CS 559: Machine Learning: Fundamentals and Applications

HW 7 Due: 4/12/2024 Friday 11:59 p.m.

- The assignment must be individual work and must not be copied or shared. Any tendency to cheat/copy evidence will lead to a 0 mark for the assignment.
- Students must only use Pandas, NumPy, matplotlib, and Spacy if the coding problem does not specify libraries/packages. Use of libraries other than the specified will be penalized.
- All problems must be submitted in a single notebook file. Do not work in the lecture notebook file.

1 KMeans and Gaussian Mixture [35 pts]

KMeans and Gaussian Mixture are algorithms in machine learning that cluster observations by their similarities.

a. Load the following data set.

```
raw_data = []
with open('hw7_data.txt') as f:
    raw_data = [line.split() for line in f.readlines()]
X = np.array([[float(x), float(y)] for [x,y] in raw_data])
m, n = X.shape
```

b. [5 pts] By visualizing the data, guess the number of clusters. Justify your answer.

c. [15 pts] In the Kmean implementation (**KMeans_im** - do not confuse with Scikit-learn Kmeans), points get clustered by the initial k value the user gives to. However, it does not provide stability (e.g., the optimized total within-cluster variance or inertia). Modify the given algorithm to return the within-cluster variance. Then, visualize how the within-cluster converges as k changes from 1 to 15. Explain which k value is the most appropriate and justify your answer.

d. [15 pts] Run the train_gmm algorithm (the implemented Gaussian Mixture algorithm) to check if the k value answered in (c.) is a good choice. Visualize the change of log-likelihood as k increases from 1 to 15. Then, compare the visualization of clusters with KMeans results from (c.). Explain how results are similar and different.

2 PCA [20 pts]

In this problem, the same data set from Section 1 will be used.

a. [10 pts] In the lecture, the algorithm of PCA and the scikit-learn example are discussed. Implement the PCA algorithm using **NumPy**.

b. [5 pts] Find the principal direction and components.

c. [5 pts] Perform the dimensional reduction using Scikit-learn LDA. Determine the hyperplane equation and projectile points to the hyperplane. Compare the visualization and explain the differences. To understand the differences, having the projectile in the same original space will be good.