SMAI (CSE 471)
Spring-2019
Assignment-9 (100 points)
Posted on: 21/03/2019
Due on: 30/03/2019, 11:55 PM

- Questions can involve a mix of writing code/scripts and answering questions or analyzing results.
- Code: Your scripts should be of the form q-x-y.py where x is the main question, y is the sub-question. For e.g., q-1-2.py is Python script for sub-question 2 within question 1. If you are submitting Jupyter notebook file (.ipynb), make sure that it is properly formatted and documented with question part numbers (Part-1, Part-2 etc.).
- Ensure that submitted assignment is your original work. Please do not copy any part from any source including your friends, seniors and/or the internet. If any such attempt is caught then serious action will be taken.
- Use suitable train-validation split for your training and validation (20% of data).
- Numpy, pandas/csvReader(for data processing) are allowed.
- Report should contain details of algorithm implementation, results and observations.

1 Question

1. (70 points) In Assignment 3 , Question 1 you had worked on the problem of anomaly detection. Continuing with the same problem, now we will explore other approaches to dimensionality reduction.

   .

   1. **Part-1**: (20 points) Now apply dimensionality reduction on the dataset using:
      - 3-Layer autoencoder consisting of input , output and bottleneck layers in which

      a. input and output layers have linear activation functions.

      b. input and output layers layers have non-linear activation functions

      - Deep Autoencoders where you increase the number of hidden layers and use non-linear activation functions.

      In all the autoencoders, the number of nodes in bottleneck layer are equal to the number of reduced dimensions you got by using PCA by keeping the tolerance of 10

2. **Part-2**: (15 points) Use the reduced dimensions from all the techniques in the first part and perform K-means clustering with k equal to five(number of classes in the data). Also calculate the purity of clusters with given class label.

3. **Part-3** (15 points) Perform GMM (with five Gaussian) on the reduced dimensions from first part and calculate the purity of clusters. You can use python library for GMM.

4. **Part-4**: (15 points) Perform Hierarchical clustering with single-linkage and five clusters on the reduced dimensions from all the techniques in the first part and calculate the purity of clusters.You can use python library for hierarchical clustering.

5. **Part-5**: (5 points) Create a pie chart comparing purity of different clustering methods you have tried for all classes for the different autoencoders.

2 Question

2. (30 points) Problem of Generating New Data: We will be working on handwritten digits dataset available in sklearn.
   Apply dimensionality reduction using PCA , to reduce the number of features to 3 values in the range 15 to 41. For all the reduced number of features do:

   1. **Part-1**: (10 points) **Kernel Density Estimation:**
      Use grid search cross validation on the reduced feature data to optimize bandwidth
      Compute Kernel Density Estimate
   2. **Part-2**: (10 points) **Gaussian Mixture Model based Density Estimation:**
      Use Bayesian Information Criteria to find the number of GMM components we should use Apply GMM using the using the above number of components
   3. **Part-3**: (10 points) Draw 48 new points in the projected spaces using both the above generative models. Use Inverse transform of PCA to construct new digits
      Plot these points from both the models.