4. Recall that the L_∞ distance $d_\infty(u,v)$ between two 2-dimensional vectors $u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ and $v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ is given by $d_\infty(u,v) = \max(|u_1 - v_1|, |u_2 - v_2|)$.

Let $X = \{x_1, x_2, x_3, x_4, x_5, x_6\}$ and $C = \{c_1, c_2\}$, where:

$$x_1 = \begin{bmatrix} 4 \\ 2 \end{bmatrix}, x_2 = \begin{bmatrix} -1 \\ 2 \end{bmatrix}, x_3 = \begin{bmatrix} -2 \\ 1 \end{bmatrix}, x_4 = \begin{bmatrix} 3 \\ 1 \end{bmatrix}, x_5 = \begin{bmatrix} -2 \\ 2 \end{bmatrix}, x_6 = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$$

$$c_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, c_2 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

(i) Calculate the new values $\bar{C} = \{\bar{c}_1, \bar{c}_2\}$ of $\{c_1, c_2\}$ after one iteration of the k-means clustering algorithm (k=2). Use the L_∞ distance measure. Show all of your calculations. **[5]**

(ii) Suppose that instead of the original values, c_1, c_2 are given by $c_1 = \begin{bmatrix} 5 \\ -4 \end{bmatrix}, c_2 = \begin{bmatrix} 4 \\ 4 \end{bmatrix}$. What problem arises in the application of k-means clustering and how would you resolve it? **[5]**

3. (i) Consider 2-dimensional data being modelled using a Gaussian Mixture Model (GMM) with two mixture components. The parameters of the GMM components are: **[7]**

Gaussian 1: mean vector $\mu_1 = \begin{bmatrix} 4 \\ 6 \end{bmatrix}$, covariance matrix $\Sigma_1 = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$, and weight $w_1 = 0.7$

Gaussian 2: mean vector $\mu_2 = \begin{bmatrix} 7 \\ 4 \end{bmatrix}$, covariance matrix $\Sigma_2 = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$, and weight $w_2 = 0.3$

Calculate the probability of the sequence of data $y = \left(\begin{bmatrix} 5 \\ 6 \end{bmatrix}, \begin{bmatrix} 8 \\ 4 \end{bmatrix}\right)$ being generated by the GMM. Show your workings.

(ii) Suppose that a class C gives rise to 1-dimensional measurements that are uniformly distributed over the interval [0,2]. Unfortunately, we do not know this and we try to model the class distribution using a GMM. Suppose that the training set consists of 5 samples. How many GMM mixture components would lead to the best fit to the training set? Would such a model be appropriate to model data arising from class C? Explain your answer. **[5]**

5. Assume there are three baskets, each containing a certain number of apples and plums. Assume the following quantity of fruits:

Basket 1: 8 apples, 2 plums
Basket 2: 5 apples, 5 plums
Basket 3: 3 apples, 7 plums

At any point in time, you select a basket at random, and then a fruit from that basket (i.e., an apple or plum) and record your finding (A for apple and P for plum). You immediately replace the fruit so that the total number of apples and plums stays the same over time, and repeat the process. Unfortunately, you forgot to write down the baskets you chose and simply have an account of apples and plums.

(i) Explain how the sequence of fruits that is chosen can be modelled as the output of a hidden Markov model. Write down the parameters of the model. **[4]**

(ii) Compute the optimal sequence of baskets corresponding to the fruit sequence $y = (A, P, A, P, A)$. That is, which basket was each piece of fruit most likely to be selected from? Show your workings. **[7]**

2. (i) A topic spotting system is designed to detect newspaper articles on Brexit (topic T). The word "border" (word w) occurs 4280 times in 2100 documents that are judged to be on-topic, and 450 times in 2500 documents that are judged to be off-topic. The average number of words in all documents is 85, and approximately 8% of documents are on-topic. Calculate the usefulness and salience of the word "border" for topic T. Use the natural logarithm in your calculations. Show your calculations. **[6]**

(ii) Suppose that the number of on-topic articles increases, but the conditional probabilities of the word "border" occurring in documents that are on-topic and off-topic are unchanged. What are the limiting values of the usefulness and salience of the word "border" as the proportion of articles that are on-topic tends towards 100%? You should show your calculations. **[4]**

1. Two documents d_1 and d_2 are defined as follows after stop word removal and stemming: **[7]**

d_1 = "*project design network sampl tool monitor network traffic*"

d_2 = "*area network secur includ network sampl method*"

Using the Inverse Document Frequency (IDF) values in Table 1, calculate the TF-IDF similarity $sim(d_1, d_2)$ between d_1 and d_2. Show your workings.

Term *t*	$\boldsymbol{IDF(t)}$
area	2.4
design	1.7
includ	0.9
method	1.4
monitor	2.4
network	4.6
project	3.3
traffic	3.1
sampl	1.1
secur	4.1
tool	2.2

Table 1: IDF values.