CHALMERS
EXAMINATION / TENTAMEN

Course code/kurskod	Course name/kursnamn			
DIT821	Software engineering for AI systems			
Anonymous code Anonym kod		Examination date Tentamensdatum	Number of pages Antal blad	Grade Betyg
297		2022-10-26	12	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	√	4.5	
2	√	4.	
3	√	5.5	
4	√	3.	
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
Bonus poäng	0.5		

a) There isn't a lot of data/insufficent data data you are collecting are too noisy and needs data cleaning 0,5

b) categorical data usually consists of textual data input. its easier for machine learning algonthms to process ~~data that are numeric~~/binary within a certain range. Which one?

One-hot encoding transforms the categorical data to numeric data.

if we have one category: Gender with two possible data: female/male one-hot encoding would transform that one category into two with a binary input 1 or 0 where 1 would represent true and 0 false ✓ still 1 ~~0,5~~

eg.

encoded_fem	encoded_mas
0	1
1	0
1	0

istadd of

gender
male
female
female

c)

IOU is used to check the accuracy of two or more annotars annotation of one specific object. eg. a car.

The formula divides RoI/area of intersection of the two boxes with the total area of the two boxes. the formula will output a number between 0 and 1. If the number is 0,95 the two annotators annotations are 95% similar

1

3

Task 1

a) underfitting is when the function/line fitting the curve does not fit/predict the data well. overfitting is when the function created fits the data too well and does not work well on generalized data.
Let's take an example to illustrate overfitting and underfitting problems, calculating house prices. If the predicted function is underfitted then the predicted output would be inaccurate and hence have a large cost. If the function is overfitted you would get good results on your training data. However, if you compare to the validation or test datasets the result would be inaccurate. Fits the training data to well basically. (1)

b) Regularization helps with overfitting by limiting high exponents terms using this term: $\lambda \cdot \sum_{i=1}^{m} w^2$ (lambda). This results in the produced function high exponents terms to be close to 0. Illustration: (1)

$$w_0 + x_1 w_1 + x_2^2 w_2 + x_3^3 w_3 + x_4^4 w_4 + x_5^5 w_5 \ldots$$

In this example the weights associated to the higher terms to be regularized or have less impact resulting in a more smooth curve.

3

Task 1

c) Gradient descent using a small alpha valve can result in many iterations and long computational time. Using the normal equation no alpha valve is neccesary.

In gradient descent you must select an alpha value which if selected wrong can either result in long computation time or overshooting the minimum. comparatively, Normal eqval no alpha valve is necesary only transpose and inverse functions of matrixes are necessary.

A large positive with gradient descent is that it works well with large datasets. Whilst for Normal Equation it does not handle large datasets well, the inverse of a matrix is very taxing or computational high the larger the dataset

d) For this large dataset I would definitly use gradient descent over Normal Equation. The reason is due to the size of the dataset 50 examples an 20000 features if larger the dataset the higher computational time Normal Equation. Gradient descent would not matter the high amount of features and still converge to the global minimum considering regularizati would be an option to limit overfitting.

3

Task 1

e) Regarding a this would be false since it not true that with a large alpha it will always decrease θ_0, θ_1 since with a large alpha is prone to overshoot a potential local optima which would mean that the values θ_0 and θ_1 would increase in that instance. 0.5

Regarding b this statement would be False. The reason for this is due to gradient descents cost formula:

$$J(w) = \frac{1}{2m} \sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)}) \cdot w_j^2$$

If we assume θ_0 and θ_1 are the same value then the only time this statement would be true is when x_0, x_1 and y_0, y_1 also have the exact same values, otherwise θ_0 and θ_1 would be slightly different after one iteration. 0.5

Task 2

a) The equation of the decision boundary would be $y = g(x^2 - 5x + 6)$ by plugging in the values. This would result in a quadratic decision boundary prior to the g function. The g function then maps the values between 0, 1 or the amount of features.

Equation of decision boundary is:

$$g = g(x^2 - 5x + 6)$$

4p

b) $L = \{(x_i, y_i)\}_{i=1}^{m}$ $m = 3$

3 vectors: $y = 0$ $y = 1$ $y = 2$

$1 = [0.53, 0.21, 0.75]$

$2 = [0.29, 0.23, 0.2]$

$3 = [0.28, 0.56, 0.05]$

If we assume that for each x it checks how much of a label is present in the data compared to all. If we think of each index in the above vector as a indication of how much of label y was present in x. For simplisity $h^{(1)} = 0$ $h^{(2)} = 1$ $h^{(3)} = 2$.

Answer

Prediction:
- $(x_1, y = 0)$
- $(x_2, y = 2)$

4p

← this was done by taking the highest number at each vector index and mapping

c) k-means algorithm requires the inputs training data and amount of clusters (k). This parameters means the amount of clusters you want to group you data in. Having k = 1 would mean all data would be grouped into one cluster, k = 2 would mean the data would be grouped into two clusters etc. (b)

d) The second step in k-means algorithm is to calculate the nearest datapoint to each cluster centroids and assign each nearest datapoint to their corresponding cluster.

The third step is to calculate the mean of all nearby data points to then move each cluster centroids to those points and the repeat step 2 and 3 until the cost function reaches a threshold. (b)

```
def k-means(X, k):
    def Initialize_cluster_centroids(X, k):
        for i in range(1, k):
            for j in range(1, m):
            closest_point[i]:euclidian_distance(x, clusterpoint)
            clust_point[i] = cluster[i]
            cluster_point = calculate_mean(x)
    repeat until convergance
```

CHALMERS 297 Anonym kod DIT821 297 (to be filled in by teacher) Poäng på uppgiften (ifylles av lärare) Löpande sid nr 6 Question no. Uppgift nr Task 2,3

Task 2

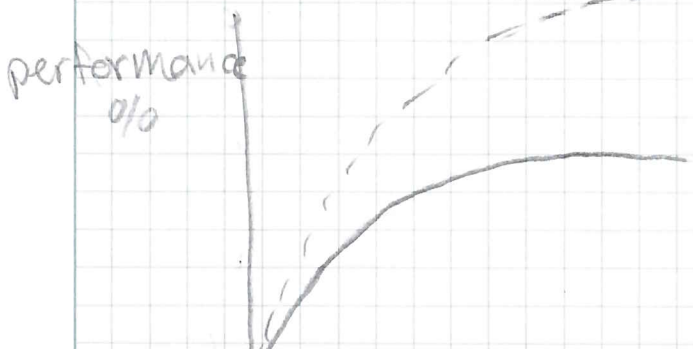
(0.3)

e) k-means cost function uses a similar cost function to Lasso regressions cost function. Where the sum of absolute value of each x point and standard deviation of that point to the center point of the centroid. The cost function for k-means gives an indication how well all of the datapoints were grouped. For example, in the instance of outliers μ_c would be large resulting in larger cost.

What about μ_c of clusters?

Task 3

a) One difference is how well they preform with large datasets. In traditional machine learning, having a very large dataset does not necessarily mean a very good permance. However, using deep learnings architecture the larger the datasets the better performance in most cases.



— = machine learning

- - - - = deep learning

Answer continues next →

Task 3

a) Second difference is the input provided to systems. Deep learning input does not need to be labelled. (supervised) The model extracts the useful features by itself in order to identify the correct digit for example. Whereas for ML models at least regarding being able to predict something, it must be labeled/supervised, which takes time to create. Of course unsupervised ML does not need labeles but then the goal is to mainly group feature not create predictions based of features.

A third difference is how ML and DL models work. ML models work by wanting to fit a line/curve to the data using a feature size amount of weights, where the weights are updated using gradient descent. In DL models, more non-linear outputs can be produced using activation functions applied on the data and the model works by extracting features and assigning neural network like weights in the system Input, Hidden, output layer nodes.

3

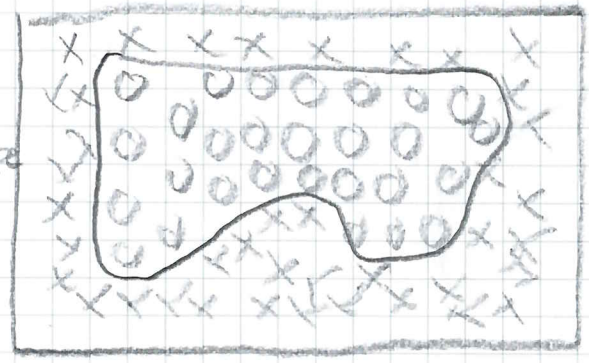
Task 3

(1)

b) The purpose of non-linearity functions such as sigmoid or ReLu is to be able to fit the data in a non linear matter since often data does not need to follow a linear trend. Example of non-linear clustering:

X = a type of data

O = another type of data



In this example having a linear function would not be able to fit the data well, hence the need for non-linearity.

(1)

c) The kernel size is often smaller than the image size since you want traverse the image and extract the import features in the image. Had the kernel size been larger then only the very high level features would be extracted no edge or corner detection would be preformed.

3

Task 3

d) The functionality of the pooling layer is to further decrease the size of the feature map by extracting the most important features using max, average or min pooling. In the example of max pooling with a pool size of (2, 2) then it would traverse the extracted features from the convolutional layer and at each 2x2 square the highest feature would be extracted.

e) vanilla recurrent neural networkes where each nodes stores it's memory

has a problem with the vanishing gradient problem. The vanishing gradient problem occurs when large sequences are inputed into system. Where the previous weights affect future weights resulting in the gradient descent becoming smaller and smaller and eventually not updating the weights at all.

$0.1 \cdot 0.1 \cdot 0.1 \cdot 0.1 = 10 \cdot e^{-4}$

Answer continues next

Task 3

e) The way RNNs handle the vanishing gradient problem is by utilizing the Long-short term memory architecture (LSTM)

The way LSTM's handle the vanishing gradient problem is by forgetting information that is not useful. Take the example of sentiment analysis, in the LSTM it does not need to consider all text only the words after the sentiment such as bad or good text. This way the whole sequence of words does not need to be considered handeling the vanishing Problem. The way LSTM's work is by having 3 gates: Forget gate, Input gate and output gate. The forget gate checks the most important info and forgets the rest. The Input gate checks if the memory needs to be updated and the output gate outputs the prediction output.

3

Task 4

a) One source of problem that can arise during data collection is the identification of certain types of features being over represented such as in face detection people with different skin not being present in the data.

A source of ptoblem after would be the identification of noisy data, null values or unlabeled data. For all of these data cleaning techniques would need to be used such as fillna, dropna and manually labeling missing labels. That's not "over time" (0,7)

b) The reason categorical data must be converted into numeric data is since the deep learning model works the best with numeric data. To convert categorical data to numeric one-hot encoding and hash encoding can be used. One-hot encoding works by turning the categorical data into columns of binary data.

Example:

	categorical
1	cat
2	dog
3	dog
4	cat

one hot encoding →

	cat	dog
1	1	0
2	0	1
3	0	1
4	1	0

(1)

3

Task 4

c) IOU means the intersection over union. The formula is used to calculate how conjoint the data is. The boxes are the areas of importance such as identifying a car in a security camera. The score is interpreted by higher IOU means the lower the boxes are joint since the area of overlapp is smaller. Low IOU means the boxes are more joint since the overlapping are will be larger and hence decrease IOU.

(1)

IOU formula: Intersection - overlapp

d) I believe this statement is always applicable since ML requirements are more data driven such "As a user I want a system that can predict cat and dog images with 99% accuracy". This way in the model evaluation must fufill the acceptance criteria (metrics) in order to be able considered as completed. Often in ML requirements the model to achieve this accuracy is often not stated. It is the developers job to determine what model would fit the data