Machine Learning 1DL034
2020-06-10 8-13.

Question	Points	Score
1	20	
2	10	
3	6	
4	5	
5	8	
Total:	49	

Exam Instructions

In order to get a 3 you have to get 80% on question 1. Your will get a 4 if you pass 50% on the remaining questions 2—5 and a 5 if you get more than 80% on the same remaining questions. Note that in order to at a 4 or a 5 you must get 80% on question 1.

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.

In the multiple choice questions, some of the questions have more than one answer. If there is more than one answer then I expect you to circle/indicate all correct answers. If there is more than one answer then I will tell you.

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.

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. I will only accept a PDF file as a submission. I will not accept a word or openoffice document as a submission. You are free to provide handwritten solutions, and 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.

I will be online during parts of the exam, and I will be able to answer questions. I will only answer questions on clarifying exam questions. You can contact me via email `justin.pearson@it.uu.se`. I can guarantee that I will be online 8:30-9:30 to answer any initial problems, and other times I will be online, but I cannot guarantee an instance response.

Questions

1. General Questions

   (a) (1 point) The goal of a machine learning algorithm will find the best hypothesis that explains the training data.

   A. True B. False

   (b) (1 point) You always use the same loss (error) function when training your machine learning algorithm and to evaluate the performance of your algorithm.
   A. True **B. False**

   (c) (1 point) One of these algorithms is an unsupervised learning algorithm. Which one is it?

   A. Linear Regression
   B. Principle Component analysis
   C. Support vector machines
   D. Naive Bayes Classification

   (d) (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. Principle Component Analaysis
   C. Naive Bayes Classification
   D. K-means clustering

   (e) (1 point) Which of the following data types are categorical variables. There is more than one answer. You must circle or indicate all possible correct answers.

   A. Lives in Sweden (Yes or No)
   B. Weight (in kg)
   C. Height (in cm)
   D. Likes Ice Cream (Yes or No)

   (f) (1 point) Which of these problems is a regression problem. There is more than one answer, please circle or indicate all possible correct answers.

   A. Predicting the probability that a message is spam or not.
   B. Labelling a message as spam or not.
   C. Predicting the final sale price of a house.
   D. Predicting if a customer to a web shop will if a student will pass or fail an exam.

   (g) (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

   (h) (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 are true (there is only one correct answer):

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

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

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

(i) (1 point) Logistic regression requires all variables to be categorical

A. True

B. False

(j) (1 point) Which of these statements best describe what is learnt when doing logistic regression:

A. A linear hypothesis $h_\theta = \theta_0 x_0 + \sum_{i=1}^{n} \theta_i x_i$ such that $h(x) = 1$ if the data point x belongs to the class and $h(x) = 0$ if x does not belong to the class.

B. A hyperplane defined by the set of values for which $\theta_0 x_0 + \sum_{i=1}^{n} \theta_i x_i = 0$ that separates the data into two regions with those points belonging to the class and those points not belonging the class.

C. The value $\theta_0 x_0 + \sum_{i=1}^{n} \theta_i x_i$ is the probability that the point x_i belongs to the class or not.

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

1. Try to run algorithm for different centroid initialisation
2. Adjust 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.**

(l) (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
N	N	Y
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. $\frac{\frac{2}{3} \times \frac{3}{5}}{\frac{3}{5}}$ **B.** $\frac{\frac{2}{3} \times \frac{3}{5}}{\frac{4}{5}}$ C. $\frac{\frac{2}{5} \times \frac{3}{5}}{\frac{2}{5}}$ D. $\frac{1 \times \frac{3}{5}}{\frac{2}{5}}$

(m) (1 point) What does a naive Bayes classifier assume?

A. Most features are independent of each other.

B. No features are independent of each other.

C. Some features are independent of each other.

D. All features are independent of each other.

(n) (1 point) To avoid over-fitting you can use some of the following techniques. There is more than one answer, please indicate all techniques that can be used.

A. Principle Component Analysis.

B. Using a separate training and validation set.

C. Using k-fold validation.

D. Using logistic regression instead of linear regression.

(o) (1 point) As part of the principle component algorithm the eigenvectors and eigenvalues are calculated of the co-variance matrix of the training data. What does the largest eigenvalue 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) (2 points) Given a d-dimensional data set with n points, $x_1, \ldots, x_n$, where each x_i belongs to $\mathbb{R}^d$ after running principle component analysis (PCA) you pick the first P principle component directions. Your dimension reduced dataset is now n points $y_1, \ldots, y_n$, where each y_i belonging to $\mathbb{R}^P$, is a P-dimensional point. Can you always reconstruct any data point x_i from y_i?

A. Yes, if $P < d$

B. Yes, if $P < n$,

C. Yes, if $P = d$,

D. It is never possible.

(q) (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 mean zero.

B. Increasing λ 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 θ_i.

D. Increasing λ will make the algorithm converge to a solution much more quickly.

(r) (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.**

2. 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?

3. Hyper-parameters.
   (a) (2 points) What are hyper-parameters? Give some examples of hyper-parameters in some of the algorithms that you have seen in the course.
   (b) (2 points) How do you implement grid-search to tune hyper-parameters of an algorithm?
   (c) (2 points) Why would k-fold cross validation be preferred over grid-search for tuning hyper parameters.

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) (3 points) Describe different techniques that you can use to find a global minimum in K-means clustering.
   (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.