CS4801: Principles of Machine Learning
Programming Assignment 3

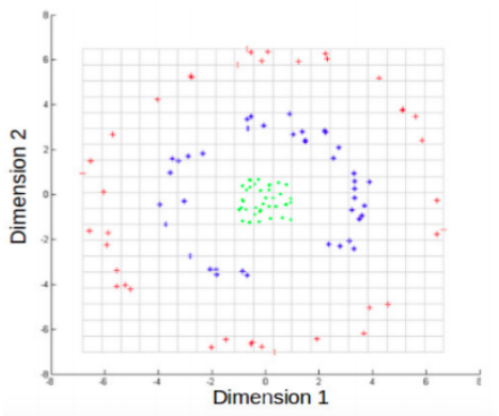
8 points
Grading will be focussed on the report you submit
Due on 26th October 2017
No request for change will be accepted

This homework consists of only programming assignment on Clustering, PCA and Random forest. A few instructions to make life easier for all of us:

- please submit your code and a elaborate discussion on your observation (preferably PDF and latex) from your experiments. Put all codes and report in a single zipped file and name it as <First-name><Last-name>.zip. Then submit it in moodle.
- Deadline for programming assignment is 17:00 pm 26th October 2017.

(3points) Exercise 1 : Compare clustering models

- Generate a 3D data-set of 3 concentric balls [where 2D projection on any two dimensions will generate the following picture]



- Comparing results of following clustering methods [plot clustering output, discuss which methods works(does not work) well and why(why not)]
  - K-means
  - Spectral with gaussian kernel (choose your kernel width)
  - GMM

(2 points) Exercise 2 : Feature extraction

- Use the same the data set you have created in Exercise 1. Generate one(and two) dimensional(s) embedding for the data using PCA and KPCA [use $K(x, y) = (x^T y+1)^2$] discuss the result [plot clustering output, discuss which methods works(does not work) well and why (why not)]

(2 points) Exercise 3 : Random Forest

- Download data from `https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data"`. This is the Breast Cancer Wisconsin (Diagnostic) Data Set from UCI repository.
- make a random train -test split of $(60\%, 40\%)$.
- Find out best performance with random forest for this data. Height of individual DT and also total number of component DT can be optimized with prediction error on out-of-box samples.
- Compare performance of random forest with performance of individual models.