UNIVERSITY OF LONDON

GOLDSMITHS COLLEGE

Department of Computing

B. Sc. Examination 2018-19

**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. 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.*

*Calculators are not allowed for this exam – you will not need one.*

**THIS PAPER MUST NOT BE REMOVED FROM THE EXAMINATION ROOM**

# Part A

**Answer all questions.** Multiple-choice questions may have more than one correct answer, and in that case **all** must be given for credit.

**Question 1**: General Questions.

(a) The process of finding parameters $\theta$ for function, $f(X;\theta) \rightarrow Y$ (using historical data X, and corresponding labels Y) is an example of: [4]

i. Unsupervised learning
ii. Training
iii. Supervised learning
iv. Testing

(b) For a classifier, $f(X^*|X;\theta) \rightarrow Y^*$, which values for $Y^*$ might be valid? [4]

i. Share price of a company on the stock market
ii. Banana, Apple, Car
iii. 1, 2, 3
iv. Any range of real numbers

(c) k-Nearest Neighbour is an example of: [4]

i. Supervised classification
ii. Unsupervised learning
iii. Logistic regression
iv. None of the above

(d) Adding regularization to an optimisation function can be used to: [4]

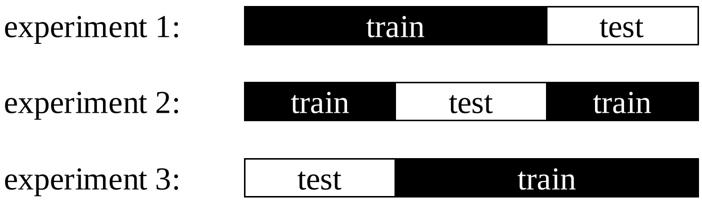
i. Increase complexity
ii. Penalise complexity
iii. Control overfitting
iv. None of the above

(e) The following is(are) example(s) of unsupervised machine learning: [4]

i. Clustering
ii. Dimensionality reduction
iii. Logistic regression
iv. None of the above

**Question 2**: Cross validation.

(a) *N*-fold cross validation is applied to a dataset by splitting the data according to the following scheme:




i. How many folds ($N$) are there here? [1]

ii. If $e_i$ is the error from experiment $i$, what is the equation for estimating the total error over all folds? [4]

iii. If nested cross validation is going to be used, what is the name of the additional data-set split that will be needed? [1]

(b) What problem does cross-validation reduce when performing model or parameter selection? [2]

**Question 3**: Evaluation.

(a) The confusion matrix below records the number of true / false negatives (TN/FN), and true/false positives (TP/FP) returned by a binary classifier.

| | | Prediction 0 | Prediction 1 |
|---|---|---|---|
| Ground Truth | 0 | TN | FP |
| | 1 | FN | TP |

i. Write the equation for overall accuracy. [2]

ii. Write the equation for recall. [2]

iii. Write the equation for precision. [2]

(b) A company claims its new fitness device, RunTrack, can automatically recognise whether you are Standing (S), Walking (W), or Running (R). The following sequence is used to test the system:

Ground Truth = [S,W,R,W,S]

This results in the following output:

Prediction = [S,S,W,W,S]

i. How many classes does RunTrack claim to recognise? [1]

ii. Draw the confusion matrix for this result. [3]

iii. Calculate the overall accuracy of RunTrack on this test set. [2]

# Part B

**Answer two questions only.**

**Question 4**: Linear regression, optimization, regularization.

(a) Given two known data points, $(x_{1,}y_1) = (2,4)$ and $(x_{2,}y_2) = (3,8)$

i. Draw a graph showing this data in 2D space. [3]

ii. Assuming a linear model, $\boldsymbol{y} = \theta_0 + \theta_1\boldsymbol{x}$, use the data to solve for parameters $\theta$. Highlight which value is the gradient, and which is the y-intercept. [5]

iii. What value for $y$ would this model predict if $x = 4$ was observed? [2]

iv. A new data point is observed at $(x_{3,}y_3) = (1,1)$. Does this new data fit our model? Would you expect such data in a real world system? Briefly justify your answer. [5]

(b) You are given an estimated regression hypothesis $h(x;\theta) = 2x^2$.

i. Given $\theta = [\theta_0, \theta_1, \theta_2]^T$, and $\boldsymbol{x} = [1, x, x^2]^T$, what are the values of $\theta_0$, $\theta_1$, and $\theta_2$ for the above hypothesis? [2]

ii. What type of equation is $h$? What is its degree? [2]

iii. Given (x,y) data observed at (1,1), (2,4), and (3,8), calculate the loss for the hypothesis function using the mean squared loss function (where m is size of the dataset): [5]

$$J(\theta) = \frac{1}{2M}\sum_{i=1}^{m}\left(h\left(x^{(i)};\theta\right) - y^{(i)}\right)^2$$

*Note: You can give your final answer as a fraction*

iv. The cost function $J(\theta)$ is used as part of the Gradient Descent algorithm to optimise the parameters, $\theta = [\theta_0, \theta_1, \theta_2]^T$. If we wanted to apply regularization to that optimisation, what changes would be needed to the cost function? Write an equation for a regularized version of $J(\theta)$. [3]

v. During optimisation, what effect does increasing the regularization parameter have on model complexity? How might this help the final model? [3]

**Question 5**: Logistic regression, linear regression, scaling

(a) A hypothesis function is defined as: $h(\boldsymbol{x};\theta) = g(\theta^T\boldsymbol{x})$
with $g(z) = \frac{1}{1+e^{-z}}$, input features $x$, and parameters $\theta$.

i. What is the function $g(z)$ commonly known as? [2]

ii. What is the range of its output? [2]

iii. What is the value of the hypothesis function when $\theta^T\boldsymbol{x} = 0$? [3]

iv. We want a classifier that outputs $y = 1$ when $(\theta^T\boldsymbol{x}) \geq 0$, and $y = 0$ otherwise. [5]
(1) How is this decision made using the hypothesis function?
(2) What does the value of the hypothesis function represent?

(b) You are asked to design a *logistic regression* classifier that detects whether there is a storm coming or not. You have data on the following:

y: {storm=1, not storm=0}
$x_1$: temperature
$x_2$: precipitation
$x_3$: air pressure

Write out, in full, a suitable logistic regression hypothesis for this problem in terms of features $x_i$ and coefficients $\theta_i$. Highlight any special cases of values for $x_i$. [5]

(c) A used car dealership asks you to build a website that lets people evaluate how much an old car might be worth. You have access to a database of cars with the following information:

y: an expert's opinion on the car's worth (in £)
$x_1$: number of km driven
$x_2$: the number of doors
$x_3$: its age (in months)

i. Write an equation for a suitable *linear regression* hypothesis function, $h(X; \theta)$, that can be trained on this database ($X = \{x_1, x_2, x_3\}$) to estimate a car's worth, y. [5]

ii. Briefly discuss why you might we want to scale the data before applying gradient descent on this function? [3]

iii. The known ranges and average values for each of the input features in the database are:

$x_1 = [0;\ 1{,}000{,}000]$ , $\overline{x_1}$=100,000
$x_2 = [0;\ 10]$ , $\overline{x_2}$=3
$x_3 = [0;\ 1{,}000]$ , $\overline{x_3}$=50

(1) Given this information, what would be an appropriate formula for scaling the data to within the range [-1, 1]? [5]
(2) Use this formula to scale the following data (*fractions will suffice*): $x_1 = 100{,}000\ km$, $x_2 = 4$ doors, $x_3 = 40$ months.

**Question 6**: Bayesian modelling

(a) In the following, we assume a binary classifier with features X, and class labels, Y = {0,1}.

i. P(Y=y) represents the probability that random variable Y takes on the value y. What does P(Y=y | X=x) represent? [3]

ii. If P(Y=1) =0.2, what is the value of P(Y=0)? Why? [2]

iii. What is the name given to describe the type of classifier that aims, like logistic regression, to learn P(Y | X) directly? [2]

iv. A generative classifier models the joint probability of a class and its associated feature data. Write this probability in terms of X and Y. [2]

(b) Bayes' theorem is defined as:

$$\mathrm{P(Y|X)} = \frac{\mathrm{P(X|Y)P(Y)}}{P(X)}$$

i. Which terms are referred to as i) the prior, ii) the posterior? [4]

ii. Use the product rule to show how Bayes' theorem can be derived from P(X,Y) [4]

(c) A model has been trained to help classify whether a person has a cold or not $Y = \{cold, not\}$. Given a new observation $X = symptom\ x$ (coughing), the following conditional probabilities are obtained:

$$P(x|cold) = 0.3$$
$$P(x|not) = 0.1$$

We assume that

$$P(cold) = P(not) = 0.5$$

Use this information and Bayes' theorem to help answer the following:

i. Calculate the probability of a person having a cold given the observed symptom, $x$, i.e. $P(cold|x)$ ? (hint: $P(X) = P(X|cold)P(cold) + P(X|not)P(not)$) [5]

ii. A comprehensive study reveals that actually 1 in 10 people typically have a cold at any one time, what does this fact do to the result above? Recalculate $P(cold|x)$. [4]

iii. Briefly describe how we might use Bayes' Theorem to build a classifier that estimates the most probable illness, given symptom $x$, from the following:
$Y = \{$cold, flu, hayfever$\}$? [4]