STATISTICS SEMINAR

Yu Shi
Master's Defense

Monday, July 27th, 2020
Via Zoom, 1:00 – 4:00 pm

Evaluation of Optimal Clusterings Found by Cluster Validation Measures

There are many measures developed for assessing clustering algorithms. However, little work has been done to determine what type of clusterings these validation measures would consider "the best." In particular, if a clustering validation measure performs well, then it should be able to identify the "correct" clustering when when presented with all possi-ble ways of clustering a dataset. We evaluate the performance of five clustering validation measures-Silhouette, Hubert-Gamma, R-squared, the Dunn family of indices, and the data Davies-Bouldin index-on five small clustered datasets. To obtain a large set of candidate clusterings, we view each dataset as a graph and form a connected bottleneck subgraph. On this subgraph, we identify all set-connected partitions-those whose blocks are connected-that satisfy a set of constraints on the number of blocks and the size of each block within the partition. We then apply the validation measure on each of the possible partitions to determine the clustering that each validation measure considers to be optimal. Based on test results, we find each measure has its own preferences. For example, the silhouette measure tends to be better at capturing connected regions, and many others measures prefer cluster-ings that contain many clusters. Finally, we compare the clusterings found by the validation measures to those obtained by other popular clustering methods including k-means, hierar-chical agglomerative clustering (HAC), density-based spatial clustering of applications with noise (DBSCAN) and ordering points to identify the clustering structure (OPTICS).