UNIVERSITY OF LONDON

GOLDSMITHS COLLEGE

Department of Computing

B. Sc. Examination 2016-17

IS51009C
Machine Learning

Duration: 2 hours 15 minutes

Date and time:

This paper is in two parts: part A and part B. You should answer ALL questions from part A and TWO questions from part B. Part A carries 40 marks, and each question from part B carries 30 marks. The marks for each part of a question are indicated at the end of the part in [.] *brackets.*

There are 100 marks available on this paper.

THIS PAPER MUST NOT BE REMOVED FROM THE EXAMINATION ROOM

Part A

Answer all questions. Each question has only one correct answer.

Question 1 General Questions

(a) Which of the following constitutes a Machine Learning task

i. Supervised Learning

ii. Unsupervised Learning

iii. Both the above

iv. None of the above.

[3]

(b) You are given a problem where the output variable y consists of a range of real values, i.e. for every data-point i in the training data, the corresponding y_i value is a real number (e.g., $y_1 = 1.5$, $y_2 = 2.3$, $y_3 = 3.4, ...$). Which of the following would be most applicable?

i. Regression

ii. Classification

iii. Any type of supervised learning

iv. Any type of unsupervised learning

[3]

(c) When applying nested cross-validation, the entire data-set is split to...

i. a training and testing set

ii. a training and validation set

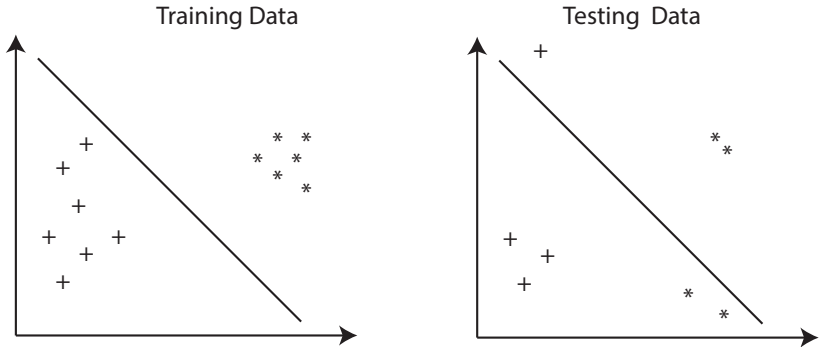
iii. a training, testing and validation set

iv. None of the above

[3]

Question 2

You are given the following linear classifier (shown on both training and testing data), with data belonging to class 1 (represented with +) or class 2 (represented with $*$). Note that everything on the left of the line is classified as class 1 (+), and everything on the right of the line as class 2 ($*$).



(a) What is the accuracy of this classifier on test data for Class 1 (+) and Class 2 (*)? [3]

i. Accuracy for Class 1 (+) is $\frac{2}{8}$ and (ii) for Class 2 (*) $\frac{2}{4}$.
ii. Accuracy for Class 1 (+) is $\frac{3}{4}$ and (ii) for Class 2 (*) $\frac{2}{4}$.
iii. Accuracy for Class 1 (+) is $\frac{3}{4}$ and (ii) for Class 2 (*) $\frac{3}{8}$.

(b) How many samples are **incorrectly** classified on the test data for Class 1 (+) and Class 2 (*)? [3]

i. 1 for Class 1, 2 for Class 2
ii. 3 for Class 1, 3 for Class 2
iii. 2 for Class 1, 2 for Class 2

(c) Which of the following confusion matrices corresponds to the this classifier applied on the test data? (with rows corresponding to actual class and columns to predicted class) [3]

i.

		Predicted Class	
		Class 1	Class 2
Actual Class	Class 1	3	1
	Class 2	2	2

ii.

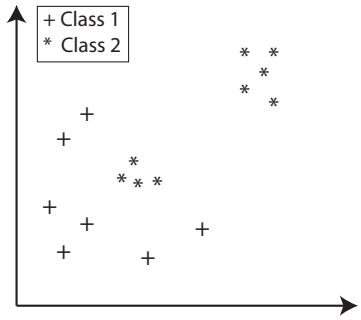
		Predicted Class	
		Class 1	Class 2
Actual Class	Class 1	1	3
	Class 2	2	3

iii.

		Predicted Class	
		Class 1	Class 2
Actual Class	Class 1	2	1
	Class 2	3	2

Question 3

You are given the 2D data illustrated below:



(a) Which of the following is **incorrect**: [3]

i. The data is linearly separable (can be separated by drawing a line)

ii. The data is not linearly separable

iii. There are two classes in the dataset

(b) To correctly classify these data, we can: [4]

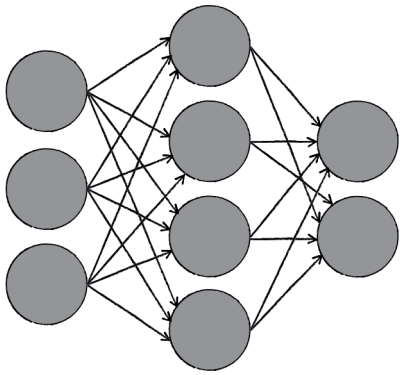
i. Use a more complex boundary than a line to separate the data points

ii. Project the points onto a higher dimensional space

iii. Any of the above

iv. None of the above

Question 4 Given the following artificial neural network.



(a) How many hidden layers does it have? [3]

i. 3

ii. 4

iii. 1

iv. 2

(b) There is a bias in the network. [3]

i. True

ii. False

iii. We do not have enough information in the diagram.

(c) Which of the following algorithms is used to fit the neural network given? [3]

i. K-means

ii. Perceptron

iii. Backpropagation

iv. Nearest Neighbour

Question 5 Overfitting

(a) Suppose we are doing linear regression and we want to make sure we are not overfitting. We use regularization in order to do this, however, our regularization parameter is too large. What will happen? [3]

i. We will underfit the data

ii. We will end up with a low bias

iii. We will end up with a high variance

iv. We will overfit the data

(b) How can we determine the optimal regularization parameter (λ) for linear regression? [3]

i. There is no need to optimize the regularization parameter λ for linear regression.

ii. By plotting a bias - variance curve over various λ's.

iii. By plotting a bias - variance curve over various iterations of the gradient descent algorithm.

iv. By plotting the cost function, J over various λ's.

Part B

Question 6 Batch Gradient Descent

The batch gradient descent algorithm for linear regression is as follows.

Repeat until convergence

$$\theta_j := \theta_j + \alpha \sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)}))x_j^{(i)} \text{ for every j} \quad (1)$$

Assume you have the following 3 training examples to fit your linear regression model: $(x_{1,1} = 1000, x_{1,2} = 100, y_1 = 5)$,$(x_{2,1} = 3000, x_{2,2} = 300, y_2 = 3)$,$(x_{3,1} = 400, x_{3,2} = 90, y_3 = 0)$.

Assume you decide to use the batch gradient descent algorithm given by equation 1, answer the following questions.

(a) What is the value of m for the training examples given? [2]

(b) Write the equation for a suitable hypothesis function for this example. [4]

(c) How many j's do you need to solve for in this example? [2]

(d) What initial assumptions do you need to make to initiate your gradient descent algorithm? [5]

(e) When has the algorithm converged? [2]

(f) Why is it necessary to scale the features? [4]

(g) In what cases will your algorithm NOT converge? [2]

(h) What should we plot to determine if the algorithm is converging over time? [2]

(i) i. What is the batch gradient descent algorithm for logistic regression? [5]

ii. What is the difference between the batch gradient descent algorithm for logistic regression and linear regression? [2]

Question 7 Logistic regression

Given the logistic function $g(z) = \frac{1}{1+e^{-z}}$ and the hypothesis function $h_\theta(x) = g(\theta^T x)$.

Answer the following questions.

(a) Draw the logistic function. [5]

(b) Write out the equation for your hypothesis function if you had 3 input variables (for logistic regression). [6]

(c) Given you have 3 input variables for your logistic regression model, how many equations will you need to solve using gradient descent? [2]

(d) Assume you have the following dataset.

hrs studying	hrs commuting	hrs working	pass
200	200	30	1
50	500	16	1
50	89	120	0

Scale the "hrs studying" input features using the scaling equation $X_i = \frac{X_i - \mu_i}{S_i}$. [6]

(e) Assume the output of your hypothesis function $h = 0.51$. What will be the final output of your logistic regression model? [4]

(f) Assume you have fit your logistic regression model to a hypothesis with a polynomial of order 10 and you are getting $J = 0$. You can assume the model is being "overfit".

i. Why is it not a good idea to overfit the model to the data? [4]

ii. How does regularization address the issue of overfitting? [3]

Question 8

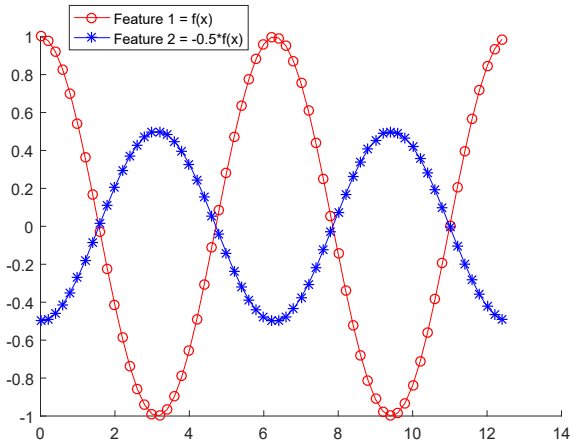
(a) Assume we are given the following 1 dimensional data $\mathbf{x} = [3, 4, 5, 6, 7]$ (i.e., $x_1 = 3, x_2 = 4, x_3 = 5, x_4 = 6, x_5 = 7$, where $N = 5$)

i. Estimate the mean ($\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$) and variance of $\mathbf{x}$ (by using $var(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$) [3]

ii. Suggest one or more elements that could be added to $\mathbf{x}$ in such way that the mean remains the same and the variance decreases. [3]

iii. Suggest one or more elements that could be added to $\mathbf{x}$ in such way that the mean remains the same and the variance increases. [3]

(b) You are given the following data, consisting of $2D$ features with T samples (i.e., data matrix $\mathbf{X} \in \mathbb{R}^{2\times T}$)



i. If we estimate the covariance of feature 1 to feature 2, will it be positive, negative or zero? [3]

ii. If we apply PCA on this data, how many components would we need to fully describe the data? [3]

(c) Assume that are given a linear system of the form

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t$$

where the matrix $\mathbf{A}$ is defined as

$$\mathbf{A} = \begin{pmatrix} 0.8 & 0.5 \\ 0.2 & 0.5 \end{pmatrix}$$

i. Starting with $\mathbf{x}_0 = [0, 1]^T$, multiply with matrix $\mathbf{A}$ to arrive at $\mathbf{x}_1 = \mathbf{A}\mathbf{x}_0$. Write down the result $\mathbf{x}_1$. What do you observe with respect to the values of $\mathbf{x}_0$? Explain your answer. [3]

ii. Assume a general linear system of the form $\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t$. How can we tell if the system has a steady state? Explain your answer. [3]

iii. Can you tell if the linear system above has a steady state by just looking at $\mathbf{A}$? Explain why. [3]

(d) Principal Component Analysis (PCA) is a method that maximizes the covariance of our data in a latent space $\mathbf{Y}$, with $\mathbf{Y} = \mathbf{W}^T\mathbf{X}$. We denote the covariance matrix as $\mathbf{S}$, i.e. $\mathbf{S} = \text{covariance}(\mathbf{X})$.

i. Assume that you are given the following optimization problem. Which value is assigned to $\mathbf{W}$? Explain your answer.

$$\arg\max_{\mathbf{W}} \text{cov}(\mathbf{Y}) = \arg\max_{\mathbf{W}} \mathbf{W}^T\mathbf{S}\mathbf{W}$$

[3]

ii. How would you augment the problem mentioned above in order to get a reasonable solution? Explain your answer. [3]