# UNIVERSITY OF LONDON

# GOLDSMITHS COLLEGE

## Department of Computing

## B. Sc. Examination 2017-18

## IS53051A
## 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. Multiple-choice questions may have more than one correct answers, and in that case all must be given for full credit.**

**Question 1** General Questions. **Multiple-choice questions may have more than one correct answers, and in that case all must be given for full credit.**

(a) You are given a problem where the output variable $y$ can only take the values of 0 or 1. Which of the following statements are correct? [4]

i. This is an unsupervised learning problem

ii. This is a supervised learning problem

iii. We could apply linear regression to this problem

iv. We could apply logistic regression to this problem

(b) Which of the following methods performs unsupervised learning? [4]

i. Principal Component Analysis (PCA)

ii. Logistic Regression

iii. Linear Regression

iv. None of the above

(c) Which of the following can help prevent overfitting? [4]

i. Regularization

ii. Using polynomial hypothesis functions

iii. Using complex hypothesis functions

iv. None of the above

(d) Which of the following are classification methods? [4]

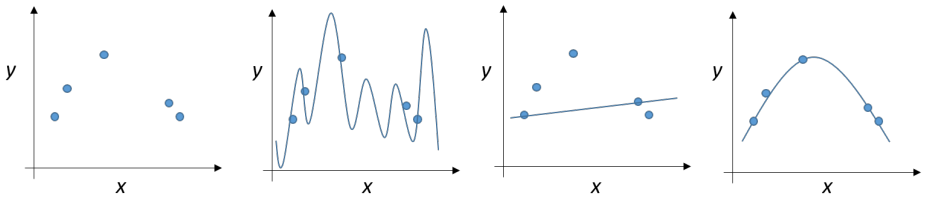
i. Logistic Regression

ii. Linear Regression

iii. k-Nearest Neighbours

iv. Principal Component Analysis

(e) You are given a set of training samples $(x^{(i)}, y^{(i)})$, illustrated in the figure below, along with a set of three hypothesis functions ($h_1$, $h_2$, $h_3$).




samples $h_1$ $h_2$ $h_3$

i. Which one of the above hypothesis functions would you choose? Briefly describe why. [2]

ii. Which of the above hypothesis functions is underfitting the data? Briefly describe why. [2]

iii. Which of the above hypothesis functions is overfitting the data? Briefly describe why. [2]

**Question 2** Cross Validation. **Multiple-choice questions may have more than one correct answers, and in that case all must be given for full credit.**

You are given a dataset consisting of $m$ samples, with each training example represented as $(x^{(i)}, y^{(i)})$. Assume that you want to train a k-nearest neighbour (k-NN) classifier on the given data.

(a) Which is the purpose of applying cross-validation on a dataset? [2]

(b) Assume you are performing 5-fold cross-validation. For each of the five iterations, this means you split your data into a few subsets (or data-splits).

i. What are the names of these subsets? [2]

ii. Which data subset (or data-split) from the above would you use in order to evaluate the performance of different parameter values during cross-validation? [2]

(c) Which of the following correspond to k-NN parameters that you can tune during nested cross-validation? [4]

i. number of neighbours

ii. distance (or, similarity) function

iii. number of data

iv. none of the above

**Question 3** Evaluation.

(a) For a given problem, the correct output labels $y_{\text{true}}$ and the predicted (by some classifier) labels $y_{\text{pred}}$ are given below. Draw the confusion matrix for this problem, by considering rows to correspond to actual labels, and columns to predicted labels. [4]

$$y_{\text{true}} = [1, 1, 0, 2, 1, 2, 0, 1]$$
$$y_{\text{pred}} = [1, 0, 0, 1, 1, 1, 0, 1]$$

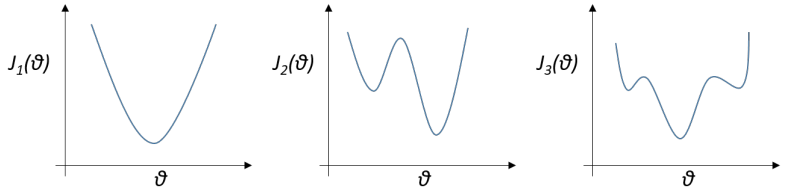
(b) Draw a $2 \times 2$ (empty) confusion matrix, indicating which entries correspond to true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Assume the first class is the positive class, and that rows correspond to actual labels and columns to predicted labels. [4]

# Part B

**Answer two questions**

**Question 4** Gradient Descent, Linear Regression

(a) Assume that we want to optimize a parameter $\theta$ with respect to the cost functions $J(\theta)$ (illustrated below). In which of these plots is gradient descent guaranteed to find the (globally) optimal value? Briefly explain why. [6]




(a) $J_1(\theta)$ (b) $J_2(\theta)$ (c) $J_3(\theta)$

(b) You are given a problem of predicting the temperature in a specific room ($y$), given the following measurements ($x$):

$x_1$ temperature in a proximal room

$x_2$ current weather at location

$x_3$ insulation properties of the room

i. Assume your goal is to perform linear regression to obtain a model for estimating the room temperature, $y$. Write the equation for a suitable hypothesis function for doing so. [5]

ii. How many parameters would you need to update if applying linear regression via gradient descent in the problem above? How many parameters would you need in general, given a problem with $n$ features? [3]

*Note: please turn over for the next parts of this question*

(c) Assume that you are trying to fit a linear regression problem, with training data shown in the table below (where, e.g., $x^{(2)} = 4$ and $y^{(2)} = 2$) .

| $y$ | $x$ |
|---|---|
| 1 | 2 |
| 2 | 4 |
| 3 | 6 |

i. Given that you are using the hypothesis function $h_\theta(x) = \theta_0 + x_1\theta_1$, with values $\theta_0 = 0$ and $\theta_1 = 1$, what would be the value of the mean squared loss function shown below? [5]

$$J(\boldsymbol{\theta}) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

*Note: giving the result as a fraction is fine.*

ii. Given the initial value of $\theta_1 = 1$, compute the new value of $\theta_1$ by applying the general gradient descent rule shown below. [3]

$$\theta^{\text{new}} = \theta^{\text{old}} - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_{\theta^{\text{old}}}(x^{(i)}) - y^{(i)})x^{(i)}$$

*Note: Use a learning rate $\alpha = \frac{1}{100}$. You don't have to give a single real number as the result, a fraction is fine.*

iii. Does the value of $\theta_1$ decrease or increase after the update you computed above? Explain your answer with respect to the direction of the gradient of $J(\theta)$. [3]

(d) How does the learning rate affect convergence (e.g., in linear regression)? Discuss any problems that may arise when (i) learning rate is too small, and (ii) learning rate is too large. Draw one plot while discussing (i) and another plot while discussing (ii). [5]

**Question 5** Logistic regression, Gradient Descent, Regularization

Answer the following questions. You are given that the logistic function is $g(z) = \frac{1}{1+e^{-z}}$ and that the hypothesis function for logistic regression is $h_\theta(x) = g(\theta^T x)$.

(a) Draw the logistic function $g(z)$. Make sure you label the value of $z$ where $g(z)$ crosses the y-axis. [2]

(b) Describe briefly why the logistic function is suitable for classification tasks. [3]

(c) Assume the output of logistic regression is $h_\theta(x) = 0.65$ for a given $x$. How can we interpret this result? How can we connect the result to the probability of having an output of 1 or an output of 0? [5]

(d) You are given a problem of predicting whether a room is too hot $y = 1$ or not $y = 0$, given the following measurements ($x$):

$x_1$ temperature in a proximal room

$x_2$ current weather at location

$x_3$ insulation properties of the room

i. Assume your goal is to perform logistic regression to obtain a model for estimating the target variable $y$. Write the equation for a suitable hypothesis function for doing so. [5]

ii. Assume that the values of $x_1$, $x_2$ and $x_3$ are in different ranges. Provide two ways of normalizing your data before applying gradient descent. Name one advantage of scaling your features. [6]

(e) Assume we are minimizing a loss (or, cost) function on a given problem, i.e.:

$$J(\boldsymbol{\theta}) = \frac{1}{2m} \sum_{i=1}^{m} \text{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

i. How would we apply regularization to the function $J(\boldsymbol{\theta})$? Write the equation of a regularized cost function. [3]

ii. Briefly describe the effect of regularization on this optimization problem, with respect to both overfitting and underfitting. [6]

**Question 6** PCA, Linear Systems

(a) Assume we are given the following 1-dimensional data $\mathbf{x} = [1, 1, 4, 7, 7]$, where $x_1 = 1, x_2 = 1, x_3 = 4, x_4 = 7, x_5 = 7$ and $m = 5$). Estimate (i) the mean ($\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$) and (ii) the variance of $\mathbf{x}$ (by using $var(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{x})^2$) [4]
*Note: fractions are fine for full marks - no need to calculate result as real number*

(b) Assume that for a given problem, we have two features (our data matrix is $\mathbf{X} \in \mathbb{R}^{2\times T}$). Assume we know the covariance matrix $\mathbf{S}$, that is given below, describing the covariance between feature 1 and feature 2 (row/column 1 and 2 respectively).

$$\mathbf{S} = \begin{bmatrix} 0.5 & -0.5 \\ -0.5 & 0.5 \end{bmatrix}$$

What can you tell regarding feature 1 and feature 2 by looking at the covariance matrix $\mathbf{S}$? [2]

(c) i. What is the purpose of applying the Principal Component Analysis (PCA) method? [2]

ii. What does the 1st principal component represent? Illustrate your answer by drawing a graph of a few 2-dimensional data ($x$, $y$), and then plotting the 1st principal component (clearly indicate this on the graph). [2]

iii. How does the second principal component relate to the first principal component? [2]

(d) Assume that we are applying PCA on a set of $N = 5$ features (dimensions), where each feature is described as feature($i$) $= i \times \sin(x)$ (where $\times$ denotes regular multiplication), for all $i = 1$ to 5 (e.g., feature(3) $= 3sin(x)$).

i. Would you expect the covariance of feature(2) to feature(5) to be positive or negative? [2]

ii. How many components would PCA recover? Briefly explain your answer. [3]

iii. Is this number of components dependent on $N$? Explain your answer [3]

*Note: please turn over for the next parts of this question*

(e) Assume you are given a linear system of the form

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

where $\mathbf{A}$ can be equal to either $\mathbf{A}_1$ or $\mathbf{A}_2$, defined as:

$$\mathbf{A}_1 = \begin{pmatrix} 0.3 & 0.7 \\ 0.7 & 0.3 \end{pmatrix}$$

$$\mathbf{A}_2 = \begin{pmatrix} 1.2 & 0.7 \\ 0.7 & 0.6 \end{pmatrix}$$

i. Does the system have a steady state if $\mathbf{A} = \mathbf{A}_1$? Briefly explain your answer. [3]

ii. How can we tell if the system has a steady state if $\mathbf{A} = \mathbf{A}_2$? Briefly explain your answer. [3]

iii. Given the matrix $\mathbf{A}_3$ and vector $\mathbf{u}_3$,

$$\mathbf{A}_3 = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 0 & 2 \\ 0 & 1 & 0 \end{pmatrix}, \mathbf{u}_3 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

Assuming that $\mathbf{A} = \mathbf{A}_3$ in the linear system above, is the vector $\mathbf{u}_3$ a steady state of the linear system? Briefly explain your answer. [4]