UNIVERSITY OF BRISTOL

January 2022 (Mock exam) Examination Period

Department of Computer Science

3rd Year Examination for the Degrees of
Bachelor in Computer Science
Master of Engineering in Computer Science

COMS30033
Machine Learning (Mock exam)

TIME ALLOWED:
2 Hours
plus 30 minutes to allow for collation and uploading of answers.

This paper contains **fifteen** questions.
All questions will be marked.
If you attempt a question and do not wish it to be marked, delete it clearly.
The maximum for this paper is **100 marks**.

Other Instructions

1. **THIS IS A MOCK EXAM!!**
2. **Instruction 1: The exam is divided into two parts (Part 1 and Part 2). The first contains 10 short questions worth 5 marks each and the second 5 long questions worth 10 marks each. Both parts cover the full material taught in the unit.**
3. **Instruction 2: Note that sharing information with colleagues is strictly forbidden and that we have a set of measures in place to identify cases of plagiarism.**
4. **Instruction 3: This is NOT a open book exam.**

Part 1: Short questions (5 marks each)

Question 1 (5 marks)
Which form(s) of learning requires a explicit target?

Question 2 (5 marks)
What is the key difference between parametric and non-parametric models?

Question 3 (5 marks)
Figure 1 shows a Bayesian network structure (i.e. directed acyclic graph). Write down all pairs of variables which are independent conditional on F.

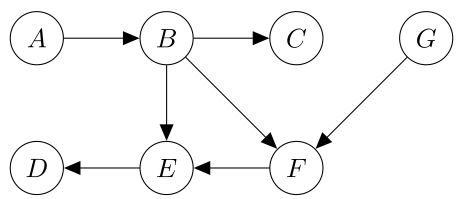

Figure 1: Directed acyclic graph for Question 3

Question 4 (5 marks)
Let $(1, 5, 2)$ and $(1, 3, 1)$ be the first two principal components computed from some dataset. Let $x_1 = (2, 6, 0)$ and $x_2 = (-6, 1, 2)$ be two datapoints. Compute a 2-d approximation for each of x_1 and x_2 using the two principal components.

Question 5 (5 marks)
Suppose you were using EM to estimate the parameters of a mixture of 3 Gaussians. Let x_i be a training datapoint where $\mathcal{N}(x_i|\mu_1, \Sigma_1) = 0.2$, $\mathcal{N}(x_i|\mu_2, \Sigma_2) = 0.1$, $\mathcal{N}(x_i|\mu_3, \Sigma_3) = 0.4$. μ_j and Σ_j are the current parameter values for the jth Gaussian. Let $\gamma_{i1} = 0.4$, $\gamma_{i2} = 0.2$, $\gamma_{i3} = 0.4$ be the current *responsibilites* for x_i. What are the current values for the mixing coefficients.

Question 6 (5 marks)
When using MCMC what is *burn-in* (2 marks) and we do we use it (3 marks)?

Question 7 (5 marks)

(a) Must kernel function always return non-negative values? If yes, explain why. If no, find values of x, y and k such that $k(x, y) < 0$, and where k is a valid kernel function.

(b) Explain why kernels must symmetric: i.e. $k(x, y) = k(y, x)$.

Question 8 (5 marks)
This question is about decision trees. We have a dataset containing three types of clothing with three features:

Data point index	X	Y	Z	Class Label
1	0	0	0	+1
2	1	0	1	-1
3	1	1	0	-1
4	1	1	1	+1

(a) What is the minimum depth of tree that would give zero training error? Explain your reasoning.

(b) In general, are there any problems that might arise if we grow a decision tree until there is zero training error? Explain your reasoning and give some ways to avoid this problem with CART trees.

Question 9 (5 marks)

Explain in words how the Adaboost method applies weights to training data instances, and why it does this?

Question 10 (5 marks)

Labelled data is important for supervised machine learning and evaluating machine learning systems. A common solution for obtaining large labelled datasets is crowdsourcing. Briefly explain two disadvantages or challenges of using crowdsourcing to obtain labelled data.

Part 2: Long questions (10 marks each)

Question 11 (10 marks)

(a) Which of the following models overfit, underfit and provide an adequate fit to the data? (3 marks)

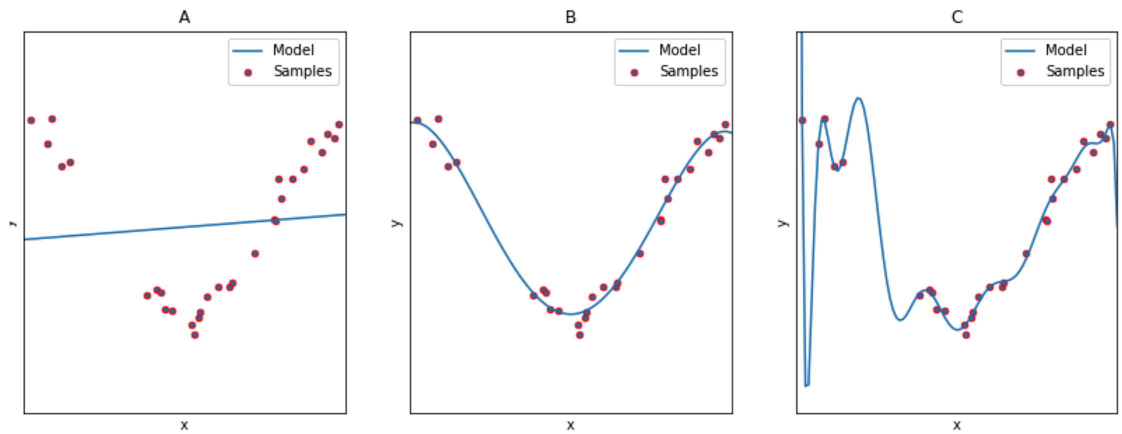

Figure 2: Examples of data (red scatter plot) with three different models (A,B,C; blue solid line).

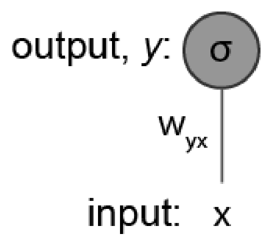

Figure 3: Schematic of a simple artificial neural network.

(b) Given a very simple neural network with only one input x and one output neuron y with a linear activation f function, where $x = 0.5$, the target $t = 1$ and a standard mean squared error ($E = \frac{1}{2}(y - x)^2$) what is the exact value for Δw_{yx} ? Assume that the current $w_{yx} = -0.5$ and that there are no bias. Your answer only needs to be approximate (e.g. 0.06). You should use the chain rule as discussed in the lectures. (7 marks)

Question 12 (10 marks)

(a) Consider running k-means on the following datapoints: $x_1 = (0, -1)$, $x_2 = (1, 2)$, $x_3 = (1, 1)$, $x_4 = (2, 1)$. Suppose $k = 2$ and assume that initially x_1 and x_2 are assigned to cluster 1 and x_3 and x_4 are assigned to cluster 2. After running one iteration of the k-means algorithm to which cluster are the datapoints assigned? (5 marks)

(b) Give two advantages of using Gaussian mixtures over k-means for clustering and one disadvantage. (5 marks)

Question 13 (10 marks)

(a) Explain what *slack variables* in support vector machines (SVMs) are. (5 marks)

(b) Is it correct to call SVMs a *nonparametric* method? Explain your answer. (5 marks)

Question 14 (10 marks)

This question is about hidden Markov models (HMM). Consider a hidden Markov model (HMM) with the following transition matrix and initial state probability estimates:

$$A = p(z_{n+1}|z_n) = \begin{matrix} 0.3 & 0.7 \\ 0.2 & 0.8 \end{matrix} \tag{1}$$

$$\pi = p(z_1) = [0.3, 0.7] \tag{2}$$

Rows in A correspond to values of z_n and columns to values of z_{n+1}.

The observations are discrete and can take values X, Y or Z. The model also has the following emission probabilities for the observations:

$$p(x_n = i|z_n = j) = \begin{matrix} x_n = i & z_n = 1 & z_n = 2 \\ \text{X} & 0.4 & 0.1 \\ \text{Y} & 0.1 & 0.5 \\ \text{Z} & 0.5 & 0.4 \end{matrix}$$

We observe the sequence X, Y.

(a) Use the parameters to compute the probability distribution $p(x_1 = X, x_2 = Y, z_2 = 1)$.

(b) Suppose we want to compute the probability that the next state is $z_3 = 1$. Briefly state explain the first order Markov assumption and how it applies to this computation.

Question 15 (10 marks)

This question is about linear dynamical systems. Suppose you are using a linear dynamical system to predict a continuous state variable, z_n. The model parameters have already been learned using expectation maximisation. For a new time-step, n, you have a Gaussian prior over z_n with mean 0 and variance 100. You observe a noisy sensor measurement, x_n, then use it to obtain a posterior distribution over z_n. The observation $x_n = 10$ with noise variance 1.

(a) What kind of distribution is the posterior distribution over z_n and which method can you use to compute it?

(b) Will the posterior mean of z_n be closer to 0 or 10 and why?

(c) If we now observe x_{n+1} and want to update the posterior over z_n, what method do we need to use and why?

END OF PAPER