Fahim Khan
CS6923 Machine Learning
5/1/20

Final Project: Nearest Centroid

Nearest Centroid is an extension for K-Nearest Neighbors.

- Nearest centroid finds the center of each class, and then classifies test data to the closest centroid's label.
- This is like a supervised version of k-means clustering, which sklearn also implements; but I used nearest centroid since K-Nearest Neighbors was a supervised algorithm.

The numpy and sklearn implementations use nearly the same methods, so accuracy was the same between both counterparts.

Nearest centroid works best in cases where the centroid (mean of each class) best represents each class data. However it does predict poorly when each class has very different variances.

K-NN can also predict poorly when one class makes up most of the data, or when test data is far away from the other classes. So the nearest neighbors can be mostly of the wrong class.

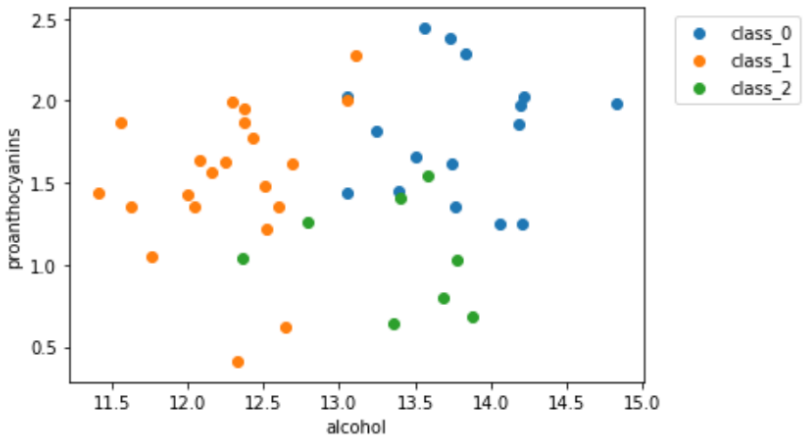
Both K-NN and NC can predict poorly when there are too many features.

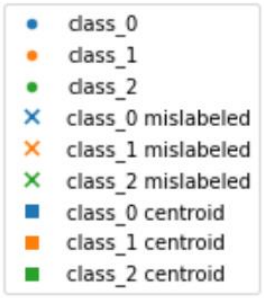
Nearest centroid also has better performance than K-NN

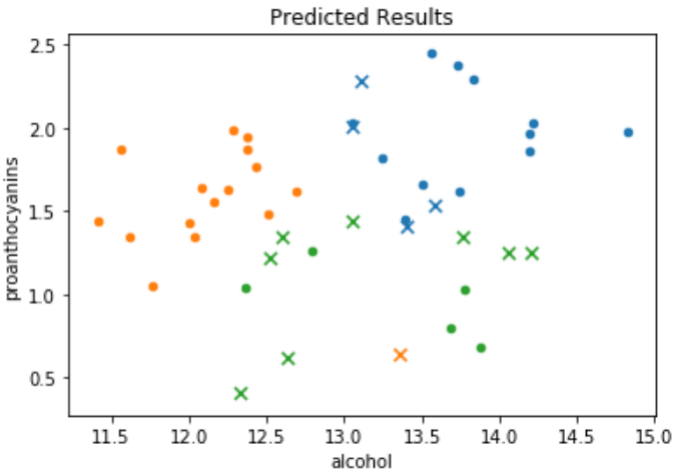
- Centroid takes the mean of each class and compares test data to each class centroid
- K-NN compares each test data all other points (brute force method)

The wine data (alcohol vs proanthocyanins) was a good case where nearest centroid was more accurate than nearest neighbors.

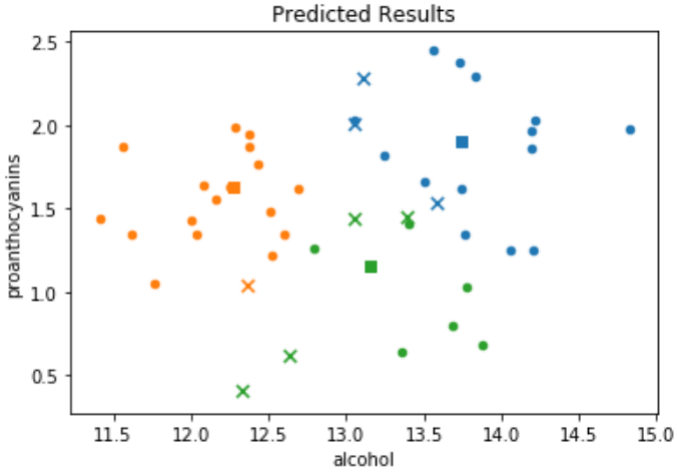
Wine Test Data



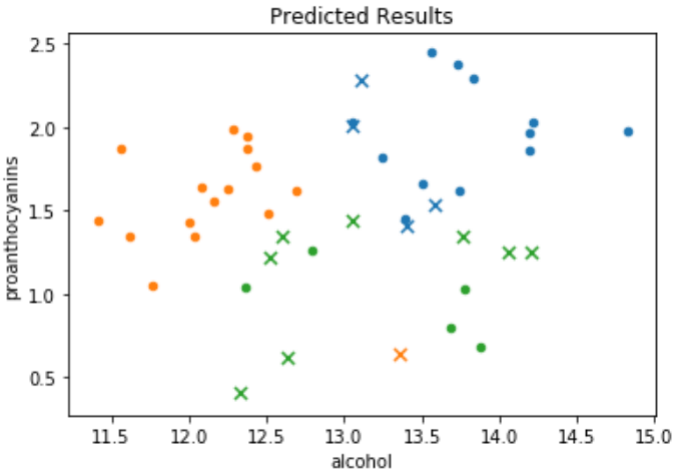




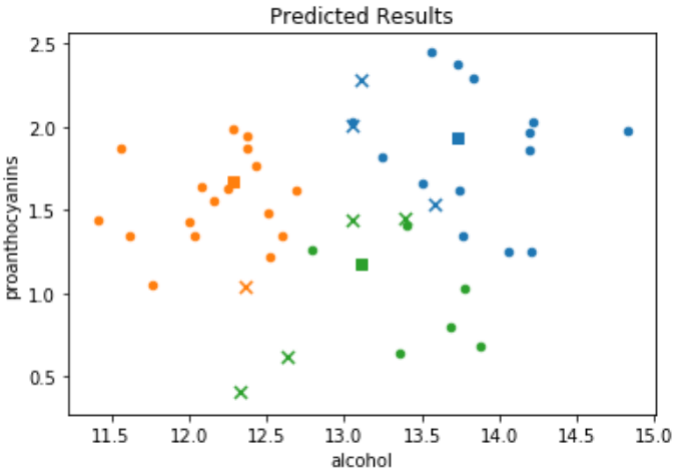
Sklearn K-NN: Accuracy = 71.11%



Sklearn NC: Accuracy = 82.22%



Numpy K-NN: Accuracy = 71.11%



Numpy NC: Accuracy = 82.22%