CS4786 Kaggle 1: Early Report

Alexander Ueki (aru5), Divyansh Garg (dg595), Jaiveer Singh (bs672), Ryan Curtis (rec284)

October 16, 2017

1 Preliminary Implementation

Our preliminary implementation will focus on using Canonical Correlation Analysis (CCA) and clustering to produce labels. Given the feature vectors (F) from Extracted_features.csv and the similarity graph (G) from Graph.csv, we will produce a new data-set S using the 6000 points that are in both F and G (F_G), such that $CCA(F_G, G) = S$. Using S, we will apply a clustering algorithm (K-means) to create 10 clusters C via $Cluster(S, K = 10) = C$. Using the 60 labeled points from Seed.csv, we can label C. The remaining 4000 points (F_L) are then labeled based on which cluster they belong to using our clustering method.

We will expand our preliminary method in the following ways to hopefully improve accuracy:

1. Use component analysis methods, like PCA or Random Projections, on F to try and remove noise before using CCA.
2. Vary clustering methods, e.g Single-Link, Hard Gaussian MM.
3. Attempt to label the F_G and use that data for supervised classification learning.

2 Individual Plans (Currently)

- Alexander Ueki
    1. Assist Preliminary Implementation.
    2. Contribute to Early Report.
    3. Contribute to Final Report.
- Divyansh Garg
    1. Complete Preliminary Implementation.
    2. Contribute to Final Report.
- Jaiveer Singh
    1. Expand Preliminary Implementation with new algorithms.
    2. Contribute to Final Report.
- Ryan Curtis
    1. Expand Preliminary Implementation with new algorithms.
    2. Contribute to Early Report.
    3. Contribute to Final Report.