Unsupervised Learning and Dimensionality Reduction

Yi Zhao

[Yzhao644@gatech.edu](mailto:Yzhao644@gatech.edu)

1. Introduction

2. Dataset

3. Clustering

3.1 Metrics to consider (silhouette, AIC/BIC,

3.2 Elbow method

4. Dimensionality Reduction

4.1 PCA

4.2 ICA

4.3 Randomized Projections

4.4 Decision Tree Feature Importance

5. Clustering after Feature Reduction

Cluster is easier after feature reduction. As it’s projected to new dimensions that are independent or orthogonal to each other. Easier to map to a cluster.

It can be shown with visualization and silhouette\_score (kmeans)increase with same number of clusters.

Except trees, as it’s not changing anything.

For kmeans tree and ori choose about 20 clusters, others about 50. Inline with em result.

For EM, it’s similar effect, as one can see ica has a sharp drop in the 2-20, as it doesn’t need many clusters to reach the elbow point, whereas pca and rp needs about 50 clusters to reach the turning point.

More feature does end up with a lower aic score. But sihousette score is low since it’s not very sepratable.

Best number of clusters has a relationship with number of clusters. If bit wise feature, number of clusters = 2\*\*n

Rp has the lowest aic, meaning it’s matching the distribution well, but the si score is lowest, meaning it’s not cluster well?

Check speed of running algorithms. Rca should be faster than others.

Clustering info could be learnt through neural net, so use original data clusters

Visualization is different between original and feature reduction.

Try different variance for EM (more is not necessary better, but we will stick with it.)

Variance reduction in pca is not significant, meaning that there are not much noise in the features to reduce.