Experiment 3

*Start Date: Nov.09, 2021
*Report Due Date: Nov. 23, 2021

Notes:

- Each experiment must be done individually. You can search material through Internet but remember to mark it.
- Write an experiment report to describe and analyze the experimental observations.
- You should pack your files including **source code, report, readme file, and other related files** in one .zip/.rar/.7z file.
- Please submit on the Web Learning platform. Do NOT Email or Wechat your report to the instructor or TAs.
- No late report is accepted. No exceptions.
- You can write the report using English or Chinese.

Task:

Task1: Visualize the data and try to use an ensemble model to classify.
Task2: Image segmentation via clustering.

Goal:

We hope you can better know about the tSNE or other methods to project the high dimension data to a 2D plane and know better about the ensemble model.

Data:

For **CS background** students, you should use the attached dataset file[1].
For **Non-CS background** students, you can use the sklearn package dataset to load the MNIST dataset.

For image segmentation task, check the data file brainimage.txt

Experiment Steps:

- Use the **tSNE** or **PCA** to visualize the data.
- Do some EDA (Explore Data Analyze) work to analyze the data. There is an example.
- Choose **one** ensemble model (like XGboost/Adaboost/Random Forest) to train and test on the dataset. Try to change the related parameters to see the change of metric you use. Record the result and try to explain it.
- Compare the result with experiment 1 and experiment 2. Give your conclusion about model choosing and parameter setting.
- (Optional, extra score for bonus) There is an imbalance between the data of different labels. Try to solve this problem by using some tricks. (Hint: you can refer to focal loss)

[1] Divide the data into train set and test set.

- There is an image of a brain in the data file brainimage.txt. There are 151 rows in the data file and 171 columns, with each data point an integer value. The image is shown below. Your task is to use EM to classify the pixels into three classes (outside brain, gray matter, white matter).

  Start by normalizing the pixel values from 0 to 1. Start your search with the following parameters (for the variance parameter, set it to the sampled variance of the entire data set). Your system should converge in about 20-30 steps. Show the log-likelihood of your model. Classify each pixel based on your posterior probability after the final step and plot three images to show each class.

  $$P(Z_1 = 1) = P(Z_2 = 1) = P(Z_3 = 1) = \frac{1}{3}$$
  $$\mu_1 = 0.3, \mu_2 = 0.5, \mu_3 = 0.7$$
  $$\sum_1 = \sum_2 = \sum_3$$

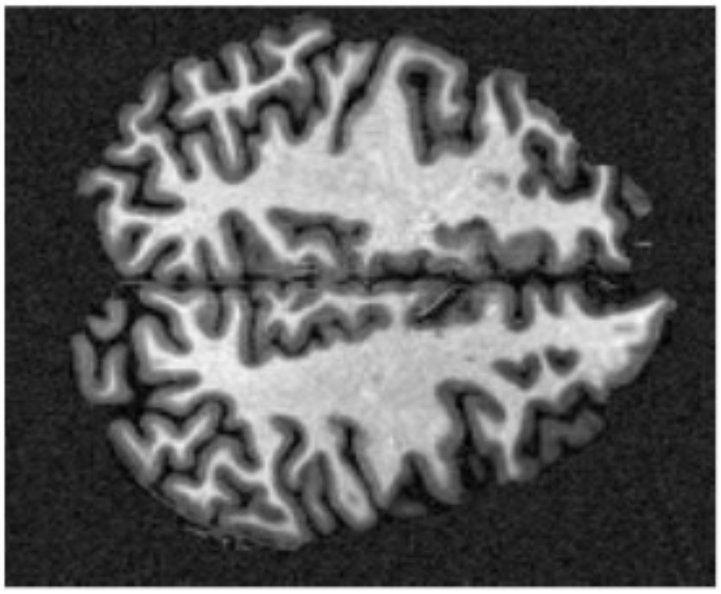

Figure 1: An image of a brain .