Examination
Software Engineering for AI Systems
DIT822

Software Engineering and Management
Chalmers | University of Gothenburg

Tuesday, October 29, 2024

Time	14:00-18:00
Location	Lindholmen
Responsible teacher	Daniel Strüber (mobile: 0760475434)
Total number of pages	**4** (including this page)
Teacher visits exam hall:	At circa 14:30 and at circa 15:30
Exam (4.5 HEC)	Max score: 20 pts
Grade limits (4.5 HEC)	3: at least 10 pts
	4: at least 14 pts
	5: at least 17 pts

ALLOWED AID:

- English dictionary
- **NOT ALLOWED:** Anything else not explicitly mentioned above (including additional books, other notes, previous exams, or any form of electronic device: dictionaries, agendas, computers, mobile phones, etc.)

PLEASE OBSERVE THE FOLLOWING:

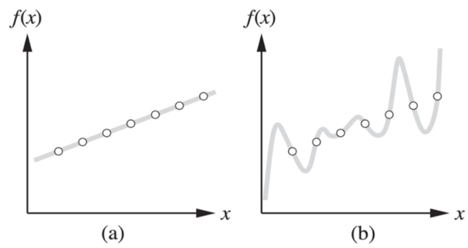
- This exam is composed by four exam tasks, divided into further sub-tasks, roughly corresponding to the four main topic areas of the lecture.
- Start each task on a new paper.
- Sort your answers in order (by task and sub-task) before handing them in.
- Write your student code on each page and put the number of the question on **every** page.
- Points are denoted for each task and sub-task. The point distribution can give you an indicator of how much time to spend on each task and sub-task.
- A few sub-tasks involve small calculations. These are of a type that can be done manually on paper, without a calculator.

Task 1: Linear Regression, Gradient Descent, Normal Equation (5 pts.)

a) In Linear Regression, we use either *gradient descent* or the *normal equation* to fit a model to data. Discuss their "pros and cons":
   i.) Explain two advantages of normal equation over gradient descent.
   ii.) Explain two advantages of gradient descent over normal equation.
   (1 pt.)

b) At what points do convex functions reach their global optimum? (1 pt.)

c) The figure below illustrates two polynomials that are consistent with the same data. Which polynomial should be preferred according to the principle of Ockham's razor? Motivate your answer. (1 pt.)



d) What characterizes a high-variance model? How does such a model perform on training and test data, respectively? (1 pt.)

e) How can we reduce high variance in a model? (1 pt.)

Task 2: Classification and Clustering (5 pts.)

Sir Jamie Lannister is trying to teach machine learning to his sister Cersei. Help Cersei by answering the following questions:

a) Logistic regression is used for binary classification. The outputs in logistic regression are real numbers between 0 and 1, which are used to assign a class label $y \in \{0,1\}$ based on whether $p \geq 0.5$ or not. What does the number p represent? (1 pt.)

b) Why should the mean squared error not be used as a cost function in Logistic Regression? (1 pt.)

c) Sir Jamie has trained a one-vs.-all model that classifies data points into four categories: $\{0,1,2,3\}$. Given a new data point x as defined below, what is the category of x? Motivate your answer. (1 pt.)

$$h^{(0)}(x) = 0.23, h^{(1)}(x) = 0.4, h^{(2)}(x) = 0.77, h^{(3)}(x) = 0.09$$

d) In the K-means algorithm, explain what the parameter K is, and explain one method of finding the optimal value for K. (1 pt.)

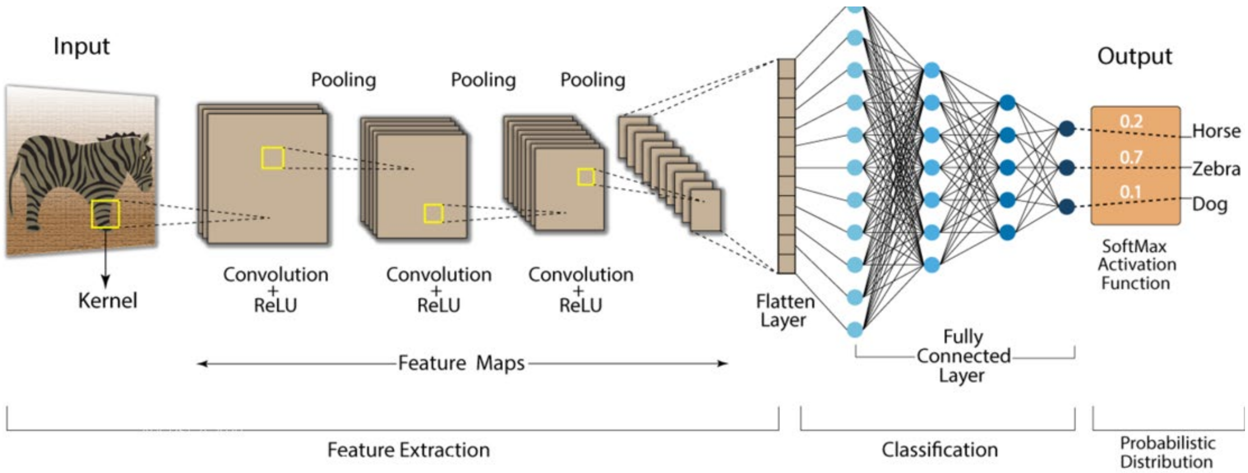
e) After initializing the centroids, the K-means algorithm recursively iterates two steps. What are these steps? **Hint**: Each step can be seen as a for-loop.
Write up your answer either as pseudo-code or as prose. (1 pt.)

Task 3: Neural networks and deep learning **(6 pts.)**

a) Compare traditional machine learning to deep learning by explaining at least three differences. These differences could be, for example, about their input, how they work, their performance, and the quality of their results. (1 pt.)

b) The following figure shows the Convolution Neural Network architecture.

i.) Explain briefly what happens in the first convolution layer in terms of inputs, outputs and operations. Mention the sizes and dimensions of the inputs and outputs and any other relative component. (1 pt.)
ii.) What are the inputs of the last convolution layer (mention the size and dimensions)? (1 pt.)
iii.) Why is the flatten function used before the fully connected layer? (1 pt.)



c) Assume you are working for a software company that wants to add neural networks to the set of AI services it provides.
The company wants to attract new customers and wants to know when it is suitable to choose from the following options: a standard *Neural Network* (NN), a *Convolutional Neural Network* (CNN), and a *Recurrent Neural Network* (RNN).
For each of these three architectures, briefly describe a scenario (2-3 sentences) where the architecture would be the most suitable choice based on the characteristics of the data and the task at hand. (2 pt.)

Task 4: ML Engineering **(4 pts. + 1 pt. bonus)**

a) Why is data quality important for machine learning? Discuss two specific data quality issues and how they might affect a model trained on the data. (1 pt.)

b) Below we see two formulas used in the context of feature engineering.

i) Which formula is used in *normalization* and which one in *standardization*?
ii) Explain the parameters in the formulas, and discuss at least one advantage of normalization and one of standardization. (1 pt.)

$$\hat{x}^{(j)} = \frac{x^{(j)} - \mu^{(j)}}{\sigma^{(j)}}$$

Formula A

$$\bar{x}^{(j)} = \frac{x^{(j)} - min^{(j)}}{max^{(j)} - min^{(j)}}$$

Formula B

c) Illustrate *bag-of-words* and *one-hot encoding* with one small example each. The examples should show the input for and the output of applying bag-of-words and one-hot encoding (drawn as tables / schemes / etc.) with specific data. (1 pt.)

d) The formula for the IOU score is:

$$IOU = \frac{Area\ of\ Intersection\ of\ two\ boxes}{Area\ of\ Union\ of\ two\ boxes}$$

Explain what the formula is used for, in the context of data management, and how it works: What are the boxes? How is the score interpreted? (1 pt.)

Bonus question:

e) How can the following visualization be used in the context of feature engineering? Give a concrete example, addressing specific data from the visualization. (1 pt.)