Assignment 8

Applied Machine Learning

Please refer to this assignment's synthetic dataset. This dataset is composed of two features belonging to observations that can be grouped into certain species. The goal is to refine the data so that we can classify species together and use their features to classify new observations.

1. [10 pts] How many species must be there in the dataset? (For the rest of this assignment, use that number as the number-of-clusters parameter in methods such as `KMeans`)

2. [10 pts] Find the rough feature ranges to classify these species correctly. It might be a good idea to do this step visually from some data plots. (An example statement, "`Species 1 has feature A in the range of [0-1.5]`")

   In the next set of problems, we will clean the points that are around the boundaries of the cluster (these points might be due to errors, anomalies, or they are simply outliers). This step is done to refine feature boundaries so that a scientist can classify the species manually, reliably, and with a high-level generalization.

3. [20 pts] Use K-means clustering to find anomalies (Hint: find cluster data points that are far from the centroids).

4. [20 pts] Use DBSCAN clustering to find anomalies in the full dataset as an alternative to Q3.

5. [30 pts] Now, using the cleaned dataset by a method of your choice (i.e., Q3. or Q4.), develop a decision tree classifier to model the species. Visualize the model decision tree (and not just by plotting lines on a scatterplot of the data).

6. [10 pts] Show that, in fact, it helped to clean the outliers before generating the decision tree. (Hint: we're looking for something more substantial than a 100% vs. 97% accuracy improvement.).