Lachlan Sinclair

3/1/2020

Week 8 Assignment 2

Reflections

a. Decision trees continually make splits in the data set based off the information gain obtained by splitting on a single attribute regarding a specific target value. This can be easily visualized in a 2D setting. Decision trees continually added vertical and horizontal lines parallel to the axis that represent decisions made and act as the partitions. This process continues until the model is properly fit to the data. Clustering evaluates feature vectors and creates clusters based off feature vectors similarity to one another and does not consider a target value. Since clusters partitions are formed by the similarity between feature vectors the partitions end up looking more fluid when compared to the cell like structure of the decision trees (in lower dimensional situations). In my opinion the most significant difference is that decisions trees are formed with a single target value dictating the partitioning, this greatly differs from using the similarity of multiple selected attributes to create partitions as seen in clustering. The partitions for a decision tree will ideally cleaning separate different values of the target attribute, and clustering will partition attribute vectors that have closely related attribute values.

b. The correct percent of a completely overfitted clustering model will be 100%. In a dataset of size n, when there are n clusters this will occur, such a model will behave like a lookup table causing it to have 100% correctness. The number of clusters defines the complexity of the model, so as we increase the number of clusters, we will gradually see the correct % increase just as we saw when we began to overfit the decision tree models. This is why it is important to consider the other attributes of the model we recorded in the assignments. For a completely overfitted model the silhouette will be 0 which is not ideal and should indicate that there is an issue. On the other end of the spectrum, the correctness of the model will approach the overall correctness of the entire dataset when the model is underfitted. With a cluster size of one, the entire dataset will be encapsulated in the cluster causing the correctness to equal the correctness of the dataset. This again matches the predictive capabilities of a underfitted tree model on the training set, when underfitted it basically adds no value to the already existing average probability of the entire set.