Examination
Software Engineering for AI Systems
DIT821

Software Engineering and Management
Chalmers | University of Gothenburg

Wednesday October 26, 2022

Time	14:00-18:00
Location	Lindholmen
Responsible teacher	Daniel Strüber (mobile: 0760475434)
Total number of pages	**6** (including this page)
Teacher visits exam hall:	At circa 14:30 and at circa 16:30
Exam (4.5 HEC)	Max score: 20 pts
Grade limits (4.5 HEC)	G: at least 10 pts
	VG: at least 15 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** paper;
- 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.
- Activities in the lectures in form of participation in breakout rooms will be considered, in the form of bonus points (1 pt. max).

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

a) Explain the *overfitting* and *underfitting* problems. (1 pt.)

b) Explain how regularization helps to address the overfitting problem. (1 pt.)

c) Compare *gradient descent* and the *normal equation* approach 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.)

d) Suppose you have a dataset with m=50 examples and n=200,000 features. You want to use multivariate linear regression to fit the parameters θ to our data. Would you prefer gradient descent or the normal equation? Explain why! (1 point)

e) Let $f(\theta_0, \theta_1)$ be a function that takes two numbers and outputs a number. Assume that f is an arbitrary smooth function, in particular, it may have local optima. Suppose we use gradient descent to try to minimize $f(\theta_0,\theta_1)$ as a function of θ_0 and θ_1. Which of the following statements are true or false? Explain with reasons. (1 pt. in total, 0.5 each)

   a. Even if the learning rate α is very large, every iteration of gradient descent will decrease the value of $f(\theta_0, \theta_1)$.

   b. If θ_0 and θ_1 are initialized so that $\theta_0=\theta_1$, then by symmetry (because we do simultaneous updates to the two parameters), after one iteration of gradient descent, we will still have $\theta_0=\theta_1$.

Task 2: Classification and Clustering (5 pts.)

a) Sir Jamine Lannister has given up sword fighting. Therefore, he has started to take up machine learning as a hobby. Sir Jamie has fit a logistic regression model of the form:

$$y = g(w_0 + w_1 x + w_2 x^2)$$

The values of the parameters he found are $w_0 = 6, w_1 = -5, w_2 = 1$. Now, Sir Jamie wants to find a decision boundary, but has forgotten how to do so. Help Sir Jamie and specify the equation of the decision boundary. (1 pt.)

b) Using logistic regression and a labelled dataset, Sir Jamie trained a one-vs-all model:

$$\mathcal{L} = \{(X_i, Y_i)\}_{i=1}^{m}$$

where the target Y can have three different labels 0, 1, 2.

Now, Sir Jamie wishes to find the labels of three new unlabeled points x_1, x_2, x_3. Help Sir Jamie predict the labels y_1, y_2, y_3 for x_1, x_2, x_3 respectively based on the following information: (1 pt.)

$$h^{(1)}(x_1) = 0.53, h^{(2)}(x_1) = 0.29, h^{(3)}(x_1) = 0.28$$
$$h^{(1)}(x_2) = 0.21, h^{(2)}(x_2) = 0.23, h^{(3)}(x_2) = 0.56$$
$$h^{(1)}(x_3) = 0.75, h^{(2)}(x_3) = 0.20, h^{(3)}(x_3) = 0.05$$

c) One important algorithm for clustering is K-means. At minimum, the K-means algorithm requires two inputs: the training data and a certain parameter. What is the parameter and what does it do? (1 pt.)

d) The first step of the K-means algorithm is to initiate cluster centroids. What are the second and third steps? You can give the answer in words or by writing it as pseudo-code. (Hint: Each step can be seen as a for-loop.) (1 pt.)

e) K-means minimizes the following cost function:

$$J\left(c_j^{(1)}, \ldots, c_j^{(m)}, \mu_1, \ldots, \mu_K\right) = \frac{1}{m}\sum_{i=1}^{m} \|x^{(i)} - \mu_{c^{(i)}}\|^2$$

Explain the cost function. (1pt.)

Task 3: Neural Networks **(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) In a neural network, what is the purpose of non-linearity? Name at least one type of function typically used for non-linearity. (1 pt.)

c) In a convolutional neural network, why is the kernel size generally smaller than the image size? (1 pt.)

d) In a convolutional neural network, what is the functionality of the pooling layer? (1 p.)

e) What is the vanishing gradient problem and how do recurrent neural networks solve it? (2 pt.)

Task 4: ML Engineering (4 pts.)

a) Name two sources of data quality problems, in particular:
- One source of problems that can appear during data collection, and
- One source of problems that only arises over time, *after* data collection. (1 pt.)

b) During feature engineering, it can become necessary to convert categorical data into numerical data. Explain why that is (with details, e.g., giving an example). Explain how one-hot encodings can be used to address this task. (1 pt.)

c) 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.)

d) Discuss the following statement about the "model requirements" and the "model evaluation" phases in the ML engineering workflow:

> Model requirements are connected to the evaluation metrics used for model evaluation, in the following way: for each model requirement, there should be at least one metric that directly measures how well the model fulfills it.

Is this statement always applicable? Give at least one example and (if there is one) one counterexample to this statement. (1 pt.)