Machine Learning 1DL034
2020-03-20

| Question | Points | Score |
|---|---|---|
| 1 | 20 | |
| 2 | 10 | |
| 3 | 10 | |
| 4 | 5 | |
| 5 | 10 | |
| Total: | 55 | |

Draft Correct Version 2020-03-27

# Exam Instructions

In order to get a 3 you have to get 80% on question 1. If do not get 80% on question 1, then I will not mark the other questions. Your will get a 4 if you pass 50% on the remaining questions 2—4 and a 5 if you get more than 80% on the same remaining questions.

This exam is an online take home exam. You can use a computer or a calculator to do calculations. You must not use scikit-learn or any other machine learning library to do calculations for you. You must show all your workings.

You are expected to treat this as a normal exam, and not refer to any of your notes or any additional material online or printed.

If possible please supply your answers as one-single PDF file. If you are running out of time, then you can send a zip file containing all your pages. You can write your answers using LaTeX/openoffice or whatever you normally use. You are free to provide handwritten solutions, but you should scan or photograph each page. If you solutions come as multiple pages, then for each page please make it obvious which questions or questions are being answered.

# Questions

1. General Questions

   (a) (1 point) Which of the following problems are more suitable for classification? There might be more than one answer, and you must indicate all correct answers to obtain 1 point.

   **A. Predicting if a job advert is an advert for a real job or a scam.**

   B. Predicting the blood alcohol content of a person based on data including the persons BMI, how many and what types of drink they have had, and how long they have been drinking.

   **C. Predicting which film a user will watch next based on their preferences and their past history.**

   > **Solution: A lot of people did not get this one, and when people were close to passing I awarded half points for one correct answer. The actual film has no numerical value. It is just a class.**

   D. Predicting the number of likes a post on a social media site will receive.

   (b) (1 point) Consider the following data-sets and problems. Which data sets would require un-supervised learning. There might be more than one answer, and you must indicate all correct answers to obtain 1 point.

   **A. An un-labelled set of pictures of animals. Your job is to classify the pictures into different classes of animals.**

   B. A labelled set of pictures of cats, dogs and automobiles.

   C. A data set containing the closing price of every house sold in Uppsala since 2000. You are to predict the final house price of a house based on its location and other data.

   (c) (1 point) Which of the following features are categorical. There might be more than one answer, and you must indicate all correct answers to obtain 1 point.

   **A. The gender of a person.**

   B. The age of a person.

   **C. Which country the person lives in.**

   D. The weight of a person.

   (d) (1 point) Consider using gradient descent to learn a hypothesis $h_\theta = \theta_0 + \sum_{i=1}^{n} \theta_i x_i$ for regression. During gradient descent an error/loss function $J$ is minimised. Does the gradient descent for linear regression *always* converge to a global minimum for *any* training set?

   **A. True**

   > **Solution: In Gradient descent for linear regression there is only one minimum which is the global minimum.**

   B. False

(e) (1 point) Again, consider using gradient descent to learn a hypothesis $h_\theta = \theta_0 + \sum_{i=1}^{n} \theta_i x_i$ for regression. During gradient descent an error/loss function $J$ is minimised. At the end of gradient descent for linear regression does the function $J$ always equal 0 for *any* training set?

A. True

**B. False**

> **Solution: You minimise $J$, but the global minimum might still be when $J$ is not equal to 0. In general, there is no straight line that goes through all the points.**

(f) (1 point) Why should the training set always be a different set from the validation set. Please indicate the correct answer.

**A. To avoid over-fitting.**

B. Learning algorithms become inefficient if the training set is too large.

(g) (1 point) You are developing a classifier to detect cancer. The algorithm should report true if the patient has cancer. You want true cancer patients not to be diagnosed as non-cancer patients. Given confusion matrix which of the following situations is better.

> **Solution:** I realise after looking at this question again, it is not phrased so well, and either answer could be correct. So you all get a free point.

**A. A high false positive rate.**

B. A high true positive rate.

(h) (2 points) Given the following data set:

| x | y |
|---|---|
| 1 | 2 |
| 2 | 2 |
| 3 | 4 |

You are 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 $J$.

> **Solution:** Without an annotation I will not accept no answer correct.

A. $J(x, \theta_0, \theta_1) = (2 - (\theta_0 + \theta_1)) + 42 - (\theta_0 + 2\theta_1))$

B. $J(x, \theta_0, \theta_1) = \frac{1}{2}((2 - (\theta_0 + \theta_1))^2 + (42 - (\theta_0 + 2\theta_1))^2 + (43 - (\theta_0 + 3\theta_1))^2)$

**C.** $J(x, \theta_0, \theta_1) = \frac{1}{2\times 3}\left((2 - (\theta_0 + \theta_1))^2 + (2 - (\theta_0 + 2\theta_1))^2 + (4 - (\theta_0 + 3\theta_1))^2\right)$
**there was a misprint in the exam.** $J(x, \theta_0, \theta_1) = \frac{1}{2\times 3}\left((2 - (\theta_0 + \theta))^2 + (2 - (\theta_0 +$

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

(i) (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 | N |
| N | 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.** $\frac{\frac{2}{3}\times\frac{3}{5}}{\frac{2}{5}}$ B. $\frac{\frac{2}{3}\times\frac{3}{5}}{\frac{3}{5}}$ C. $\frac{\frac{2}{5}\times\frac{3}{5}}{\frac{2}{5}}$ D. $\frac{1\times\frac{3}{5}}{\frac{2}{5}}$

(j) (1 point) Consider using logistic regression with a linear hypothesis to classify 2-dimensional data points separated into two classes. Which of the following statements is true, note that there might be more than one correct answer and you must indicate all correct answers to get full marks:

A. Logistic regression is able to classify any data set.

B. Logistic regression can only classify a data set if and only if there is no overlap in the classes.

> **Solution:** I put an 'if and only if' in the question. So this is not true.

**C. Logistic regression requires the two classes to be linearly separable.**

D. Logistic regression can only classify the two classes if and only if it is possible to use linear regression to separate two classes.

> **Solution:** Think back to the lectures, you need the logistic function to do classification.

(k) (1 point) Given two probabilities $p_1$ and $p_2$. If you are told that the entropy $H(p_1, p_2) = -(p_1 \log_2 p_1 + p_2 \log_2 p_2)$ is 0, then what are the possible values for $p_1$ and $p_2$. There maybe more than one answer. You must indicate all correct answers to get full marks.

> **Solution:** Technically $p_1 = (1 - p_2)$ is also correct if $p_1 = 0$. Although it is not true for all values of $p_1$. I've accepted it as a correct answer as long as you have the two other correct answers. On its own I won't take it as a correct answer. Don't forget that with entropy $0 \log 0$ is defined as 0.

A. $p_1 = (1 - p_2)$

B. $p_1 = \frac{1}{2} = p_2$.

**C.** $p_1 = 0$**,** $p_2 = 1$**.**

**D.** $p_1 = 1$**,** $p_2 = 0$**.**

(l) (1 point) The ID3 algorithm for constructing decision trees always constructs the smallest decision tree possible for *any* training set.

A. True

**B. False**

> **Solution: Constructing the smallest decision tree is NP hard. ID3 is just a heuristic.**

(m) (2 points) 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 dataset

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

(n) (2 points) 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. The exam had a misprint:-

A. $H(P) - (H(P|\text{Windy} = \text{Yes}) + H(P|\text{Windy} = \text{Yes}))$
B. $H(P) - (\frac{3}{7}H(P|\text{Windy} = \text{Yes}) + \frac{4}{7}H(P|\text{Windy} = \text{Yes}))$
C. $H(P) + (\frac{1}{3}H(P|\text{Windy} = \text{Yes}) + \frac{1}{4}H(P|\text{Windy} = \text{Yes}))$
D. $H(P) + (\frac{3}{7}H(P|\text{Windy} = \text{Yes}) + \frac{4}{7}H(P|\text{Windy} = \text{Yes}))$

> **Solution:** Instead I should have written. I will either except no answer correct or the highlighted choice.

A. $H(P) - (H(P|\text{Windy} = \text{Yes}) + H(P|\text{Windy} = \text{No}))$
**B.** $H(P) - (\frac{3}{7}H(P|\text{Windy} = \text{Yes}) + \frac{4}{7}H(P|\text{Windy} = \text{No}))$
C. $H(P) + (\frac{1}{3}H(P|\text{Windy} = \text{Yes}) + \frac{1}{4}H(P|\text{Windy} = \text{No}))$
D. $H(P) + (\frac{3}{7}H(P|\text{Windy} = \text{Yes}) + \frac{4}{7}H(P|\text{Windy} = \text{No}))$

(o) (1 point) As part of the principle component algorithm the eigen-vectors and eigen-values are calculated of the co-variance matrix of the training data. What does the largest eigen-value tell you? There is only one correct answer.

A. The number of dimensions your reduced data set will have.
**B. The direction to project in order to maximise the variance in that dimension.**
C. A normalising coefficient for the first feature that you need to avoid over-fitting.

(p) (1 point) For logistic regression it is possible to use a regularisation parameter $\lambda$ 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 mean zero.

B. Increasing $\lambda$ reduces the considered dimension of the training data.

**C. It avoids over-fitting on the training data by forcing gradient descent to learn small weights $\theta_i$.**

2. (10 points) Using the training data from Question 1 part m use the ID3 algorithm to construct a decision tree. You should show your workings.

**Solution:** Use Chuong Chu's solution. I've emailed him to ask for permission.

3. K-fold validation.

   (a) (3 points) Describe $k$-fold cross validation.

   (b) (2 points) Describe how you can use $k$-fold cross validation to tune hyper-parameters of an algorithm.

   (c) (2 points) How does $k$-fold cross validation help to reduce over-fitting?

   (d) (3 points) Why is grid-search alone not enough to avoid over-fitting? Describe how would you combine grid-search with cross-validation to reduce over-fitting?

4. (5 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.

5. Clustering and Nearest Neighbour classifiers.

   (a) (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.

   (b) (5 points) You are to implement a system the recommends films (movies) that a person is likely to watch based on their and other users preferences, past viewing history. Sketch how you implement such a system using clustering. `based on their preferences and their past history.`

(c) (2 points) Given images that contain handwritten digits (numbers) it is possible to use nearest a neighbour classifier to learn the different classes. If you want to improve your system so that it is invariant under rotations, then there are two possible approaches. Describe the two approaches and explain the advantages and disadvantages. A system is said to be invariant rotations an image and the same image rotated by any number of degrees will be in the same class.