UNIVERSITY COLLEGE LONDON

EXAMINATION FOR INTERNAL STUDENTS

MODULE CODE : **COMP0081**

ASSESSMENT PATTERN : **COMP0081A7UA; COMP0081A7PA**

MODULE NAME : **Applied Machine Learning**

LEVEL: : **Undergraduate Masters; Postgraduate**

DATE : **30 April 2019**

TIME : **14:30**

TIME ALLOWED : **2 hrs 30 mins**

This paper is suitable for candidates who attended classes for this module in the following academic year(s):

Year
2018-19

EXAMINATION PAPER CANNOT BE REMOVED FROM THE EXAM HALL. PLACE EXAM PAPER AND ALL COMPLETED SCRIPTS INSIDE THE EXAMINATION ENVELOPE

Hall Instructions	.
Standard Calculators	N
Non-Standard Calculators	N

TURN OVER

EXAM PAPER ERRATA

2018/2019 Examination Period

Note to Candidate:
You have been provided with this errata as there is an amendment to the printed question paper. This errata replaces the question that is printed in the main question paper.

Do not turn this paper over until the examination starts.

Read this paper before the examination paper.

Module	COMP0081 Applied Machine Learning
Assessment Pattern	COMP0081A7UA; COMP0081A7PA
Date of Exam	30 April 2019
Time of Exam	2:30 PM

Question:

Question 4, all parts.

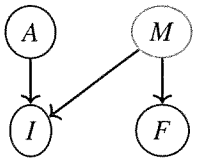
Revised Question:

4.

Consider an hypothetical music-degree scenario where individuals are tested for music aptitude at the beginning and at the end of the degree. Assume that this scenario can be expressed by the Bayesian network in the figure with associated model

$$I = \alpha A + \beta M,$$
$$F = \gamma M.$$

with $\mathbb{E}(A) = 0$, $Var(A) = 1$, $p(M) = \mathcal{N}(0,1)$. The variable A represents gender, M represents music aptitude and is unobserved, I represents initial test score, and F represents the degree final test score. The terms α, β, γ are unknown scalar parameters of the model.



- Individuals with higher music aptitude M are more likely
  - To obtain higher final-test score F ($M \to F$).
  - To obtain higher initial-test score I ($M \to I$).
- Due to discrimination women are assigned a lower initial-test score than men for the same aptitude level ($A \to I$).
- Gender does not influence music aptitude M ($A \not\to M$) nor the final-test score F

Question 4 cont. over page

Question 4 cont.

a. Is F independent of A ($F \perp\!\!\!\perp A$) in this model? Is F independent of A given I ($F \perp\!\!\!\perp A|I$)?

[5 marks]

b. Would a least-squares predictor of F of the form $\hat{F} = \theta_I I$, that is minimising $\mathbb{E}\left[(F - \theta_I I)^2\right]$, be independent of A? Justify the answer mathematically.

[5 marks]

c. Give an extension of the linear prediction above that will result in a predictor $\hat{F}$ that is independent of A.

[5 marks]

d. If the final-test score is also biased against women $F = \gamma M + \eta A$, would the predictor $\hat{F}$ in (c) be independent of A?

[5 marks]

[Total:20 marks]

Applied Machine Learning, COMP0081 (A7U, A7P)

Main Summer Examination Period, 2018/19

Answer ALL FOUR questions.

Marks for each part of each question are indicated in square brackets

Calculators are NOT permitted

1. a. Suppose that you take part in a Kaggle competition where the task is binary classification and the quality of predictions is measured by accuracy. You build a sophisticated Neural Network model. Is it possible to directly optimise accuracy using gradient descent? Support your arguments.

[5 marks]

b. Suppose that in another competition the quality of predictions is evaluated using average log-loss. You suspect that the train-test split is such that the fraction of positive examples in the training set is significantly different from that in (the public part of) the test set.

i. How can you verify this suspicion of yours?

[5 marks]

ii. Suppose that it is indeed the case. What would that mean for your approach to validation?

[5 marks]

c. Suppose that you are designing an image recognition competition. Your dataset consists of satellite images showing various cities and towns. The goal is to predict whether a given image contains a certain type of roads (gravel tracks, small roads, motorways, etc.). How would you split your dataset into train and test part?

[5 marks]

[Total: 20 marks]

2. a. In t-SNE the joint distribution of pairs of points in the high-dimensional space is obtained by symmetrising the conditional probabilities: $p_{ij} = \frac{p_{i|j}+p_{j|i}}{2N}$, where

$$p_{j|i} = \frac{\exp(-\left\|x_i - x_j\right\|^2/2\sigma_i)}{\sum_{k\neq i}\exp(-\left\|x_i - x_k\right\|^2/2\sigma_i)}$$

The alternative suggestion would be to define it as

$$p'_{ij} = \frac{\exp(-\left\|x_i - x_j\right\|^2/2\sigma)}{\sum_{k\neq l}\exp(-\left\|x_k - x_l\right\|^2/2\sigma)}$$

Suppose that there is an outlier in the dataset. Which choice of the joint distribution would ensure that this point is not ignored and why?

[5 marks]

b. The points x_i are mapped to the points y_i in a low-dimensional space. The joint pairwise probabilities in the low-dimensional space are defined as Student t-distribution with one degree of freedom:

$$q_{ij} = \frac{\left(1+\left\|y_i - y_j\right\|^2\right)^{-1}}{\sum_{k\neq l}\left(1+\left\|y_k - y_l\right\|^2\right)^{-1}}$$

Explain the intuition for using the heavy-tailed distribution in the low-dimensional space.

[5 marks]

c. The objective of t-SNE is to minimise the KL-divergence $C(Y) = KL(P||Q) = \sum_{ij} p_{ij}\log\frac{p_{ij}}{q_{ij}}$ as a function of the points y_i in low-dimensional space. The components of the gradient of this cost function are

$$\frac{\partial C}{\partial y_i} = 4\sum_j (p_{ij} - q_{ij})(y_i - y_j)(1+\left\|y_i - y_j\right\|^2)^{-1}$$

What is the complexity of computing the gradient?

[5 marks]

d. In the original t-SNE paper, this computation cost restricted the size of the datasets to several thousands. Describe the random walk t-SNE approach that allows to use the information from the whole dataset to visualise the subset of 'landmark' points.

[5 marks]

e. Describe the idea behind Barnes-Hut approximation, the structures used in it and how it helps to speed up the t-SNE computation.

[10 marks]

[Total 30 marks]

3. a. Explain what the problem of low-rank matrix factorization is and give the examples of its applications.

[2 marks]

b. Explain what singular-value decomposition (SVD) is and how you can use it to solve the low-rank matrix factorization problem.

[3 marks]

c. Explain the relation between Principal Component Analysis and SVD.

[3 marks]

d. Write what the drawbacks of using SVD for the problem of topic modeling are.

[3 marks]

e. This question goes through some of the details of Non-Negative Matrix Factorisation problem.

i. Explain what the problem of Non-Negative Matrix Factorisation is. Is the solution to this problem unique? If yes, justify your answer, if no, provide a counter-example.

[3 marks]

ii. Describe what standard algorithms you know to solve the Non-Negative Matrix Factorisation problem in general case.

[3 marks]

iii. Explain what the 'anchor words' (or separability) assumption is in terms of matrix decomposition and what it means in terms of topic modeling.

[3 marks]

iv. Show that if the "anchor words" assumption holds for matrix M, there exists an efficient algorithm for solving non-negative matrix factorization problem. Describe the algorithm.

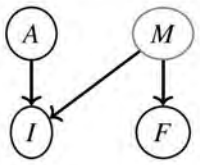
[10 marks]

[Total: 30 marks]

4. Consider an hypothetical music-degree scenario where individuals are tested for music aptitude at the beginning and at the end of the degree. Assume that this scenario can be expressed by the Bayesian network in the figure with associated model

$$E = \alpha A + \beta M\,,$$
$$F = \gamma M\,.$$

with $Var(A) = 1$, $p(M) = \mathcal{N}(0,1)$. The variable A represents gender, M represents music aptitude and is unobserved, I represents initial test score, and S represents the degree final test score. The terms α, β, γ are unknown scalar parameters of the model.



- Individuals with higher music aptitude M are more likely
  - To obtain higher final-test score F ($M \to F$).
  - To obtain higher initial-test score I ($M \to I$).
- Due to discrimination women are assigned a lower entrance-test score than men for the same aptitude level ($A \to I$).
- Gender does not influence music aptitude M ($A \not\to M$) nor the final tests score F

a. Is F independent of A ($F \perp\!\!\!\perp A$) in this model? Is F independent of A given I ($F \perp\!\!\!\perp A|E$)?

[5 marks]

b. Would a least-squares predictor of F of the form $\hat{F} = \theta_E E$ be independent of A? Justify the answer mathematically.

[5 marks]

c. Give an extension of the linear prediction above that will result in a predictor $\hat{F}$ that is independent of A.

[5 marks]

d. If the final-test score is also biased against women $F = \gamma M + \eta A$, would the predictor $\hat{F}$ in (c) be independent of A?

[5 marks]

[Total: 20 marks]