SMAI (CSE 471)
Spring-2019
Assignment-3 (100 points)
Posted on: 09/02/2019
Due on: 17/02/2019, 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) **Problem of anomaly detection:** You are given the dataset of network user activity, and the task is to classify each user activity as normal or an attack. Attacks are also categorized as follows-

   - Denial of Service (dos): Intruder tries to consume server resources as much as possible, so that normal users can't get resources they need.
   - Remote to Local (r2l): Intruder has no legitimate access to victim machine but tries to gain access.
   - User to Root (u2r): Intruder has limited privilege access to victim machine but tries to get root privilege.
   - Probe: Intruder tries to gain some information about victim machine.

   Download dataset from here (`http://researchweb.iiit.ac.in/~murtuza.bohra/intrusion_detection.zip`). Dataset contains 29 numerical features and five classes(one normal and four attacks).

1. **Part-1**: (20 points) Do dimensionality reduction using PCA on given dataset. Keep the tolerance of 10% (knee method), meaning reconstruction of the original data from the reduced dimensions in PCA space can be done with 10% error. You are only allowed to use eigen decomposition or SVD function from python library(do not use library function to compute PCA directly).

2. **Part-2**: (15 points) Use the reduced dimensions from 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. Purity is the fraction of actual class samples in that cluster. You are not allowed to use inbuilt function for K-means.

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. Also calculate the purity of clusters. Create a pie chart comparing purity of different clustering methods you have tried for all classes. You can use python library for hierarchical clustering.

5. **Part-5**: (5 points) Original data of network user activity is available here(`https://www.kaggle.com/what0919/intrusion-detection`). Original data also contains categorical feature. If you were to do dimensionality reduction on original data, could you use PCA? Justify. Write a paragraph in report for your explanation/justification.

2 Question

2. (20 points) **Question carry forwarded from assignment-2**. Use the Admission dataset to perform the following task. Dataset can be downloaded from `http://preon.iiit.ac.in/~sanjoy_chowdhury/AdmissionDataset.zip`

   1. **Part-1**: (10 points) Implement logistic regression model to predict if the student will get admit.

   2. **Part-2**: (5 points) Compare the performances of logistic regression model with KNN model on the Admission dataset.

   3. **Part-3**: (5 points) Plot a graph explaining the co-relation between threshold value vs precision and recall. Report the criteria one should use while deciding the threshold value. Explain the reason behind your choice of threshold in your model.

3 Question

3. (10 points) Implement logistic regression using One vs All and One vs One approaches. Use the following dataset `http://preon.iiit.ac.in/~sanjoy_chowdhury/wine-quality.zip` for completing the task. Report your observations and accuracy of the model.