Problem Set 9

Data Science 602: Data Analysis and Machine Learning

Spring 2022

1. k**-means**. In the 602 shared drive, the file "/data/cluster.txt" contains a features matrix $\mathbf{X} \in \mathbb{R}^{10,000 \times 20}$. (You can use `np.loadtxt` to read the file as a numpy matrix.) This features was generated from scikit-learn's `make_blobs` function with parameters of n_features=20 and centers uniformly distributed in $[-5, 5]$ along each axis. The dataset was produced with at least 5 but no more than 15 clusters. Find the number of clusters used to generate the dataset (i.e., k that best clusters the data), and justify your conclusion using silhouette plots.

2. **Gaussian Mixture Models**. The Fashion-MNIST dataset is a dataset of 10,000 grayscale images of size 28×28. Each image depicts an article of clothing. Load the dataset from OpenML (the dataset name is "Fashion-MNIST") and retain only the first 5,000 images. Find an optimal number of Gaussian components based on the Akaike Information Criterion (AIC), and use the AIC scores to cluster the data using a GMM. Visually inspect the model outcomes. Is the clustering the model selected meaningful (e.g., are the cluster separations intuitive)?

3. **Outlier Detection**. Import the MNIST-784 dataset (handwritten numbers), and keep only observations labeled as '7'. Using an outlier detection method of your choice, identify outlier observations, i.e., observations that do not resemble other observations in the dataset. Display a sample of 5-10 detected outlier observations.