Machine Learning 1DL034
2022-08-24 08:00 - 13:00

Question	Points	Score
1	20	
2	2	
3	4	
4	5	
5	9	
Total:	40	

Anonymous Exam Code : : ....

Exam Instructions

Please read the following instructions:

- In order to get a 3 you have to get 80% on question 1. You will get a 4 if you pass 50% on the remaining questions 2—3 and a 5 if you get more than 85% on the same remaining questions. Note that in order to at a 4 or a 5 you *must* get 80% on question 1.
- In the multiple-choice questions, each question has *only one* answer. You are to pick the option that best answers the question.
- It is possible that I have made a mistake in one of the multiple-choice questions, and there is no correct answer. If this is the case, then please add a note saying why you think the question is incorrect and you will get full marks assuming that I have made a mistake. It is my intention that there should be no mistakes or trick questions in the exam.
- When answering the multiple choice questions just circle *the* correct answer on the exam paper. For questions 2 and 3 write your answers on the supplied paper.
- I will not come to the exam to answer questions. If you are unsure about any of the questions then please state any assumptions that you have made when answering the question. If your assumptions are reasonable then you will still get full marks.

Multiple Choice Questions

1. General Questions (you need to get 80% on these questions to get a 3)

   (a) (1 point) Which one of the following data types is **not** categorical:

   A. The blood type of a person: A , B , AB or O.

   B. The type of Car somebody owns.

   C. The student Nation somebody belongs to.

   D. The weight of somebody.

   (b) (1 point) Which one of the following data types is best represented as categorical data:

   A. The age of a person.

   B. The number of times somebody has done the Champagnegalopp on the last of April in Uppsala.

   C. The blood type of a person.

   D. The blood alcohol content of somebody the morning after the last of April.

   (c) (1 point) Which of the following algorithms is suitable for supervised learning of categorically labelled data:

   A. Hierarchical clustering.

   B. Support Vector Machines

   C. K-means clustering.

   D. Linear regression.

   (d) (1 point) Mark the algorithm that is **not** suitable for classification:

   A. Linear Regression

   B. Naive Bayes

   C. Support Vector Machines

   D. Logistic regression.

   (e) (1 point) Which of the following tasks is best solved using a regression algorithm. problem:

   A. Classifying if a message is spam or not.

   B. Predicting the number of points a student will get on their exam based on their performance and attendance during the course.

   C. Deciding if a message is written in Swedish or English.

   D. Classifying an image as a car or a pedestrian.

   (f) (1 point) Which of the following statements is true for a classification algorithm.

   A. It is always best to optimise for the best false positive and true negative rates.

   B. It is best to maximise the F score.

   C. Optimising for the best false positive, false negative, true positive or true negative depends on your application.

   D. Improving precision is always better than improving recall.

(g) (1 point) Which of the following statements best describe hyper-parameters:

A. The number of hyper-trees that are used in random forest algorithms.

B. Parameters of your machine learning algorithm that should not be learned from your training set.

C. The number of GPUs required for deep learning.

D. The value of the regularisation constant used in logistic regression.

(h) (1 point) When training and evaluating a machine learning algorithm it is a good idea to split your data up into a training set and a test set. Why should you do this?

A. It is too inefficient to train a model on the whole data set.

B. To avoid overfitting.

C. The test set is used to see if the error or loss on the training set is correct.

D. To avoid being laughed at by fellow machine learning engineers.

(i) (1 point) Suppose that your model is overfitting. Which of the following is **NOT** a valid way to reduce overfitting:

A. Increase the amount of training data.

B. Decrease the model complexity.

C. Tuning your hyper-parameters to reduce the error/loss on your training data.

D. Reduce the noise in the training data.

(j) (1 point) Which of the following statements best describes logistic regression:

A. It is an algorithm that can divide data into two classes provided that the two classes are linearly separable.

B. It is a regression algorithm that first passes the data through a logistic function.

C. It is a type of deep learning for linear regression.

D. It is an application of the logistic kernel function for support vector machines.

(k) (1 point) If you are training logistic regression using gradient descent then it is *always* possible to converge to a solution where the cost/error function $J(\theta)$ equals 0.

A. True B. False

(l) (1 point) Assume that you are using gradient descent to train your learning algorithm, and you are given an error or loss function J used during training of a machine learning algorithm. If $\theta_0, \ldots, \theta_n$ are the weights or parameters of the algorithm and

$$\frac{\partial J(X, \theta)}{\partial \theta_i} = 0$$

for all i then which of the following statements is true:

A. There is not enough training data and you must collect more before you can train the algorithm further.

B. J is zero.

C. The values of $\theta_0, \theta_1, \ldots, \theta_n$ are the values of a local minimum of the function J.

D. The values of $\theta_0, \theta_1, \ldots, \theta_n$ are the values of the global minimum of the function J.

(m) (1 point) For logistic regression it is possible to use a regularisation parameter λ as follows:

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} -y \log(\sigma(h_\theta(x))) - (1-y) \log(1 - \sigma(h_\theta(x))) + \lambda \sum_{i=1}^{n} \theta_i^2$$

Why is regularisation used?

A. It normalises the input data so that each feature has a mean of zero.

B. It speeds up learning.

C. It avoids over-fitting on the training data by forcing gradient descent to learn small weights θ_i.

D. Increasing λ reduces the considered dimension of the training data.

(n) (1 point) Suppose that in the general population. the percentage of people having Covid is 4%. You are taking a test. It either reports `positive` indicating that you have Covid or `negative` indicating that you do not have Covid. The test is of course not perfect. Let C be the event that somebody has Covid and $\overline{C}$ be the event that somebody does not have Covid. After clinical trials, it has been established that $P(\texttt{positive}|C) = 0.99$ and $P(\texttt{negative}|\overline{C}) = 0.85$.

If you pick a random person in the general population and administer the test what is the probability that you get a positive result, that is the value of $P(\texttt{positive})$?

A. 0.99

B. $0.99 \cdot (1 - 0.85)$

C. $0.04 \cdot 0.99 + (1 - 0.04) \cdot 0.85$.

D. I do not have enough information to calculate the probability.

E. $0.04 \cdot 0.99 + (1 - 0.04) \cdot (1 - 0.85)$

(o) (1 point) Using the same probabilities as in the previous question.

What is the correct expression for $P(C|\texttt{positive})$ (the probability that you have Covid given that test returns `postive`)?

A. I do not have enough information to calculate the probability.

B. 0.99

C. $0.99 \cdot (1 - 0.85)$

D.

$$\frac{0.99 \cdot 0.04}{0.04 \cdot 0.99 + (1 - 0.04) \cdot 0.85}$$

E.

$$\frac{0.99 \cdot 0.04}{0.04 \cdot 0.99 + (1 - 0.04) \cdot (1 - 0.85)}$$

(p) (1 point) You have the following data concerning the occurrence of words in spam email:

Spam (Y/N)	'Home' Occurs	'BitCoin' Occurs
N	Y	Y
Y	Y	Y
Y	N	N
Y	Y	Y
Y	N	N

You are given an email that contains the word "BitCoin" which of the following is the correct value of $P(\text{Spam} \mid \text{BitCon})$

A. $\log_2(2/3) + \log_2(3/5)$ B. $\frac{(2/4)\cdot(4/5)}{(3/5)}$ C. $\frac{(2/5)\cdot(3/5)}{(3/5)}$ D. $\frac{(3/5)}{(2/5)}$

(q) (1 point) Which of the following options can be tried to get global minima in K-Means Algorithm?

1. Try to run the algorithm for different centroid initialisations.
2. Adjust the number of iterations
3. Find out the optimal number of clusters

A. 1 and 3 B. 3 C. 2 and 1 D. 1, 2 and 3.

(r) (1 point) You have a data set consisting of two classes `Spam` and `NonSpam`, the data set has entropy close to 0. Which one of the following statements is always true?

A. The probability of `Spam` is roughly equal to the probability of `NonSpam`.

B. The data set is almost all `Spam`.

C. There are no `NonSpam` messages in the data set.

D. The data set is either almost all `Spam` or almost all `NonSpam`.

E. The data set is almost all `NonSpam`.

(s) (1 point) You are using the ID3 algorithm to construct a classifier to work out if somebody can play golf or not. You are given the following data:

Humidity	Sunny	Windy	**Play**
L	N	N	True
L	N	Y	True
H	Y	N	True
L	Y	Y	True
H	Y	Y	False
L	Y	N	False
H	N	N	False

Consider the probability of playing golf and the probability of not playing golf. Which of the correct value of entropy of our data set

A. $-(\frac{3}{7}\log_2\frac{4}{7}+\frac{3}{7}\log_2\frac{4}{7})$ B. $-(\frac{1}{4}\log_2\frac{4}{7}+\frac{1}{3}\log_2\frac{3}{7})$ C. $-(\frac{4}{7}\log_2\frac{4}{7}+\frac{3}{7}\log_2\frac{3}{7})$

(t) (1 point) Again using the ID3 algorithm and the same data-set as in the previous question part. Which of the following is the correct expression for the information gain for splitting on the attribute "Windy"? Note that $H(P)$ is the entropy of the whole data set calculated in the previous question part.

A. $H(P)+(\frac{3}{7}H(P|\text{Windy}=\text{Yes})+\frac{4}{7}H(P|\text{Windy}=\text{No}))$

B. $H(P)-(\frac{3}{7}H(P|\text{Windy}=\text{Yes})+\frac{4}{7}H(P|\text{Windy}=\text{No}))$

C. $H(P)-(H(P|\text{Windy}=\text{Yes})+H(P|\text{Windy}=\text{No}))$

D. $H(P)+(\frac{1}{3}H(P|\text{Windy}=\text{Yes})+\frac{1}{4}H(P|\text{Windy}=\text{No}))$

Optional Questions for a higher grade

2. (2 points) What is one-hot encoding and why is it necessary?

3. (4 points) You want to build a classification algorithm that classifies images into different categories “Dog”, “Cat”, “Tesla”, and “Pedestrian”. You plan to use support vector machines. Your support vector machine implementation can only classify one class at a time. That is you can train the support vector machine to recognise if something belongs to a class or does not (a binary classifier). Describe two schemes for combining classifiers in order to classify into the four different classes above. For the two different schemes describe the advantages and disadvantages of both schemes.

4. (5 points) You are going to build a recommender system that predicts what film/movie a user is most likely to watch based on their past history. You are free to use supervised or unsupervised learning or any combination. Describe what sort of approach you would use. In particular, describe what sort of data you would collect and what attributes you would use.

5. Naive Bayes Classification and Classiying Spam.

   (a) (2 points) What is naive about a naive Bayes classifier? Explain how the relevant probabilities are calculated. In particular, explain what assumptions are made for the calculations to be correct.

   (b) (5 points) When building a Bayesian spam filter there are at least two approaches to modelling the occurence of words in messages. One approach only cares if a word occurs in a message or not, and the other approach counts how many times a word occurs. Describe these two approaches and describe how the relevant probabilities are calculated and estimated from the data set.

   (c) (2 points) How and why is Laplacian smoothing used in a naive Bayes calssifier?