Gian Thomas Wrobel

Foundations of Computational Data Analysis

Week 8

EM models benefit from being interpretable in a probabilistic way and actually revealing a “model” which has generated the clusters we are observing. For example, if a cluster has a more tightly packed core of values with very few outliers, the standard deviation parameter could be appropriately adjusted to reflect this fact. K-means has no such flexibility and merely assigns data to its nearest center without consideration of whether that assignment is particularly “likely” given other observed data.

EM models more generally are also just a “class” of clustering algorithms and so there is greater flexibility in how clusters are determined and assigned. Fuzzy clustering, for example, is a powerful tool for considering the fact that data could potentially belong to multiple clusters. K means will assign data to a single cluster and cannot take into account the possibility of multiple assignments.