Visual Analytics Homework-03

Nikhil Yadav

24th September 2017

Github link to jupyter notebook: `https://github.com/nikhil15iitd/visual_analytics_hw03/blob/master/hw03.ipynb`
The notebook has embedded bokeh server, so should run on local machine with default notebook_url = 'localhost:8888' without calling bokeh serve command.
Visualization is done using scatter plots, with selectors & sliders for specifying number of clusters, affinity(distance metric to consider for clustering) & domain transform like PCA & t-SNE(Stochastic Neighbour Embedding).
Five different clustering algorithms were chosen, performance measured in the domain of t-SNE, not original domain of 6 attributes:

- K-Means: performs good clustering, euclidean based, performance dependent on cluster parameter, non deterministic, highly reliant on initial placement of centroids.
- Spectral Clustering (slower to compute because rbf kernel transforms are expensive, so page may hang): Performs the neatest clustering of points (reason is because it uses RBF kernel which has infinite degree if it is expanded using taylor series)
- Agglomerative Clustering: Performs poorly in t-SNE domain using cosine distance metric, euclidean metric works better
- DBSCAN: considers all points as noise, only in t-SNE domain does it assign some points as clusters, does not have n_clusters parameter, hard to tune, the reason for its bad clustering is because data is homogeneous so no good clustering can be figured out based on density
- AffinityPropagation: performs good clustering, euclidean based, cannot be changed, however it requires no cluster parameter like K-Means or Agglomerative Clustering, so is easy to use, also robust & deterministic

Out of all affinities(euclidean, l1, l2, manhattan, cosine), euclidean works the best in t-SNE domain, & almost all clustering algorithms by default use euclidean as distance metric (affinity) for clustering data points. In PCA domain or default domain of 6 attributes, it is difficult to visualize the performance of clustering algorithms, because it is difficult to display clusters in 6 dimensions. PCA suffers in dimensionality reduction because its first principal component preserves the maximum variance & second holds no information, so again comparison is difficult. However, t-SNE embeds all dimensions in 2D, preserving information compared to PCA, so visualization can be done easily compared to other domains.
Log normalize, x-axis, y-axis & point size selectors work only when domain transform = "None".
Snapshots of visualization:

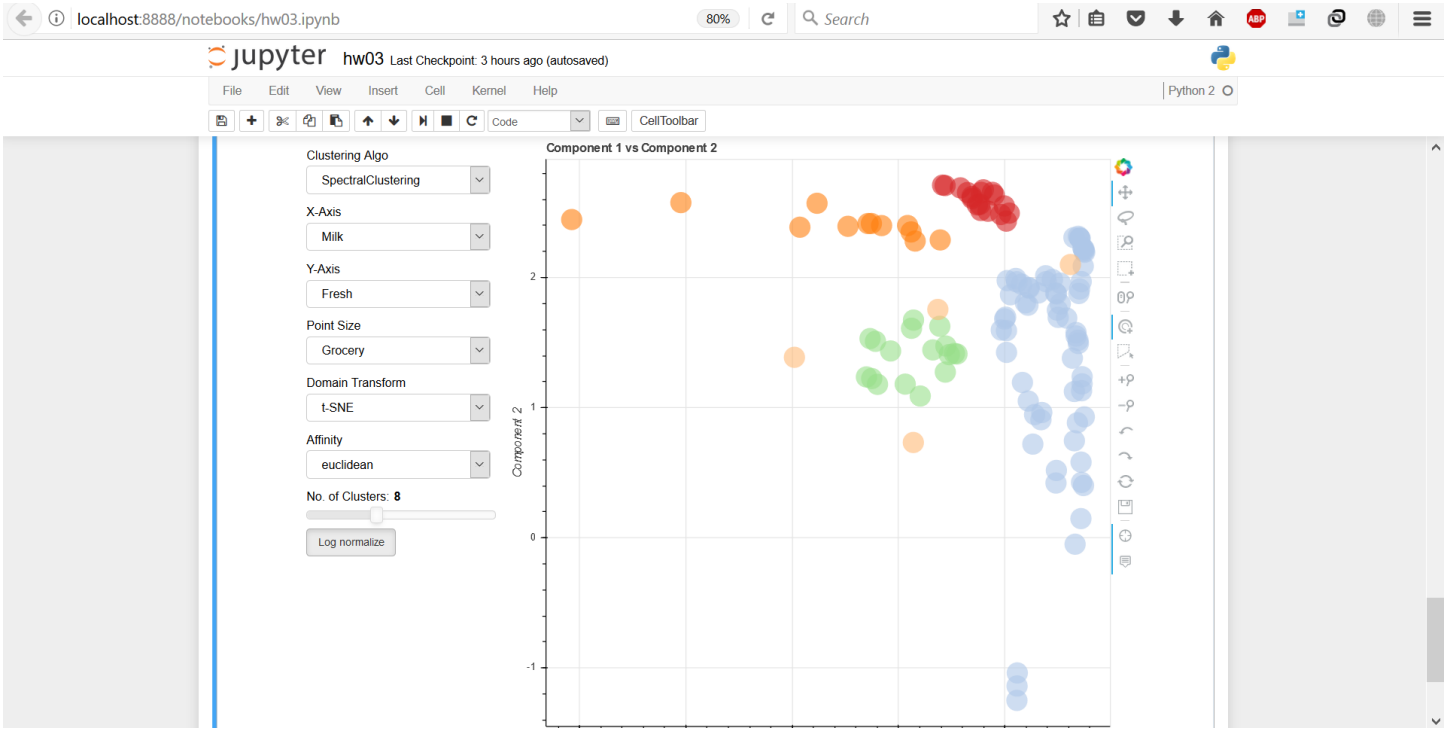

Figure 1: Scatter plot: Spectral clustering, t-SNE domain

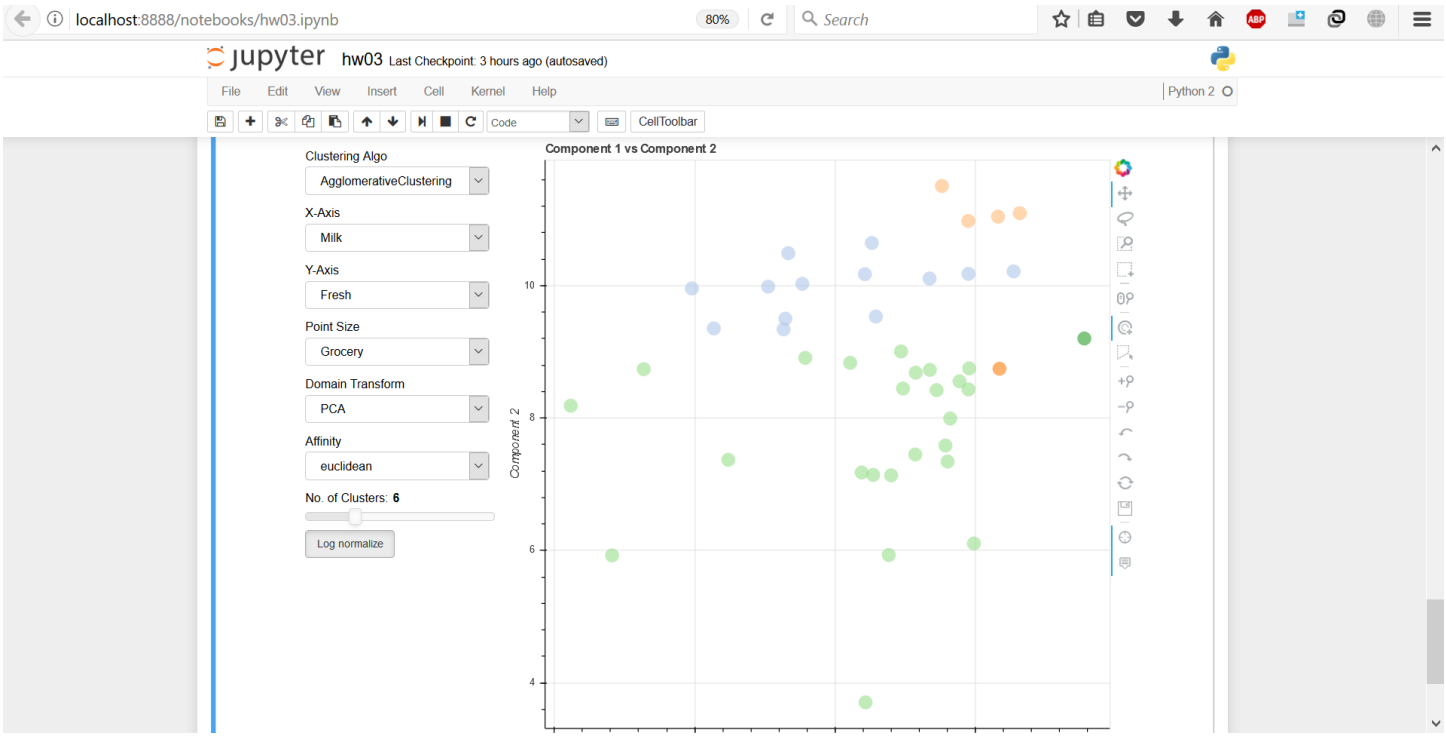

Figure 2: Scatter plot: Agglomerative clustering, PCA domain