# Homework 2

## Machine Learning

Spring 2024

Due date: May 27, 2024

Upload your answers as a pdf file to your google drive directory. For the programming questions, in addition to your source code, include an example input and output, and include a short explanation of your code.

You can research your answers online or using textbooks, and you can discuss your solutions with your classmates; but you need to disclose all the resources that you used in your report. If you use tools like ChatGPT, include your prompt and the answer in your report.

1. **Comparative Clustering of Shape and S-Set Datasets:**

   **Datasets:**

   - From the “Shape Sets” (from here: https://cs.joensuu.fi/sipu/datasets/) download Pathbased, Spiral, Jain, and Flame
   - From the “S-Sets” download S1 and S4 datasets

   **For each dataset:**

   a. Cluster the data using the following methods. Use the same number of clusters as specified in the dataset. For each method, describe what distance or similarity metric you have used.

      i. K-means

      ii. Hierarchical clustering (average linkage)

      iii. Hierarchical clustering (single linkage)

      iv. Hierarchical clustering (complete linkage)

      v. Spectral clustering. Describe how you defined the graph.

   b. Visualize the resulting clustering (use the scatter plot of the data points and color the points by cluster assignment).

c. For datasets from the “Shape Set” where the true cluster is known, evaluate each clustering result using the purity index.

For each dataset, compare the results of each pair of clustering methods using the Rand index. Visualize the results in a 5x5 heatmap.

2. **PCA on MNIST dataset**

**Dataset:**

- Load MNIST dataset (could be accessed using from keras.datasets in python)
- Separate them by label into 10 smaller sets

**For each set:**

a. Flatten the pictures and apply PCA

b. Plot first PC vs. Second PC

c. Assume the points in this scatter plot are spread between $(x_1, x_2)$ and $(y_1, y_2)$ (which are the min and max of PC1 and PC2). Split this space into a 5x5 grid, and for each cell select a point that is closest to the center of that cell. Highlight these points in the scatter plot from the previous step.

d. Draw the original pictures corresponding to the 25 selecting points.
(See figure 14.23 of Element of Statistical Learning for an example).

3. **PCA and Clustering on gene expression data**

**Dataset:** Download the expression data from here.

The data is collected from 3 different SRA datasets and samples corresponding to two tissues (Leaf and Root) are combined into one expression matrix. The CPM normalization and log transformation was applied to the gene expression data, and batch effect was removed using combat function in sva package in R. This zip file should include:

- The information about each sample: Leaf_Root_annotation.csv
  This should include Project label (accession ID of the original dataset) and Tissue label (tissue of each of the samples).
- The log-transformed raw expression matrix: Leaf_Root_raw_data.csv
- The normalized and batch effect corrected expression matrix: Leaf_Root_normalized_data.csv

**For each expression matrix:**

a. Perform clustering (with 2 and 3 clusters) on the samples and compare the results to Project and Tissue label in the annotation file.
b. Perform PCA, and color the samples once with Project label and once with the Tissue label.

Compare the results of raw and normalized data.

4. **Classification on MNIST dataset**

**Dataset:**

- Load MNIST dataset (could be accessed using from keras.datasets in python)

**For this dataset:**

Use the following classification methods:

a. Logistic Regression
b. MLP with one hidden layer of size 128.
c. MLP with two hidden layers of sizes 256 and 128.
d. CNN with two “convolution + max pooling” blocks and a dense network with one hidden layer of size 128.

Split the data into train and test sets, fit the models using training data and evaluate the models on the test set. Report accuracy, and plot the confusion matrix.