Assignment 3

Deadline: Thursday, March 23, at 9:59am.

Submission: You need to submit the final PDF file, and any Python scripts, as a compressed folder (.zip or .tgz) on Quercus. If you used Google Colab, including a link is sufficient.

Neatness Point: One point will be given for neatness. You will receive this point as long as we don't have a hard time reading your solutions or understanding the structure of your code.

Late Submission: 10% of the marks will be deducted for each day late, up to a maximum of 3 days. After that, no submissions will be accepted.

Assignments are individual work. See the Course Information handout[1] for detailed policies.

1. **[1.5 pts] Revisiting single-cell RNA-seq**

   In this question, we will re-use the single-cell RNA-seq data in A2. You can download the data from the course website.

   (a) **[1 pts]** Run K-means and GMM with different K values [2,4,6,8] and report the ARI values for all of the 8 models [2].

   (b) **[0.5 pts]** In less than 3 sentences, comment on the differences between K-means and GMM in the above exercise in terms of speed and performance. (more than 3 sentence will not be considered)

2. **[1 pts] Hierarchical Clustering**

   The following matrix shows the distance between every pair of units that you want to cluster. Use single and complete link agglomerative clustering to group the data described by the following distance matrix. SHOW EVERY STEP OF THE COMPUTATIONS AND THE FINAL DENDROGRAMS.

	A	B	C	D
A	0	1	5	6
B		0	3	8
C			0	4
D				0

3. **[2 pts] Adaboost Implementation**

   Please complete this Colab notebook.

4. **[3 pts] PCA**

   For this question we will also use the RNA-seq dataset from Q1. You do not need to split the data.

[1] https://lmp1210-uoft.github.io/2023/assets/misc/syllabus.pdf

[2] you can use the function in https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_rand_score.html

(a) **[0.5 pt]** Run Principal Component Analysis (PCA) on your data with 15 components, and report their variance explained. Which component is most important for representing the data?

(b) **[0.5 pt]** Assuming that we want at least 80 percent of data variance to be captured by principal components, what is the minimum number of principal components that we need? You may get part of the mark if you explain how the answer is obtained.

(c) **[0.5 pt]** Make a scatter plot the largest principal component (y-axis) vs the second largest (x-axis). Colour the points based on their class label. Do you notice a pattern? Repeat this process with the bottom two principal components.

Note: see the Lecture 7 Colab book for hints

(d) **[0.5 pt]** Train a logistic regression classifier on the top two principal components, and add this line to your plot. Do the same for the bottom two principal components. Comment on the accuracy of each model.

(e) **[1 pt]** For k from 1 to 10, train a logistic regression model that uses the top k principal components (i.e. the first model uses the top 1, the second uses the top 2, and so on). Make a line plot of each model's accuracy, where the x-axis is the k value and the y-axis is the accuracy of the model that uses the top k principal components. What pattern do you observe?

5. **[3.5 pts] Multi-Omics Analysis for Cancer Subtyping** You are given a simulated multi-omic dataset for 832 cancer patients with 10 subtypes. The dataset consists of bulk RNA-seq (see file “A3RNAseq.csv”) and DNA Methylations (see file “A3Methylation.csv”). The ground truth can be found in file “label.csv”.

(a) **[1 pts]** Use three different visualization methods (PCA, tSNE, and UMAP) to project the RNA-seq data into a 2D space in which each dot represents a patient color-coded with the cancer subtypes. Further, run k-means with $k = 10$ and report the ARI value.

(b) **[1 pts]** Use three different visualization methods (PCA, tSNE, and UMAP) to project the Methylation data into a 2D space in which each dot represents a patient color-coded with the cancer subtypes. Further, run k-means with $k = 10$ and report the ARI value.

(c) **[1.5 pts]** Develop an unsupervised multi-modal machine learning algorithm which takes both RNA-seq and Methylation data and learns a latent representation of patients. Once the patient representations are learned, use UMAP to visualize them in 2-D space in which each dot represents a patient color-coded with the cancer subtypes. Further, run k-means with $k = 10$ and report the AMI value. The higher your final AMI is, the more marks you will get. **Note: this is an open-ended question. You can use any external package you can possibly find to solve this question. Please submit your code too with a Colab link so that we can verify the result.**

6. **[3 pts] Convolutional Neural Networks (CNN)** Please include detailed calculations in your solutions.

(a) **[1 pts]** What is the output size of a convolution layer, given an input RGB image of size 64x32x3 (Width*Height*Depth), with 12 filters of kernel size 6, stride 2 and padding 1?

(b) **[1 pts]** How many trainable parameters are in a 2D Convolution layer with input channel 5, output channel 6, and kernel size 3?

(c) [**1 pts**] What is the output size of a 2D pooling layer, given an input RGB image of size 128x128x3, with kernel size 8, stride 4?

7. [**1 pts**] **Obama Video** State the procedure (including potential model architectures) through which the Obama video in Lecture 1 was generated. An additional 2-point bonus will be given if you can generate the same video (please include the codes in the submission).