Machine Learning 2

Master in Data Science

Beuth Hochschule für Technik, Berlin

Prof. Tim Downie

Example Exam – January 2020

Surname, forenames:

Matriculation nr.:

Third attempt? ☐ Yes ☐ No

Allowed material: Provided formulae, 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.

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

Do not write in this section:

Question	1	2	3	Project	Total*	Grade
Marks						
Maximum	x	x	x	30	100	

NB: marks per question are not allocated for the Example Exam
* if necessary rounded up

Provisional grading scheme

1.0	90	–		2.0	75	–	79	3.0	60	–	64	4.0	45	–	50
1.3	85	–	89	2.3	70	–	74	3.3	55	–	59				
1.7	80	–	84	2.7	65	–	69	3.7	50	–	54	5.0	0	–	44

Question 1 Local (loess) Regression **x Marks**

An outcome variable y_i was measured every second $x_i = i$ for $i = 1, \ldots, 100$ and is plotted against time in a scatter plot. It seems that there was noticeable measurement error in the y_i values.

(a) Give a detailed explanation of how loess (locally estimated least squares) regression can be used to obtain a predicted value for y after $x_0 = 60$ seconds ($f(60)$). You can assume that the loess degree is 1, the kernel function used is the Tricube function and the span is equal to $\frac{1}{2}$. You should define and explain these terms and explain the effect of the span on the resulting function.

(b) Is the predictor function at $x = 60$ equal to the predictor function at $x = 60.1$? Give a reason for your answer.

(c) How can one adapt the method to represent the predictor function for $1 \leqslant x \leqslant 100$ on the scatter plot?

(d) Why is loess regression called a smoothing method?

Question 2 Linear Support Vector Machines **x Marks**

Assume that the data consist of a binary outcome variable z and two predictor variables x and y.
Explain the following terms in the context of a linear support vector machine model.

(a) What is a vector in this context?

(b) What are the possible values of z in the SVM model?

(c) The data are separable.

(d) The boundary has the form $\beta_0 + \beta_1 x + \beta_2 y$, and the accompanying classification rule.

(e) The maximal marginal hyperplane, including the definition of *the margin* and a mathematical explanation of how the maximal marginal hyperplane is found.

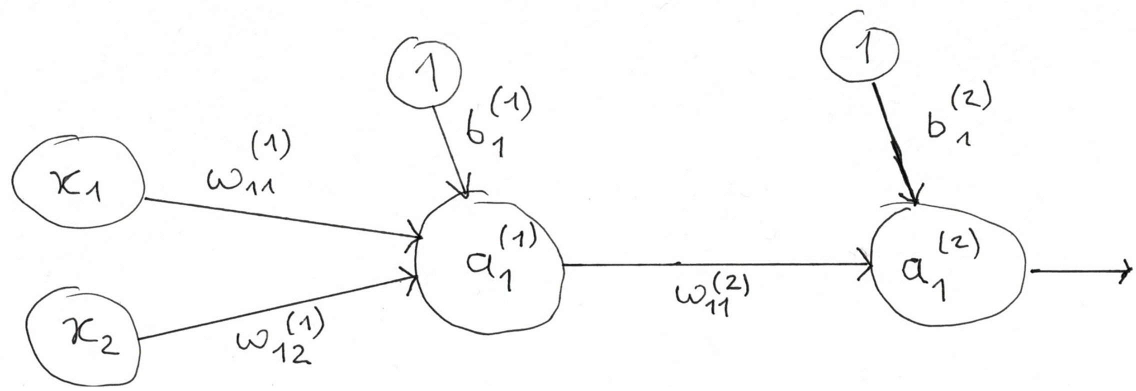
(f) The difference between a maximal marginal classifier and a support vector classifier.

(g) If the smoothing parameter C has the value 5, what can we infer from the resulting classifier?

(h) Express the classifier for a linear in terms of a kernel function, defining any terms you introduce.

Question 3 Back propagation **x Marks**

A simple regression Neural Network has the following structure with given weights and biasses. The input variables are x_1 and x_2 and the output variable is y. The activation function at the hidden layer is the sigmoid function, with $\sigma(v) = (1 + e^{-v})^{-1}$ and derivative $\sigma'(v) = e^{-v}(1 + e^{-v})^{-2}$. The activation function at the output node is the identity function and the squared error loss R is to be minimised.



(a) Derive following formula for the partial derivative of R with respect to w

$$\frac{\partial R}{\partial w_{11}^{(1)}} = -2\left(y - a_1^{(2)}\right) w_{11}^{(2)} e^{-z_1^{(1)}} \left(1 + e^{-z_1^{(1)}}\right)^{-2} \cdot x_1,$$

with $z^{(1)} = w_{11}^{(1)} x_1 + w_{21}^{(2)} x_2 + b_1^{(1)}$, and $a_1^{(1)} = \sigma(z^{(1)})$.

(b) Explain the role of $\dfrac{\partial R}{\partial w_{11}^{(1)}}$ in the vector "Grad R", $\nabla R(\boldsymbol{\theta})$.

End of exam paper