Advanced Algorithms COMP4121, Term 3 of 2022
FINAL EXAM QUESTIONS

Exam closes at 5:00pm. If you are a registered ELS student, your exam closes at 6:45pm.

Submit your answers by email to shuiqiao.yang@unsw.edu.au. You can send a scan or a photo but make sure your handwriting is clear. To email, use ONLY your UNSW email account. Please make sure your file is less than 5MB in size, otherwise it might not go through. Please follow the format "first name last name zID" to name the file of your submission.

If you have any questions email me and I will try to answer them as promptly as I can. Problems are worth as indicated. You can use freely the slides of the course. You must show your work and justify your answers. Good luck!

1. Suppose you have a Gaussian random variable X with zero mean ($E[X] = \mu = 0$) and variance $V[X] = v$, its probability density function is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi v}} e^{-\frac{x^2}{2v}}.$$

   Assume you use such X to generate independently the coordinates $(x_1, \ldots, x_d)$ of a random vector $\vec{x}$ from $\mathbb{R}^d$. Please find the value of v such that the expected value of the length of $\vec{x}$ is equal to 1. **(15 points)**

2. You are asked to use QuickSort to sort an array containing the birth month of the students from COMP4121. Since there are only 12 possible months, you can expect that each month will show up many times in the array. We thus have an array with many duplicate values. Except for using a Randomised QuickSort, can you figure out some other strategies to accelerate the sort? **(15 points)**

3. Suppose you have a simple dataset contains some data points in 2D Euclidean space as shown in Table 1. Now, you are asked to use K-means algorithm to find the clusters in the dataset.

Table 1: A simple Dataset.

Points	x	y
P1	0.5	0.5
P2	0.5	1.5
P3	1	1
P4	3	3
P5	3.5	4
P6	4	3
P7	4	4

   (a) If the initial cluster centers are set as **P2** and **P5**. Please find the final coordinates of the cluster centers and the members in the corresponding clusters. **(10 points)**

   (b) Please find two types of 2-dimensional data distributions where K-means may fail to find accurate clusters and explain the reasons, you can sketch the data distribution in 2D Euclidean space to help you explain. **(10 points)**

4. You are asked to design a basic latent factor method for movie recommendation. You need to find two latent factor matrices $F \in \mathbb{R}^{m \times k}$ and $E \in \mathbb{R}^{n \times k}$ that can satisfy the following equation:

$$r_{ij} = f_i \cdot e_j,$$

   where f_i is the i-th row of F, e_j is the j-th row of E and r_{ij} is the existing rating for user j on movie i. Suppose you are provided with an incomplete ratings matrix R (Table 2) and two partially completed latent factor matrices F and E:

Table 2: Rating matrix.

	User1	User2	User3	User4
Movie1		2	3	
Movie2	2			2
Movie3		4		3
Movie4	3		4.5	

$$F = \begin{bmatrix} 1 & 2 \\ - & 1 \\ 3 & - \\ 3 & - \end{bmatrix}$$

$$E = \begin{bmatrix} 1 & - \\ 1 & - \\ - & 1 \\ 1 & - \end{bmatrix}$$

(a) Find the missing entries (denoted by $-$) for the matrices of F and E such that $r_{ij} = f_i \cdot e_j$ can be satisfied. **(10 points)**

(b) After you have the completed latent factor matrices of F and E, you can predict the ratings. Please predict the unobserved ratings of **User2** to Movie2 and Movie4. **(10 points)**.

(c) Based on the latent factor matrices F and E, can you find two users who have the most similar preferences for movies and two movies which have the most similar features? **(5 points)**

5. On a small island lives a type of animal which appears either on a tree or in a nest. The behavior of the animals is closely related to the weather on the island. On the island, the weather has been either sunny or rainy over the years. We noticed that on average 3 out of every 4 sunny days, the animals would stay on the tree. 2 out of every 4 rainy days, the animals stay in their nest. We also noticed that on average 2 out of every 4 sunny days are followed by another sunny day and that 1 out of every 3 rainy days is followed by another rainy day.

(a) Estimate the fraction of sunny and rainy days on the island. **(5 points)**

(b) For a given sequence of observations of an animal, you can use Hidden Markov Model (HMM) to infer the hidden states that are most likely to generate the observations. Please show the emission probabilities of your HMM. **(5 points)**

(c) You observed one animal for 4 consecutive days and had the following observed behavior for that animal: stayed on a tree (day 1), rested in a nest (day 2), rested in a nest (day 3) and stayed on a tree (day 4). Find the actual weather of the 4 consecutive days that was most likely to cause the sequence of your observations. **(15 points)**