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

Question	Points	Score
1	20	
2	6	
3	5	
4	4	
5	6	
Total:	41	

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. It is my intention that there should be no mistakes 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. You can write your answers using LaTeX, openoffice/word or whatever you normally use. If you do use some piece of software, then I will only accept a PDF file. 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 `mailto:justin.pearson@it.uu.se`. I can guarantee that I will be online 8:00-9:00 to answer any initial problems. If by some chance the student portal goes down then you can email me the exam as long as the timestamp is within the exam time. Please only email me the exam if you have trouble with student portal. There is no point submitting the exam twice to me. Despite its problems student portal is quite reliable when handling assignment submissions.

Questions

1. General Questions

   (a) (1 point) On of these algorithms is an unsupervised learning algorithm. Which one is is?

   A. Linear Regression
   B. [✓ K-means clustering]
   C. Support vector machines
   D. Naive Bayes Classification

   (b) (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. There is more than one answer, please circle or indicate all possible correct answers.

   A. Linear Regression
   B. [✓ Logistic Regression]
   C. [✓ Naive Bayes Classification]
   D. K-means clustering

   (c) (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. [✓ Gender (Male or Female)]
   B. Weight (in Kg)
   C. Age
   D. [✓ Member of Gotland Nation Yes or No.]

   (d) (1 point) Which of these classification problem 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 somebodies exam grade based on the number of lectures they attend and their score on assignments given during the course.]
   D. [✓ Deciding if a patient has cancer or not.]

   (e) (1 point) You are building a self driving car, and you are building a system to recognise pedestrians (fotgängare). If the car kills pedestrians then you will go to prison. The algorithm output positive if there is a pedestrian in the path of the car. Which of the following are you trying to minimise?

   A. [✓ False Negative]
   B. True Positive
   C. False Positive

   (f) (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.

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

x	y
1	2
2	3
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.

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. [✓ $J(x, \theta_0, \theta_1) = \frac{1}{2\times 3}((1 - (\theta_0 + 2\theta_1))^2 + (2 - (\theta_0 + 3\theta_1))^2 + (3 - (\theta_0 + 4\theta_1))^2)$]

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

(h) (1 point) Logistic regression is a regression algorithm:

A. [✓ True]

B. False

(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) (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. $\frac{1\times\frac{3}{5}}{\frac{2}{5}}$

(l) (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.]

(m) (1 point) Which of the following options can be used 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.]

(n) (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.

(o) (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.

(p) (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.

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

A. True

B. [✓ False]

2. 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.

3. (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.

4. (4 points) What is over fitting? Describe some techniques that you can use to avoid over fitting.

5. The question concerns the derivation and meaning of principle component analysis (PCA).

(a) (3 points) Giving your data set which is a collection of K , n dimensional vectors $x_1, \dots, x_k \in \mathbb{R}$. Let $\overline{x}$ be the mean vector

$$\frac{1}{K}\sum_{i=1}^{K} x_i$$

and σ^2 be the vector of variances

$$\sigma^2 = \frac{1}{K}\sum_{i=1}^{K}(x_i - \overline{x})^2$$

What is the goal of PCA with respect to the variance of data set and the variance of the transformed data set?

(b) (3 points) When doing PCA you compute a special matrix C the covariance matrix and look for solutions to the following matrix equation w is a vector and λ is a real number

$$C.w = \lambda w$$

If you find vectors $\vec{w}_i$ and real numbers λ_i satisfying the equations then what do these vectors and numbers tell you? How will they be used when doing PCA.