# Examination
# Software Engineering for AI Systems
# DIT822

## Software Engineering and Management
## Chalmers | University of Gothenburg

Wednesday, January 3, 2024

| | |
|---|---|
| Time | 08:30-12:30 |
| Location | Lindholmen |
| Responsible teacher | Daniel Strüber (mobile: 0760475434) |
| Total number of pages | **4** (including this page) |
| Teacher visits exam hall: | At circa 9:00 and at circa 11:00 |
| | |
| **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) You have trained a linear regression model to predict house prices based on the number of bedrooms and square footage.
The model parameters are $\theta_0 = 100$, $\theta_1 = 50$ (*bedrooms*), and $\theta_2 = 10$ (*square footage*).
Calculate the predicted price for a house with 3 bedrooms and 2000 square feet. (1 pt.)

b) Describe how each of the following methods works, and how it helps to reduce overfitting.

- L1 (Lasso) and L2 (Ridge) regularization
- Splitting the dataset into training and test sets (1 pt.)

c) Describe three differences between gradient descent and normal equation. (1 pt.)

d) Suppose you have m=14 training examples with n=3 features (excluding the additional all-ones feature for the intercept term). The normal equation is $\theta=(X^TX)^{-1}X^Ty$. For the given values of m and n, what are the dimensions of θ, X, and y in this equation? (1 pt.)

e) Suppose you are tasked with optimizing a complex, arbitrary function $f(\theta_0, \theta_1)$ using the gradient descent algorithm, where $\theta_0$ and $\theta_1$ are parameters. This function may contain local optima. Explain and discuss the following aspects:

- The trade-off between the learning rate in gradient descent and convergence speed. Provide an example to illustrate the trade-off.
- The importance of initializing the parameters ($\theta_0$ and $\theta_1$) in gradient descent. How can improper initialization affect the optimization process? (1 pt.)

## Task 2: Classification and Clustering (5 pts.)

Since Sir Jamie gave up on sword fighting, he has successfully learnt Polynomial and Logistic Regression. It now time for him to learn the $K$-means algorithm.

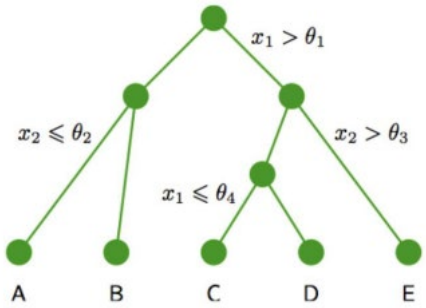
He has been given the following training data:
$x^{(1)} = 1, x^{(2)} = 302, x^{(3)} = 306, x^{(4)} = 3, x^{(5)} = 5, x^{(6)} = 304, x^{(7)} = 901. x^{(8)} = 903,$
$x^{(9)} = 909, x^{(10)} = 911$

By inspection, Jamie has chosen to use $K = 3$ clusters. The cluster centroids are randomly initialized as $\mu_1 = 900, \mu_2 = 1, \mu_3 = 300$.

a) For each data point $x^{(i)}$, find the index $c^{(i)}$ of the closest cluster centroid. (1 pt.)

b) Based on the closest cluster centroids you found, calculate the updated centroids for the 3 clusters after one iteration. (1 pt.)

c) We now move to decision trees. Given input data $x = (x_1, x_2, x_3, x_4)$, Sir Jamie can predict the label $y \in \{A, B, C, D, E\}$ using the trained decision tree below.




We have $\theta_1 = 100, \theta_2 = 25, \theta_3 = 33, \theta_4 = 20$.
Given a data point $x_1 = 101, x_2 = 1003, x_3 = 5, x_4 = 21$ what will be the label $y$? Explain the thought process that leads to your answer. (1 pt.)

d) Sir Jamie has fit the logistic regression model of the form:

$$y = g(w_0 + w_2 x^2 + w_3 x^3)$$

The values of the parameters he found are $w_0 = 2, w_2 = 3, w_3 = 5$. Specify the equation of the decision boundary. (1 pt.)

e) Sir Jamie has trained a one vs all model that predicts three labels {1,2,3}. Having already trained the model, he wants to predict labels for three new instances $x_1, x_2$. Calculate the labels of these, given that:

$$h^{(1)}(x_1) = 0.79, h^{(2)}(x_1) = 0.36, h^{(3)}(x_1) = 0.75$$
$$h^{(1)}(x_2) = 0.99, h^{(2)}(x_2) = 0.47, h^{(3)}(x_1) = 0.02$$

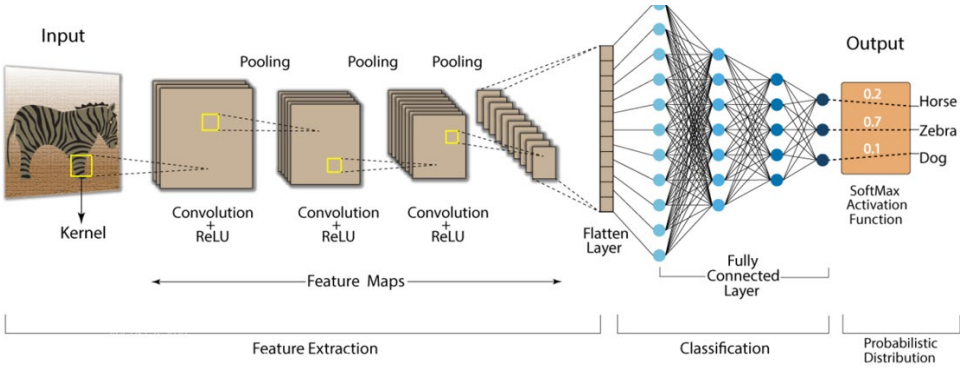
(1 pt.)

**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) In training feed-forward neural networks, what is the primary role of backpropagation algorithm? Explain briefly (in a few sentences) how the algorithm works. (1 pt.)

c) The following figure shows the Convolution Neural Network architecture. Explain briefly (in a few sentences) each part of the architecture. (2 pt.)




d) Explain the main difference between Feed-Forward Artificial Neural Networks and Recurrent Neural Networks. (1 pt.)

e) What is the vanishing gradient problem and how do Recurrent Neural Networks solve it? Answer both questions briefly (in a few sentences). (1 pt.)

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

a) For each of the following feature engineering operations, briefly describe the purpose (why is it used, in the context of ML engineering?) and how it works:

**1. normalization (or feature scaling)**
**2. standardization**
**3. binarization**
**4. one-hot encoding**
**5. bag-of-words**
**6. feature selection** (2 pt.)

b) Name a quality problem that can occur in data pipelines, and a potential solution that can help developers to avoid or address the problem. (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.)

Bonus question:

d) Present the overall workflow of developing machine learning software. Draw an overview picture and briefly (1-2 sentences) describe each activity. (1 pt.)