Noise Reduction Engine for SCoNEs

Aim

Implement a basic noise reduction function for log Ratio signal of the 2 samples contained in the test data.

Implementation

For noise reduction, i implemented a basic function which uses DBSCAN clustering algorithm for noise detection. Then i removed the detected noisy data points from the input test data and generates an output file that is noise free. Due to the following factors DBSCAN is useful in noise reduction.

- It can determine noisy points and is robust to outliers
- It can itself determine the number of clusters.
- It can find clusters of arbitrary shapes and sizes
- The parameters **minPts** and **eps** can be set by a domain expert, if the data is well understood (so here i guess my mentors experience will be very helpful to me)

Results

Parameter Values: **eps= 0.10** and **minPts= 11**
Total data points before noise reduction: **15488**
Noisy points detected: **60**
Total data points before noise reduction: **15428**

File Type	File Name
Souce Code	noiseReduction.R
Plots (Black denotes noise)	Rplots.pdf
Input File	SCONES_test.tsv
Output File	noiseless.tsv

Conclusion and Future Work

I set the parameter values by using trial and error method. But one possible way for doing this as suggested in DBSCAN paper and DBSCAN R documentation is to compute a k-distance plot of the dataset. We compute the k-nearest neighbors (k-NN) for each data point to understand what is the density distribution of our data, for different k. Once we choose a minPTS (which strongly depends on your data), you fix k to that value. Then we can use as epsilon the k-distance corresponding to the area of the k-distance plot (for your fixed k) with a low slope.

Besides using DBSCAN we can also use,

- **Manifold Learning** as it works well to determine noisy data points in high dimensions.
- **Regression (Curve Fitting)** will help us to learn the signal function from the sample data and hence based on generated function we can predict the values. Values that deviates more than the threshold could be labelled as noise.

By:

Saket Maheshwary
MS by Research CSE
Center for Data Engineering
IIIT Hyderabad, India