CHALMERS

EXAMINATION / TENTAMEN

Course code/kurskod	Course name/kursnamn
DIT 821	Software engineering for AI systems

Anonymous code Anonym kod		Examination date Tentamensdatum	Number of pages Antal blad	Grade Betyg
DIT821-0008-FLW		2023-01-03	5	VG

* I confirm that I've no mobile or other similar electronic equipment available during the examination.
Jag intygar att jag inte har mobiltelefon eller annan liknande elektronisk utrustning tillgänglig under eximinationen.

Solved task Behandlade uppgifter No/nr		Points per task Poäng på uppgiften	Observe: Areas with bold contour are to completed by the teacher. Anmärkning: Rutor inom bred kontur ifylles av lärare.
1	X	5	
2	X	5	
3	X	3,5	
4	X	3,5	
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
Bonus poäng	0	17	

1. a) When we have a model that is too high in complexity so it fits the data too well and can't generalise ~~3~~, we consider that model to be overfitting.✓ On the other hand, if the model is too simple and it can't capture the data well enough, then that model suffers from underfitting. We aim to have models that are complex enough to capture all relevant data and ~~generalise well~~ generalise well.✓ 1

1. b) Regularization addresses overfitting by introducing a regularization factor that minimizes the effect of large value features ~~3~~. The larger the regularization factor, the larger effect it has on the features. 1

1. c) Gradient descent optimizes the weights iteratively, by subtracting the gradient of the cost function from the current weight, until it converges. The rate of convergence for gradient descent depends on the learning rate. If the learning rate is too high, it might fail to converge. In most cases, gradient descent at least reaches the local minima. We might reach a different minima depending on the initial weights. Normal equation is a non-iterative approach that calculates the optimal weights by using feature and target vectors directly. With a reasonable amount of features, the normal equation approach has a good performance. If the number of features gets too large, it becomes too computationaly expensive to use this approach because of the matrix inverse that needs to be computed. 1

CHALMERS

Points for question (to be filled in by teacher)
Poäng på uppgiften (ifylles av läraren)

Löpande sid nr 2

Anonym kod DIT821-0008-FLW

Question no. Uppgift nr 1

1. d) In this case, I would prefer to use gradient descent. Since the number of features is quite large, it might become too expensive to calculate the matrix inverse needed for the normal equation. 1

1. e) a. False. The learning rate affects how fast we approach the local minimum. As we get closer, we need fewer steps to converge. So if the learning rate is too large, we might overshoot the minimum and start increasing the values of $f(\theta_0, \theta_1)$

✓

b. False. While the initial values of θ_0 and θ_1 might be the same, the ~~fat~~ feature values they're associated with might not be the same. This can lead to 2 different gradients, meaning a different θ_0 and θ_1 after one iteration of gradient descent. ✓

1

2.a) $g(h(x)) \geq 0,5$ $\qquad h(x) = 6 + (-5)x + x_2^2$

$h(x) \geq 0$ $\qquad = 6 - 5x + x^2$

$h(x) = 0$ - DB

$6 - 5x + x^2 = 0$

$x^2 - 5x = -6$ (1)

$5x - x^2 = 6$

2.b) $y_1 = 0$

$y_2 = 2$ ✓ (1)

$y_3 = 0$

2.c) The other paramheter is the number of clusters, it ~~the~~ determines the number of cluster centroids that will be created when initialising K-means. (1)

2.d) The second step is to go over every data point and assign it to the closest centroid. The third step is to go trough every cluster and move the centroid in each of those clusters to thee average position of the data points that belong to it (1)

2.e) The cost function calculates the mean of ~~[illegible]~~ the sum of all squared distances between all centroids and data points that belong to it. ~~(The lower the cost)~~ lower cost means that we have better formed clusters. (1)

3. a) Deep learning differs from traditional machine learning with the introduction of a neural network. Where a machine learning algorithm might be just a simple function taking in the features and weights, a neural network works on a principal of layers and perceptrons. At the very least, a neural network contains an input layer, an output layer and a hidden layer. Each layer has a set of perceptrons that performs an activation function and passes on the result to the following layer.

Deep learning != NN 0

3. b) Some functions used for non-linearity are the sigmoid function, ReLU. ✓ Non-linearity is required to achieve non-linear outputs from a neural network. what for? 0,5

3. c) The kernel size determines the quality of the produced feature maps. (~~The smaller the kernel~~) With a smaller kernel we can capture more detail. 1

3. d) The pooling layer reduces the size of feature maps while still keeping all of the relevant features of the map. 1

3. e) The vanishing gradient problem occurs during backpropagation trough a neural network. When performing gradient descent during backprop, the gradient tends to reach a very small value. Recurrent neural networks solve this by implementing a hidden state that is calculated at every step in the neural network. More details required

1

CHALMERS

Anonym kod
DIT821-0008-FLW

(to be filled in by teacher)
Poäng på uppgiften
(ifylles av lärare)

Löpande sid nr 5

Question no.
Uppgift nr 4

4. a) During data collection, a problem we might encounter is ~~[illegible]~~ invalid or incomplete data. For instance, some data might be missing same relevant features. ✓

After data collection, a problem we might encounter is the data becoming unusable due to, for instance, drastic changes in some regulations or the law. ✓

1

4. b) Machine learning algorithms are in ~~[illegible]~~ a lot of cases designed to work with numerical data. ~~Y~~ Categorical data has a textual form quite often, so it is necessary to convert it. One-hot encoding addresses this problem by converting the categories into a series of bits. Every category becomes a separate feature and its presence is denoted by a 0 or a 1.

1

4. c) The IOU score is used to test the inter-annotator agreement. The boxes represent the annotated features. The IOU score ranges from 0 to 1 (higher is better) and it tells us how well the annotations from 2 (or more) different annotators match. The close the score is to 1, the better the match. 1

4. d) This statement is applicable in most cases. For instance, if a model requirement is to keep predictions under x amount of ms, there needs to be a corresponding metric to evaluate if this requirement was fulfilled. 0,5

Counterexample?