Arman Irani

Isolation Forest

After preprocessing our dataset to contain only a select amount of features, an Isolation Forest will be designed in Python from scratch in order to discover “anomalies” in our dataset of quasar features. Currently the algorithm is in production, and as is with a standard Isolation Forest will consist of an ensemble of randomly generated Isolation Tree’s and an anomaly will be defined using a depth-based approach as those with short average paths, with a score between [0,1]. The instances with the highest anomaly score will be considered to be outliers. This will hold true under the assumption that anomalies will require less splits than normally required. The Forest will be designed to deal only with continuous data. Parameter tuning will be utilized to optimize the subsampling value and the number of trees during the training period. It will be difficult to tune these parameters as there is no efficient way to confirm our outlier results so our parameters will be selected carefully.