Machine Learning 1DL034
2023-03-13 08:00-13:00

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

Anonymous Exam Code : : ....

Exam Instructions

In order to get a 3 you have to get 70% on question 1. Your will get a 4 if you pass 50% on the remaining questions, 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 70% on question 1.

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. Each multiple choice question only has *one* correct answer.

When answering the multiple choice questions just circle the correct answer on the exam paper. For the remaining questions write your answered on the supplied paper.

I will probably not be able to 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 70% on these questions to get a 3)

   (a) (1 point) Which of the following statements best describes supervised learning?

   A. An algorithm with a human in the loop. After each training run, the human evaluates how well the algorithm has done.

   B. A learning algorithm that requires training data where is item is labelled with either a value or a class. The goal is to train the algorithm to minimise the prediction error on the training set.

   C. An algorithm such as k-means clustering, where each cluster represents a class.

   D. An algorithm that works on unlabeled that must be supervised to ensure that the correct classification has been made.

   (b) (1 point) Which of the following problems is a regression problem?

   A. Deciding if a message is Spam nor non-Spam.

   B. Deciding if an image is a Dog,Cat or a Wombat.

   C. Predicting the probability that a message is a Spam message.

   D. Deciding if a Student will pass or fail an exam based on their performance on the assignments and project.

   (c) (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 care somebody drives

   C. The type of web browser that a client is using.

   D. The age of somebody.

   (d) (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 to inefficient to train a model on the whole data set.

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

   C. The test set is used to decide the values of the hyper-parameters of your algorithm.

   D. To avoid overfitting.

   (e) (1 point) Which of the following algorithms is suitable for classification using supervised learning.

   A. K-means clustering

   B. Linear Regression

   C. Principle Component Analysis

   D. Support Vector Machines

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

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

B. Improve recall is always better than improving recall.

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) You have a data-set labelled into two classes "True" and "False". Which of the following machine learning algorithms could you try without any modification:

A. Linear Regression

B. Hierarchical Clustering

C. Logistic Regression.

D. Principle Component Analysis

(h) (1 point) Your task is to build a supervised learning algorithm that classifies data into multiple classes. Which of the following algorithms could you use unmodified for such a task:

A. K-nearest neighbours.

B. Linear Regression

C. Logistic Regression

D. Support Vector Machines

(i) (1 point) You are building a supervised learning algorithm for classification and all your features are categorical. Which of the following algorithms could you use unmodified for such a task:

A. Linear Regression

B. Logistic Regression

C. Support Vector Machines

D. Decision Trees

(j) (1 point) Given the following data set:

x	y
1	2
2	3
4	7

Using gradient descent to fit a linear regression model: $h_{\theta_0,\theta_1}(x) = \theta_0 + \theta_1 x$. Which of the following expressions is the correct value of the loss (or error) function J.

A. $J(x, \theta_0, \theta_1) = \frac{1}{2\times 3}\left((2-(\theta_0+\theta_1))^2 + (3-(\theta_0+2\theta_1))^2 + (7-(\theta_0+4\theta_1))^2\right)$

B. $J(x, \theta_0, \theta_1) = \frac{1}{2\times 3}\left((2-(\theta_0+\theta_1))^2 + (3-(\theta_0+2\theta_1))^2 + (7-(\theta_0+3\theta_1))^2\right)$

C. $J(x, \theta_0, \theta_1) = \frac{1}{2\times 3}\left((2-(\theta_0+\theta_1)) + (3-(\theta_0+2\theta_1)) + (7-(\theta_0+4\theta_1))\right)$

D. $J(x, \theta_0, \theta_1) = \frac{1}{2\times 3}\left((2-(\theta_0+\theta)) + (2-(\theta_0+2\theta_1)) + (4-(\theta_0+3\theta_1))\right)$

(k) (1 point) Suppose you have a training set and a test set. You use gradient descent to train your machine learning model on your training set. After training your error or loss is extremely low (very close to zero), but when you test your learned model on the test set the error or loss is extremely high. Which of the following statements best characterises your situation:

A. You have the wrong model.

B. You have overfitting.

C. You need to adjust the training/test split so that your test set is larger.

D. You have the wrong random seed for your learning algorithm and you should retrain the model.

(l) (1 point) What does the regularisation parameter do in logistic regression?

A. Tries to avoid over fitting by forcing the model to learn small weights.

B. Allow a percentage (depending on the parameter) of miss classifications during training.

C. Improves the convergence of the training algorithm by having the step size depend on the average value in the data set.

(m) (1 point) Suppose that in the general population the percentage of people having Covid is 4%. You are taking a test. It either reports `postive` 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 $P(\texttt{postive}|C) = 0.99$ and $P(\texttt{positive}|\overline{C}) = 0.1$. What is the correct expression for $P(C|\texttt{postive})$ (the probability that you have Covid given that test returns `postive`).

A. 0.99

B.

$$\frac{0.99 \cdot 0.04}{0.04 \cdot 0.99 + (1 - 0.04) \cdot 0.1}$$

C.

$$\frac{0.99 \cdot 0.05}{0.1 \cdot (1 - 0.04)}$$

D. $0.99 \cdot (1 - 0.95)/0.04$

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

(n) (1 point) You are classifying a set of images as either `Cat`, `Dog` or `Wombat`. After analysing the data you estimate that $P(\texttt{Cat}) = 0.3$, $P(\texttt{Dog}) = 0.5)$ and $P(\texttt{Wombat}) = 0.2$ which of the following expressions is the correct value of the (binary) entropy of the data set.

A. $\frac{0.3 \cdot (1-0.3) + 0.2 \cdot (1-0.2) + 0.6 \cdot (1-0.6)}{0.3 \cdot 0.6 \cdot 0.1}$

B. $0.3 \log_2(-0.3) + 0.5 \log_2(-0.5) + 0.2 \log_2(-0.2)$

C. $0.3 \log_2(0.3) + 0.6 \log_2(0.6)$

D. $-0.3 \log_2(0.3) - 0.5 \log_2(0.5) - 0.2 \log_2(0.2)$

(o) (1 point) You have a data set consisting of two classes `Spam` and `NonSpam`, the data set has entropy close to 1. 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. The data set is almost all `NonSpam`.

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

(p) (2 points) You have the following data concerning the occurrence of words in spam email:

Spam (Y/N)	'Home' Occurs	'BitCoin' Occurs
Y	Y	Y
N	Y	Y
Y	Y	N
Y	N	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. $\frac{\frac{2}{3}\times\frac{3}{5}}{\frac{3}{5}}$ B. $\frac{\frac{2}{3}\times\frac{4}{5}}{\frac{3}{5}}$ C. $\frac{\frac{2}{5}\times\frac{3}{5}}{\frac{2}{5}}$ D. $\frac{1\times\frac{3}{5}}{\frac{2}{5}}$

(q) (1 point) What is PCA and how is it used?

A. It is a pre-processing algorithm that picks categorical attributes that can be used for k-means clustering.

B. It is a type of unsupervised learning that can do linear regression.

C. It is a pre-processing algorithm that can reduce the dimension of your input data set by finding linear relationships in your data.

D. It is a way to fill in missing data in your training data.

(r) (1 point) During PCA the eigenvalues and vectors of a covariance matrix are calculated. What do these tell you?

A. The largest eigenvalue tells you which categorical feature to pick.

B. The D largest (for some value of D) eigenvalues tells you which eigenvectors to project to maximise the variance of the projected data.

C. The eigenvectors and values define a hyperplane that can be used as an initial model when solving linear regression algorithm by gradient descent.

D. The eigenvalues are used to reconstruct missing data.

(s) (1 point) What best describes the algorithm ID3:

A. ID3 is an algorithm for learning decision trees that is not guaranteed to find the optimal tree.

B. ID3 is a post processing algorithm for PCA that finds the best three vectors to project the data onto.

C. ID3 is a method to give compressed decision trees.

D. ID3 is a method to produce small optimal decision trees.

Optional Questions for a higher grade

2. General questions

   (a) (2 points) Often you are working with a data set that is not perfect, and there will be some entries that contain missing values. Describe two strategies to deal with missing values.

   (b) (2 points) You are trying to build a classifier that classifies multiple classes using a classification algorithm that can only classifier data into two classes (for example logistic regression). Describe two strategies for combining classifiers.

   (c) (1 point) Explain with an example what one-hot encoding is, and explain why it is necessary?

3. k-means clustering.

   (a) (2 points) With k-means clustering you have to decide how many clusters to cluster your data into. Describe how to use the elbow method to try and find the optimal number of clustering. Hint it has something do with looking at how the error function improves with a different numbers of clusters.

   (b) (3 points) The K-means clustering works by gradient descent. The gradient descent algorithm is not always guaranteed to find the global minimum. Explain why this is the case.

4. Overfitting and Hyper-parameters.

   (a) (1 point) Define hyper-parameters.

   (b) (1 point) How can you test if your machine learning model is suffering from overfitting.

   (c) (3 points) Describe two different strategies to avoid overfitting. For each strategies you must try to explain how it avoids overfitting.

5. Entropy and ID3.

   (a) (2 points) Explain the mathematics behind your answer to question 1 (o).

   (b) (3 points) Explain how a feature is chosen at the root of a decision tree in the ID3 algorithm. You should explain the mathematics and give an intuitive explanation of what the formulas mean. You can refer to your answer to part (a) of this question.