Instructions: This assignment covers topics of classification.

Submission Requirements:

In *w4*, this part, you will be asked to walk through methods and calculations **manually** or display understanding of concepts and topics. In *a4-R* or *a4-python*, you will use packages and libraries to implement classification models, data reduction and text mining methods.

You must prepare your solution to this part as a document using LaTeX. I have provided a template for this assignment, where you can create your answers.

For this assignment, you are to work in your **groups**. I highly suggest using Overleaf to work on your submission together in your group.

Follow the submission template where work for each question must **start on a new page**. Do not put work for multiple problems on the same page.

Questions:

1. Representing Text
   Consider the following documents:

   Doc 1: pat sat on rat
   Doc 2: cat sat on mat
   Doc 3: cat was rat
   Doc 4: pat sat on rat mat
   Doc 5: pat was fat
   Doc 6: fat cat was rat

   (a) (12 points) Construct the Document-Term Matrix
   (b) (9 points) Construct the last threes rows of the TF-IDF Matrix (document-term form - `ltn` SMART Notation). Use $\log_{10}$ in calculation.

2. Text Classification

   Consider the problem of classifying the following documents:

Sample	Document	Class
1	`aries gemini gemini cancer libra`	1
2	`cancer leo leo libra pisces`	2
3	`aries gemini leo leo aquarius pisces`	2
4	`leo libra aquarius aries aries`	3
5	`aries libra aquarius`	3

   We are going to want to predict the class of a test document of "`gemini aries aquarius pisces`" using the Bernoulli and multinomial Naïve Bayes model.

   (a) (14 points) Assume that we use a Bernoulli (i.e., binary) Naive Bayes model. Compute the following conditional probabilities (use laplace smoothing). Assume you are providing the probability of when the token is present.
   Report as unreduced fraction.

Token	$\hat{P}(t\|c=1)$	$\hat{P}(t\|c=2)$	$\hat{P}(t\|c=3)$
aquarius			
aries			
cancer			
gemini			
leo			
libra			
pisces			

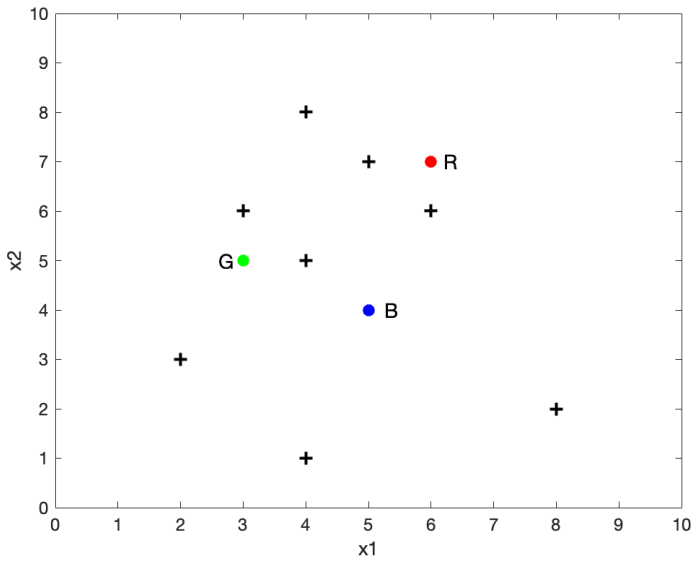
(b) (14 points) Assume that we use a multinomial NB model instead. Compute the following conditional probabilities: (use laplace smoothing).
Report as unreduced fraction.

Token	$\hat{P}(t\|c=1)$	$\hat{P}(t\|c=2)$	$\hat{P}(t\|c=3)$
aquarius			
aries			
cancer			
gemini			
leo			
libra			
pisces			

(c) (18 points) Compute and show the calculation for predicting the class of a test document of '`gemini aries aquarius pisces`' using the Bernoulli Naive Bayes model. Use Laplace smoothing for conditional probabilities.
Show the calculations, writing out the probabilities that go into making the prediction.

(d) (18 points) Compute and show the calculation for predicting the class of a test document of '`gemini aries aquarius pisces`' using the multinomial Naive Bayes model. Use Laplace smoothing for conditional probabilities.
Show the calculations, writing out the probabilities that go into making the prediction.

3. K-means Clustering
The plot below illustrates data (8 samples over 2 variables) that are being clustered using Kmeans (k=3) with Euclidean distance. The initial cluster centers are indicated with the points shown with filled in circles and labelled B-blue, G-green, and R-red.



(a) (16 points) Show the assignment of each sample to it's cluster. Report the distance of the sample to the centroid of group red ($d2cR$), the distance of the sample to the centroid of group blue ($d2cB$), and the distance of the sample to the centroid of group green ($d2cG$). Report the cluster labels (R, B, G) for each observation (what group that sample will be assigned for the next iteration).

(b) (6 points) Calculte the centroids for the next iteration. Report centers to 2 digits after the decimal point.

(c) (2 points) Will any points change their assignments in the next iteration? (YES / NO)

4. Hierarchical Clustering
   Suppose you have 6 samples (`1`, `2`, `3`, `4`, `5`, `6`), with a dissimilarity / distance matrix of:

	1	2	3	4	5	6
1	−−	0.3	0.4	0.35	0.6	0.5
2	0.3	−−	0.5	0.8	0.2	0.25
3	0.4	0.5	−−	0.45	0.4	0.9
4	0.35	0.8	0.45	−−	0.35	0.6
5	0.6	0.2	0.4	0.35	−−	0.4
6	0.5	0.25	0.9	0.6	0.4	−−

(a) (16 points) Trace hierarchical clustering **manually** with **complete linkage**.
For each step, say what are the next two items (samples or clusters) to be combined and report the new distance matrix (have samples & groups ordered numerically). You only need to report the lower triangular part of the distance matrix.

Example: Assume in step 1 samples 2 and 4 combine and assume in step 2, 1 and 5 combine.

Step 1. Combine **2** and **4**.
Report new distance matrix between groups, ordered numerically (e.g., if just combined `1` and `2` - then distance matrix should be ordered - `1`, `24`, `3`, `5`, `6`).

	1	24	3	5	6
1	0	-	-	-	-
24	?	0	-	-	-
3	?	?	0	-	-
5	?	?	?	0	-
6	?	?	?	?	0

Step 2. Combine **1** and **5**.

	15	24	3	6
15	0	-	-	-
24	?	0	-	-
3	?	?	0	-
6	?	?	?	0

(b) (3 points) Sketch the dendrogram resulting from Q4(a) by hand (scan or include a clear photo of this in your sumbmission). Be sure to include height estimates in the dendrogram from the distances.

(c) (16 points) Repeat Q4(a), with **single linkage**.

(d) (3 points) Sketch the dendrogram resulting from Q4(c). Estimate the heights in the dendrogram from the distances.

(e) (3 points) Use the dendrogram from (b) and (d), cut the dendrograms to form three clusters. Which samples are in each cluster?