Exam Pattern Recognition
Wednesday, April 19, 2023
13.30-16.00 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: Regression (15 points)

We analyse data on $N = 100$ employees in a particular occupation. Interest centers on investigating the factors that explain salary differences with a view to addressing the issue of gender discrimination in this occupation. The data set contains the following variables: Salary (measured in thousands of dollars), Education (measured in years of schooling), Experience (measured in years of employment), and Gender (1 for male, 0 for female). Fitting a linear regression model with the method of least squares yields the following results:

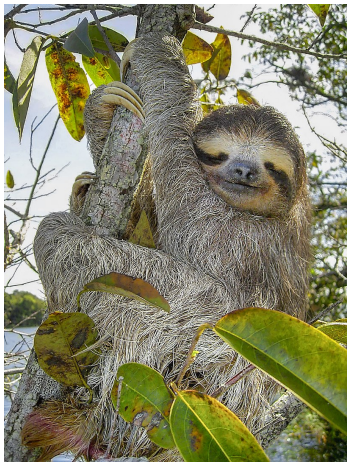
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 20.12763    0.14446 139.331   <2e-16
Education    0.50460    0.02840  17.765   <2e-16
Experience   0.76641    0.05596  13.695   <2e-16
Gender      -0.26689    0.28511  -0.936    0.352
---
Multiple R-squared:  0.8622
```

(a) (5 pts) Give an interpretation of the estimated coefficient of Experience. Does the sign of the coefficient make sense? Explain.

(b) (4 pts) What percentage of the total variation in Salary is explained by the model? Explain how you determined the answer.

(c) (6 pts) If you had to base your conclusions on this limited analysis, would you say there is significant evidence of gender discrimination in this occupation? Explain your answer.

Question 2: Logistic Regression (20 points)



The brown-throated sloth (*Bradypus variegatus*) is a species of three-toed sloth found in Central and South America. In order to characterize the "ecological niche" of this mammal, we have collected data on temperature and rainfall for locations where it has been found (Present = 1), and locations where it has not been found (Present = 0). The total number of locations in the training data is $N = 8950$. We specify a logistic regression model with predictor variables annual mean temperature in degrees Celcius (temperature), and annual rainfall in mm. (rainfall). Fitting the model with the

method of maximum likelihood yields coefficient estimates as given in the table below:

Coefficient	Estimate
(Intercept)	-2.724
temperature	-0.035
rainfall	0.0016

(a) (5 pts) Use the fitted model to estimate the probability that *Bradypus variegatus* can be found on a location with mean annual temperature of 10 degrees Celsius, and annual rainfall of 1500 mm.

(b) (5 pts) According to the fitted model, what is the effect of a rise in temperature of one degree Celcius (keeping rainfall constant) on the odds of *Bradypus variegatus* being present?

(c) (5 pts) Use the fitted model to construct a linear prediction rule for the presence of *Bradypus variegatus*, assuming we always predict the class with the highest estimated probability.

In a follow-up analysis, we add the interaction term temperature $\times$ rainfall as an additional predictor. This yields the following result:

Coefficient	Estimate
(Intercept)	-6.14
temperature	0.10
rainfall	0.004
temperature $\times$ rainfall	-0.0001

(d) (5 pts) According to the model with interaction term, for which values of rainfall does a rise in temperature increase the probability that *Bradypus variegatus* is present?

Question 3: Support Vector Machines (15 points)

We receive the following output from the optimization software for fitting a support vector machine with linear kernel and perfect separation of the training data:

n	$x_{n,1}$	$x_{n,2}$	t_n	a_n
1	3	5	-1	0
2	4	2	-1	0
3	6	6	-1	$\frac{2}{5}$
4	6	10	$+1$	0
5	7	8	$+1$	$\frac{2}{5}$
6	9	9	$+1$	0

Here $x_{n,1}$ denotes the value of x_1 for the n-th observation, t_n denotes the class label of the n-th observation, etc.

You are given the following formulas:

$$b = t_s - \sum_{n=1}^{N} a_n t_n \mathbf{x}_s^\top \mathbf{x}_n \qquad \text{(for any support vector } \mathbf{x}_s \text{ with label } t_s\text{)}$$

$$y(\mathbf{x}) = b + \sum_{n=1}^{N} a_n t_n \mathbf{x}^\top \mathbf{x}_n$$

Answer the following questions:

(a) (6 pts) Compute the value of the SVM bias term.

(b) (6 pts) Give the equation of the maximum margin linear decision boundary.

(c) (3 pts) Which class does the SVM predict for the data point $x_1 = 8, x_2 = 6$? Show your calculation.

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

Question 4: Feedforward and Convolutional Networks (25 points)

(a) (8 pts) In classification tasks with more than two classes, the loss function of choice is often the Cross-Entropy loss. Define this loss function, and explain why it is more appropriate for classification than a squared error loss.

(b) (8 pts) Regularisation is defined as any modification we make to a learning algorithm to prevent it from overfitting. Choose two regularisation techniques that are commonly used with neural networks. Explain how each of the two techniques works.

(c) (4 pts) In Convolutional Neural Networks (CNNs), convolution operations are typically followed by pooling. Suppose the following is the matrix that results from applying a 2×2 convolution filter to an input, with stride 2:

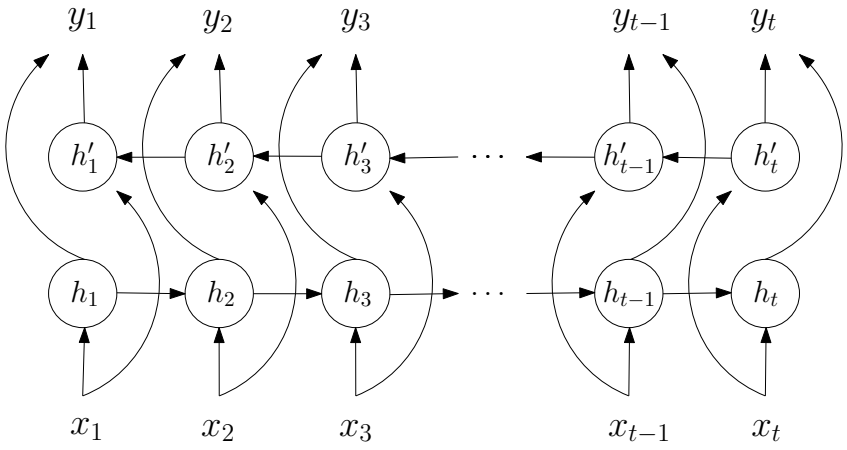
1	5	2	8
0	3	8	7
1	0	3	6
3	2	2	4

Write down the matrix that results from applying max pooling to the above.

(d) (5 pts) Some convolutional network architectures, including the classic ResNet by He et al. (2016), incorporate "residual" or "skip" connections. What is the purpose of such connections?

Question 5: Recurrent neural networks and attention (25 points)

(a) (6 pts) The following is a schematic diagram showing a bidirectional recurrent neural network (RNN). Note that the diagram does not display the weight matrices.



Explain how, at any timestep i, the output value y_i is computed as a function of the two hidden states. You may use plain English or give a formal definition.

(b) (8 pts) How does gating overcome the vanishing gradient problem in RNNs? Explain this with reference to either the Long Short-Term Memory (LSTM) architecture, or the Gated Recurrent Unit.

(c) (7 pts) In a standard encoder-decoder architecture, the encoder provides the "context" for the decoder, where the "context" is the encoder representation of the entire input. Explain why this creates a bottleneck, and how the use of an attention mechanism can overcome this bottleneck.

(d) (4 pts) Transformer models are able to capture essential information about sequences (for example, sentences in natural language). The core mechanism in such models is self-attention. What is the intuition behind self-attention, and why does it work so well to capture meaningful relationships between elements of a sequence?