UNIVERSITY OF LONDON

GOLDSMITHS COLLEGE

Department of Computing

B. Sc. Examination Spring 2019-20

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 –questions involving calculation do not require exact answers (fractions and approximations will be accepted).

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 full credit.

Question 1: General Questions.

(a) The function, $f(X) \rightarrow Y,$ is created using only historical data X and some measure of similarity. This is an example of: [4]

i. Unsupervised learning
ii. Supervised learning
iii. Clustering
iv. Linear regression

(b) Which of the following data labels are most suitable for a regression task? [4]

i. Yellow, Red, Blue
ii. Temperature in degrees Celsius
iii. Time in milliseconds
iv. Trees, Lorries, Trucks, Cars

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

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

(d) We can use cross validation when evaluating a classifier to: [4]

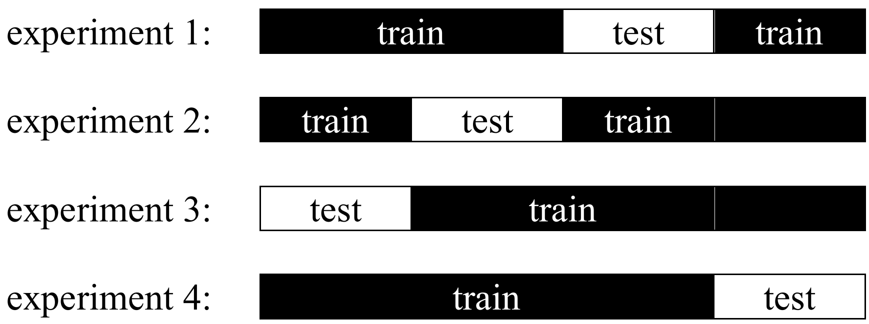
i. Ensure separation of training and test data
ii. Ensure 100% classifier accuracy
iii. Estimate classifier response to unknown data
iv. None of the above

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

i. Linear regression
ii. Logistic regression
iii. Principle component analysis
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. What does N represent, and what number is it in this example? [2]

ii. If e_i is the error from experiment i, what is the general equation for estimating overall mean error? [3]

iii. Given experiment errors of: $e_1 = 0.4, e_2 = 0.8, e_3 = 0.5$, and $e_4 = 0.3$, calculate the mean error? [2]

iv. What type of cross-validation might we be using if we add an additional validation split? [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. [1]

ii. Write the equation for precision. [1]

iii. What is the name of the performance metric given by $\frac{TP}{TP+FN}$? [1]

(b) A multi-class classifier is evaluated. From a sequence labelled as:

Ground Truth = [1, 2, 1, 2, 2, 3, 1, 2, 2]

It outputs the following:

Prediction = [1, 1, 2, 2, 1, 1, 2, 2, 1]

i. Draw a fully labelled confusion matrix for this result. Label rows as actual ground truth, and columns as prediction output. Show the ground truth totals. [4]

ii. Calculate the accuracy (class recall) for each class. [3]

Part B

Answer two questions only.

Question 4: Conditional probability, Naïve Bayes

(a) Given probabilities P(A) and P(B), with variables A, B:

i. What is meant by the term P(A, B)? [2]

ii. What is meant by the term P(A | B)? [2]

iii. If B is a binary random variable, i.e. B = {0,1}, and the probability P(B=0) = 0.7, what is the value of P(B=1)? Show your working. [3]

iv. Use the product rule to write out two equivalent formulas for P(A, B) in terms of conditional probability. [4]

v. Given P(B=0) = 0.7 and P(A=0 | B=0) = 0.1, what is the probability of both A and B being zero? [3]

(b) Given Y represents a class, and X represents observed features, we can specify the following terms: the prior probability of class Y, P(Y), the likelihood, P(X|Y), and the marginal observation probability P(X).

i. Use the above terms to write out the equation for the posterior probability, P(Y|X). [3]

ii. What is the name of this equation? [2]

(c) You are asked to build a Naïve Bayes classifier.

i. What is the "naïve" assumption in Naïve Bayes? [3]

ii. Given class variable, Y, and multiple feature variables X_1, X_2, and X_3, write out the Naïve Bayes version of Bayes' theorem, $P(Y \mid X_1, X_2, X_3) = \ldots$ [4]

iii. Your NB classifier is designed to recognise 3 different classes of animal, Y={dog, cat, goat}, given observations, X. Briefly describe how you would build such a classifier and show how the classifier would decide on the class of a new observation, X=*x*. [4]

Question 5: Linear regression, optimization, regularization.

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

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 θ. Show your working. [4]

iii. Which parameter is the gradient, and which is the y-intercept? [2]

iv. Calculate the point where the line intercepts with the x-axis? Show your working (you can use *fractions*). [2]

v. A new data point is observed at $(x_{3,}y_3) = (1, 4\,)$. Does this new data fit our model? Would you expect such data in a real-world system? Briefly justify your answer. [4]

(b) You are given a regression hypothesis $h(x;\theta) = 3x^2 + 2$.

i. Given $\theta = [\theta_0, \theta_1, \theta_2]^T$, and $\boldsymbol{x} = [1, x, x^2]^T$, what are the values of θ_0, θ_1, and θ_2 for the above hypothesis? [2]

ii. Given (x,y) data observed at (-2,15) , (-1,4), (1,5), and (2,13) calculate the loss for the hypothesis function using the mean squared loss function (*as a fraction*). Show all your working. [5]

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

iii. What do we need to do to regularize the above loss function, $J(\theta)$? Write the full equation for an L2 regularized loss function $J'(\theta)$. Identify and name any [5] new (hyper)parameters.

iv. During optimisation, describe the effect of changing the regularization parameter on model complexity. [3]

Question 6: Logistic regression, scaling

(a) A logistic regression 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 θ.

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

ii. Draw a labelled graph of $h(\boldsymbol{x};\theta)$, with the x-axis as $\theta^T \boldsymbol{x}$. [5]
Make sure to indicate the range of $h(\boldsymbol{x};\theta)$, and indicate any other significant points, like where $h(\boldsymbol{x};\theta) = 0.5$.

iii. When trained as part of a logistic regression classifier (i.e. that outputs a class as y=1 or y=0), what can the numeric value of $h(\boldsymbol{x};\theta)$ be used to represent? [2]

iv. Briefly describe how you might build a binary classifier using logistic regression. Specify how an output decision of y=1 or y=0 might be made, with reference to $h(\boldsymbol{x};\theta)$. [4]

v. If Naïve Bayes can be described as a Generative classifier, what is the name used to describe classifiers like Logistic Regression? Briefly explain the difference between the two types. [4]

(b) You are asked to design a *logistic regression* classifier that detects whether there may be ice on the road or not. You have the following data:

y: {icy=1, not icy=0}
x_1: ground temperature (degrees Celsius)
x_2: precipitation (mm)

i. Write out, in full, a suitable logistic regression hypothesis, $h(\boldsymbol{x};\theta)$, for this problem in terms of features x_j and coefficients θ_j. Highlight any special cases of values for x_j. [4]

ii. The logistic regression parameters are optimized using Batch Gradient Descent. The following equation defines the update rule: [6]

$$\theta_j^{new} = \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h(\boldsymbol{x};\theta) - y^{(i)}\right) x_j^{(i)}$$

(1) For the icy road example, what are the possible values of *j* (highlight any special cases)?
(2) What does *m* represent?
(3) What does α represent?
(4) Briefly describe the effects on algorithm convergence if α is too low or too high?

iii. What does scaling do to the data, and what advantages are there to scaling a dataset before using Gradient Descent? [3]

END OF EXAMINATION