CMPE 544 Pattern Recognition
Spring 2025 Assignment 1
Due April 13th by midnight

Expectation Maximization

In this task, you will implement the Expectation-Maximization (EM) algorithm for the Gaussian mixture model. Please download the synthetic data, `dataset.npy`, provided on Moodle. The synthetic dataset is a mixture of three Gaussians. Implement and run the EM algorithm for the synthetic data.

1. (5 pts) Plot the scatter plot of the data.
2. (10 pts) After the convergence, plot the cluster assignments denoted by different colors. Report the estimated values for the mean and the covariance matrix.
3. (5 pts) Please plot the Gaussian distributions you estimated.

Classification

In this task, you will learn classifiers to categorize images in the Quick Draw dataset. The dataset originally comprises 50 million drawings of 345 categories collected from the players of the game called Quick Draw. For this assignment, a subset of the dataset was downloaded from Hugging Face [1]. You may download the subset prepared for this assignment `https://drive.google.com/file/d/1oG0HnQEb7VVwpdYsn8rShfAQ8Vq4hA9E/view?usp=sharing`. The subset has 20k grayscale images of size 28×28 in the training set and 5k in the test set. An equal number of images were selected from 5 categories in training and test sets. A README is provided in the dataset folder for how to load and visualize the images.

1. (10 pts) Please investigate the dataset, extract features, and comment on intra-class variations and inter-class similarities of your features. Explain your preprocessing and feature extraction procedure in detail. (NOTE: You are not allowed to extract features from pre-trained deep networks.)
2. (20 pts) Implement a k-nearest neighbors (k-NN) classifier from scratch. You may choose distance metrics that are different from Euclidean distance. You need to tune the k value. You are free to use an accelerated version of k-NN. Please discuss how you choose the distance metric. Report the test classification accuracy.
3. (20 pts) Please implement the Naive Bayes classifier from scratch. You can assume that the class conditional densities are Normal. Report the test classification accuracy.
4. (20 pts) Please implement the logistic regression classifier from scratch. You are free to add a regularizer if necessary. Report the test classification accuracy.

[1] `https://huggingface.co/datasets/google/quickdraw`

5. (10 pts) Compare the performances of k-NN, Naive Bayes, and logistic regression. Please discuss the advantages and disadvantages you experienced training a non-parametric, parametric, and linear discriminant function classifier. Please comment on whether your features are linearly separable by only comparing the performances of the classifiers.

Submission

- A PDF report addressing the above points, including your name, student number, references, and drive link to the code, should be submitted.
- You should explain how you implemented the classifiers and solved issues in your report.
- Please include the link to your code on the first page of your report.
- Your experiments should be reproducible.
- A readme should be provided.

IMPORTANT NOTES:

- There is no report template. However, your reports should address all the bullet points above.
- You should not forget to cite your references. If you followed a GitHub repository, do not forget to cite it.
- Please note that you should be ready to answer questions regarding your code.