Retake Pattern Recognition
Wednesday, April 20, 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: Mixed Questions (20 points)

(a) (5 pts) A cholesterol lowering drug is tested through a clinical trial. We expect cholesterol reduction to be proportional to drug dosage, and that women respond differently to the drug than men.

Specify a regression equation that fits this description, where cholesterol reduction is the dependent variable, and drug dosage and gender (and any variable that may be derived from them) are potential predictor variables. Motivate your answer.

(b) (5 pts) Describe the two steps that are alternated by the K means clustering algorithm. When does the algorithm terminate?

(c) (5 pts) What is the curse of dimensionality? Give an example of this phenomenon. How can one fight this curse?

(d) (5 pts) In binary classification, if the training data is linearly separable, then there are many different lines that give perfect separation. Describe in general terms which particular line is preferred by Support Vector Machines, and explain the intuition behind this preference.

Question 2: Linear Regression (20 points)

Distinguished Utrecht University researcher A. Feelders has collected the grades for the written exams (first attempt) of data mining and pattern recognition for students who took both courses in academic year 2021. Only students who participated in the first attempt on both occasions were included in the sample; the sample contained $N = 50$ students. Since the data mining exam comes first in time, we aim to predict the grade for pattern recognition from the grade for data mining. The average grade for pattern recognition was 7.73. Fitting a linear regression model with the method of least squares produces the following output:

```
Coefficients:
            Estimate   Std. Error t value Pr(>|t|)
(Intercept)     3.70      0.48743   7.586 9.37e-10
dm              0.66      0.07588   8.677 2.13e-11
---
R-squared:  0.61
```

Here `dm` is the grade for the data mining exam.

(a) (4 pts) Give the formula for the squared error loss function for this example. Use `dm` to denote the grade for data mining, and `pr` to denote the grade for pattern recognition.

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

(c) (4 pts) How much of the total variation in the grades for pattern recognition is explained by the regression model? Explain how you determined the answer.

(d) (4 pts) What grade for pattern recognition would the model predict for a student who scored 5 on the data mining exam?

(e) (4 pts) What was the average grade for data mining?

Question 3: Logistic Regression (20 points)

We consider data that were collected in a marketing campaign for a new financial product of a commercial investment firm. The campaign consisted of a direct mailing to customers of the firm. The firm wants to identify characteristics that might explain which customers are interested in the new product (i.e. respond to the mailing), and which customers are not. The target variable $t_n = 1$ if the nth customer responded to the mailing, and $t_n = 0$ otherwise. The explanatory variables are gender (male=1, female=0), activity (1 for customers that are already active investors, 0 otherwise), age (in years) and the square of age (divided by 100). The data set contains $N = 925$ customers of whom 470 responded, and the remaining 455 did not respond. The results of fitting a logistic regression model using maximum likelihood estimation to this data set are summarized below:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.49       0.88999  -2.796  0.00517
gender        0.95       0.15818   6.029 1.65e-09
activity      0.91       0.18478   4.945 7.61e-07
age           0.07       0.03561   1.964  0.04948
age^2/100    -0.07       0.03410  -2.015  0.04394
```

(a) (5 pts) Give an interpretation in plain language of the sign of the coefficients of `gender` and `activity`.

(b) (5 pts) According to the fitted model, what is the probability that a 30 year old male who is not yet an active investor will be interested in the new financial product?

(c) (5 pts) What is the effect of including the square of age (divided by 100) as an explanatory variable, in addition to just age? For the fitted model, describe in qualitative terms what is the overall relation between the age of the customer and the probability the customer responds to the mailing.

(d) (5 pts) Which coefficients would be judged "significantly different from zero" at significance level $\alpha = 0.05$? Explain how you determined the answer.

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 sigmoid function and explain its limitations as an activation function for neural networks.

(b) (5 pts) Regularization involves modifying a learning algorithm to reduce its generalization error, but not its training error. Give two examples of regularization techniques and briefly explain each one.

(c) (10 pts) Suppose you want to train a network on the task of image classification (i.e. labelling the objects in an image). You decide to do this with a convolutional neural network. What are the advantages of using convolutions, compared to using feedforward networks, for this task?

Question 5: Recurrent neural networks (20 points)

(a) (8 pts) Recurrent neural networks (RNNs) differ from feedforward networks, in that they incorporate a recurrence formula to compute their hidden state. Explain what this recurrence formula does, and how it differs from the way hidden states are computed in feedforward nets. You may use mathematical notation or plain English, as long as your answer is precise.

(b) (8 pts) GRUs were designed to overcome a problem with "vanilla" RNNs, namely, that RNNs tend to "forget" elements in a sequence, as sequences grow longer. Explain the main reason why RNNs exhibit this forgetful behaviour. Give an informal explanation of the properties of GRUs which help to overcome this.

(c) (4 pts) RNNs lend themselves to a variety of architectures. Explain the difference between a one-to-many architecture, and a many-to-one architecture. Feel free to use diagrams and examples to clarify your answer.