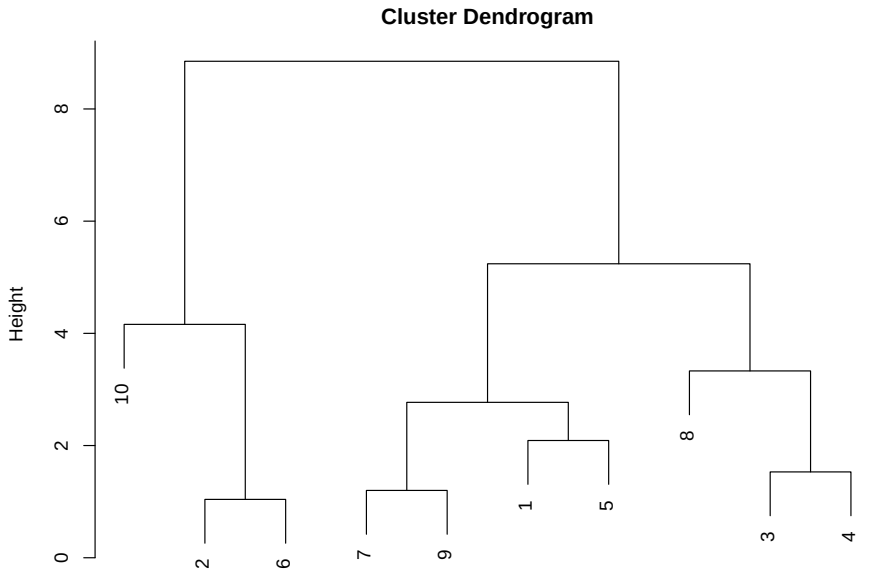
(a) Ten data points in two dimensions $X1$ and $X2$ are to be modelled using a hierarchical clustering model with complete linkage. The euclidean distance matrix is given below. To help you, the sorted distances, and a scatter plot with the labelled points are provided.

Find the first **five** joins in this hierarchical clustering model. For each join list the leaf elements involved and the complete linkage distance between the two joining clusters.

```
> dist(dX)
       1     2     3     4     5     6     7     8     9
2   7.41
3   5.89  4.49
4   8.06  2.02  3.51
5   7.22  2.12  2.70  0.89
6   2.43  5.24  5.11  6.28  5.55
7   3.52  9.46  9.16 10.66  9.93  4.38
8   5.86  1.65  3.28  2.47  1.90  3.86  8.22
9   4.66  6.04  7.62  7.71  7.22  2.91  4.25  5.32
10  4.52  6.43  7.83  8.06  7.54  2.97  3.86  5.64  0.42
> sort(dX)
 [1]  0.42  0.89  1.65  1.90  2.02  2.12  2.43  2.47  2.70  2.91
[11]  2.97  3.28  3.51  3.52  3.86  3.86  4.25  4.38  4.49  4.52
[21]  4.66  5.11  5.24  5.32  5.55  5.64  5.86  5.89  6.04  6.28
[31]  6.43  7.22  7.22  7.41  7.54  7.62  7.71  7.83  8.06  8.06
[41]  8.22  9.16  9.46  9.93 10.66
```



(b) For a different dataset, the following denrogram was obtained. Four clusters are to be defined from this model. List the elements in each cluster. Indicate graphically with a short explanation how you defined these clusters,



Question 2 Classifiers

A new home diagnosis kit is being developed to identify vitamin D deficiency. To use the kit, the user adds a paper strip to a urine sample. If the amount of "chemical X" excedes a set limit x_0 (in mg/l) then the paper strip turns purple. The company producing the strip can set the level x_0 at which the colour change occurs.

To calibrate the method, the company collect data using 60 test subjects, 25 of whom are known to have vitamin D deficiency. The remaining 35 test subjects are known to have acceptable levels of vitamin D. A logistic regression model was fitted to the results and the following output was obtained.

```
glm(formula = Def ~ chemx, family = "binomial")

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -5.750      1.454  -3.955 7.67e-05 ***
chemx          4.419      1.109   3.986 6.73e-05 ***
```

(a) What amount of chemical X gives the probability that *the test paper turning purple is equal to a half*? In mathematical notation: find x_0 such that $P(\text{Def} = 1|x_0) = 0.5$

(b) What value of x_1 corresponds to the probability that *the test paper turning purple is equal to at least 0.4, when the amount of chemical X is greater than x_1*, ie $P(\text{Def} = 1|x_1) = 0.4$.

(c) From this experiment 6 of the participants with Vitamin D deficiency and 7 of the participants without Vitamin D deficiency were predicted to have the wrong result. Display these results in a classification matrix and calculate the sensitivity and specificity of this classifier.

(d) The company director decides he wants to increase the sensitivity to 90%. Give one disadvantage in terms of the effectiveness of the test if the sensitivity is increased.

(e) A statistician reports that the area under the curve for these study participants and the above logistic regression model is 0.78. Explain what this means and how might you criticise his reporting this AUC=0.78?

Question 3 Ridge Regression

(a) A colleague is struggling with the model selection process for a multiple regression model (linear model) with many variables. Explain to the colleague what a ridge regression model is and how this supervised learning method might help. Include in your answer what the advantages and disadvantages of ridge regression compared to the multiple regression model are.

(b) Explain how cross-validation is applicable to the ridge regression model (you do not need to explain the cross validation method itself).

(c) How does the lasso model differ from the ridge regression model and what effect does this have on the resulting final model.

Question 4 Cross validation

i	x	y	fitted	CV prediction
1	2	12	11.82	11.44
2	5	13	13.52	13.75
3	9	16	15.79	15.73
4	10	17	16.36	16.10
5	12	17	17.50	17.99

A linear regression of the form $y = \beta_0 + \beta_1 x$ was fitted to the x and y data in the above table. The estimated coefficients are $\widehat{\beta}_0 = 10.69$ and $\widehat{\beta}_1 = 0.57$. The corresponding predicted values for this model are given in the table in the column called *fitted*. The *CV prediction* column gives the leave-one-out cross-validation (LOOCV) predicted value for each observation.

(a) Explain how the LOOCV predicted values are obtained.

(b) Calculate an estimate of the MSE for this model using the cross validation method.

(c) Why is this method often used to estimate the model MSE?

(d) How does the K-fold cross-validation method differ from the LOOCV method? Give one advantage and one disadvantage of K-fold CV in comparison to LOOCV.