Habib University
CSE 351 - Artificial Intelligence
Fall 2018
Assignment 2

Q-1 - Classification

Kaggle is a platform for predictive modeling and analytics competitions on which companies and researchers post their data and statisticians and data miners from all over the world compete to produce the best models. This assignment will expose you to different applications of Machine learning.

You will attempt the following two questions. For each question, you have to:

- understand the problem thoroughly
- download the dataset
- Write python code to:
    - Load data from the file
    - Perform necessary pre-processing to get the data in your desired format
    - Partition the data into training and test splits
    - Use the training split to learn multiple classifiers (try different classifiers available in scikit-learn)
    - Test these classifiers on the test split
    - Report accuracy of these classifiers in terms of accuracy, precision, recall, and f-measure
- Consolidate your results in a pdf that you will submit along with the python code.

Problem- 1.1: [10 Points]

To familiarize yourself with the environment, you will first go through the scenario of 'Titanic: Machine Learning from Disaster' and will build a model to predict the survival of passengers.

Problem – 1.2: [25 Points] Choose **any one** of the given problems and build a predictive model for that problem using scikit-learn library.

- Gender Recognition by Voice
- Twitter US Airline Sentiment
- Detecting keypoints on face images
- Spooky Author Identification
- SMS Spam Collection
- Confused Student EEG Brainwave data

Problem - 1.3 [15 Points]

You are given a Fashion-MNIST dataset comprising of 28x28 grayscale images of 70,000 fashion products from 10 categories, with 7,000 images per category. The training set has 60,000 images and the test set has 10,000 images. The dataset is freely available here.

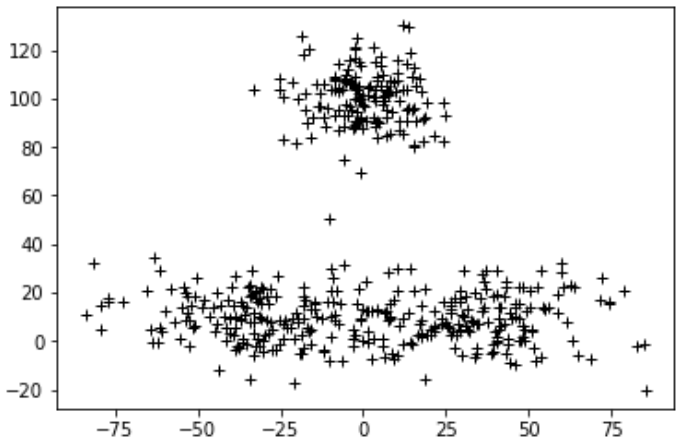
You have to build a scikit-learn based classifier for this dataset to classify each input image into one (out of 10) categories. You are required to experiment with different classifiers and report accuracy of each of them.

Q 2 – Clustering

2.1 Implement k-means algorithm to cluster a set of 1000 2D points. The data points (x,y) can be generated via Gaussian distribution using the sample code below:

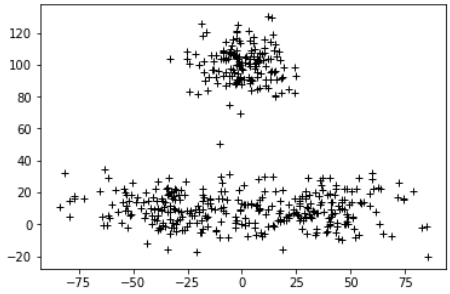
```
points = []
def initializePoints(count):
     for i in range(int(count/3)):
         points.append([random.gauss(0,10),random.gauss(100,10)])
     for i in range(int(count/3)):
         points.append([random.gauss(-30,20),random.gauss(10,10)])
     for i in range(int(count/3)):
         points.append([random.gauss(30,20),random.gauss(10,10)])
```

This code generates following data points:

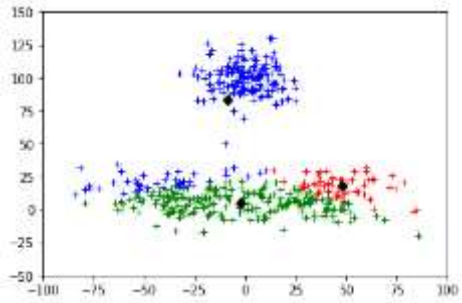


1. **[15 Point]** Implement K-means algorithm to cluster these points. K-means will take a set of points and the number of clusters (K) and will cluster the points into given number of clusters. The algorithm will stop if there is no significant change coming in positions of centroids. You can try giving different mean and variance to Gaussian distribution to have different dispersion of points.
2. **[10 Points]** Provide graphical visualization of the process of formation of clusters via K-means (as shown on the next page).
3. **[10 Points]**Instead of running K-means once, run it 10 times and give the best clustering formation achieved. You need to identify the criteria to assess the quality of clusters and then choose the best one.
4. **[15 points]** K-means has this limitation that you have to provide number of clusters (K) to be formed as an input. Can K-means be modified or supplemented with another algorithm that itself estimates optimal number of clusters in an unlabeled dataset? Do some research on this and write a summary of possible approaches in your own words (with references).
   (Note: This part does not require any implementation).

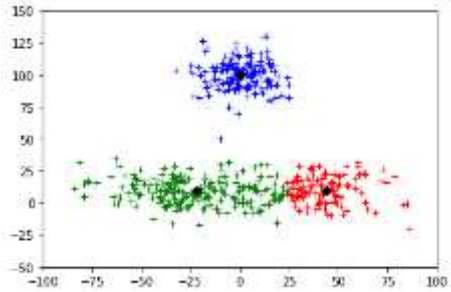
Iteration 0



Iteration 1



Iteration 2



.
.
.

Iteration 9