1 Introduction

This assignment serves as an introduction to python programming as well as machine learning with scikit-learn. To demonstrate scikit-learn's utilization, K-Means clustering, an unsupervised machine learning algorithm, was used on the Wisconsin Breast Cancer Data Set, a publicly available data set including breast cancer tumor samples, sample attributes, and malignant vs benign behavior was obtained. K-means was used to partition data into a variable number of classifications, and classifications were visualized with matplotlib.

Methods

The breast cancer data set was loaded into a PANDAS data frame using the sklearn library. This data set was compiled by Dr. Wolberg et. al in 1995 from fine needle aspiration samples of 569 breast cancers. For each sample, data was collected from digitized images of the tumor cell nucleus. These 30 columns consists of the mean, worst, and standard error values for each of 10 attributes: radius, texture, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. Each sample also has a numerical encoding for benign or malignant classification. Descriptions of the data set as well as summary statistics are available within the loaded dictionary data-structure, keyed by the string "DESCR".

Dimensionality reduction was achieved by using a simple approach of utilizing the first two attributes, the mean radius and mean texture. More specifically, mean radius refers to the average distance to center of nucleus from the perimeter points, and texture refers to the standard deviation of gray-scale values. These two attributes were intentionally selected because they encode information with regards to size and complexity of the tumor nuclear appearance. A scatter plot of the these two attributes was thus constructed using matplotlib. A second scatterplot with points color coded based on benign/malignant behavior is also given.

Unsupervised machine learning was achieved using a K-means approach implemented in Python sklearn library. Clustering was applied to these first two attributes. This was repeated for 2, 4, and 8 clusters. Each of the resulting classifications based on K-mean clustering was visualized with a matplotlib scatter plot.

Results

The first two attributes in the data, mean radius and mean texture, were graphed on a scatter plot with and without color coding by malignant vs benign classification

K-mean clustering was applied to the data set using training data obtained from the first two attributes in the data frame, mean radius and mean texture. K-mean clustering was performed using 2,4, and 8 clusters and are shown in figures 3-5.

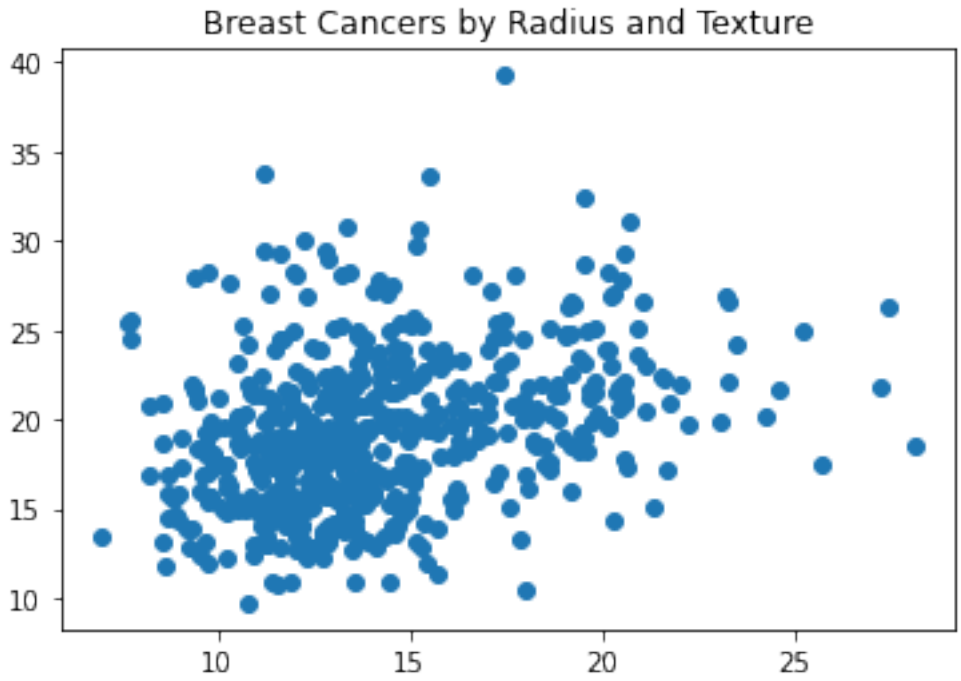

Figure 1: scatter plot of x (radius) against y (texture) of breast cancers

Discussion

When color-coded to actual malignant vs benign classification, we observe that malignant tumor samples trends towards larger mean nuclear radii whereas benign trends towards the smaller values. There is an area of overlap between the two populations at radii values around the interval between 11 and 15. Texture plotted on y axis does not appear separate malignant vs benign populations.

The actual malignant vs benign classification (figure 2) was more independent of texture, whereas the k-mean prediction model took prediction into account, dividing the two populations on a diagonal in the scatter plot (figure 3). This suggests that the true biology of tumor nuclei is not well encapsulated by the texture measurement, whereas the k means algorithm simply computes averages based on the values it receives.

Ultimately, this construction of using the first two attributes alone did not separate the breast cancer data into two clearly distinct populations of tissues. For this assignment, dimensionality reduction was performed using a simple approach that required knowledge of the data and data attributes. Principal component analysis may be able to utilize the remaining attributes in the data set, and a scatter plot of the first two principal components may be able to distinguish the data better.

K-mean clustering with additional clusters yielded similarly sized groups with similar variances within each group, but do not clearly separate distinct populations tumors into groups that are biologically meaningful. That is to say, while we can coerce k-means to

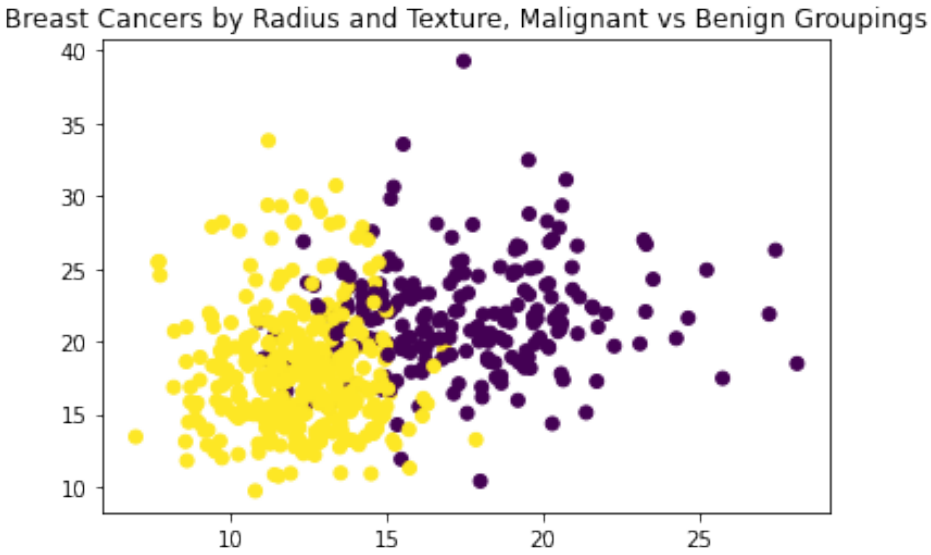

Figure 2: scatter plot of x (radius) against y (texture) of breast cancers color coded by malignant vs benign behavior. Malignant is purple, benign is yellow

separate the data into 4 or 8 groups, it is improbable that these actually correlate to a meaningful difference in tumor behavior.

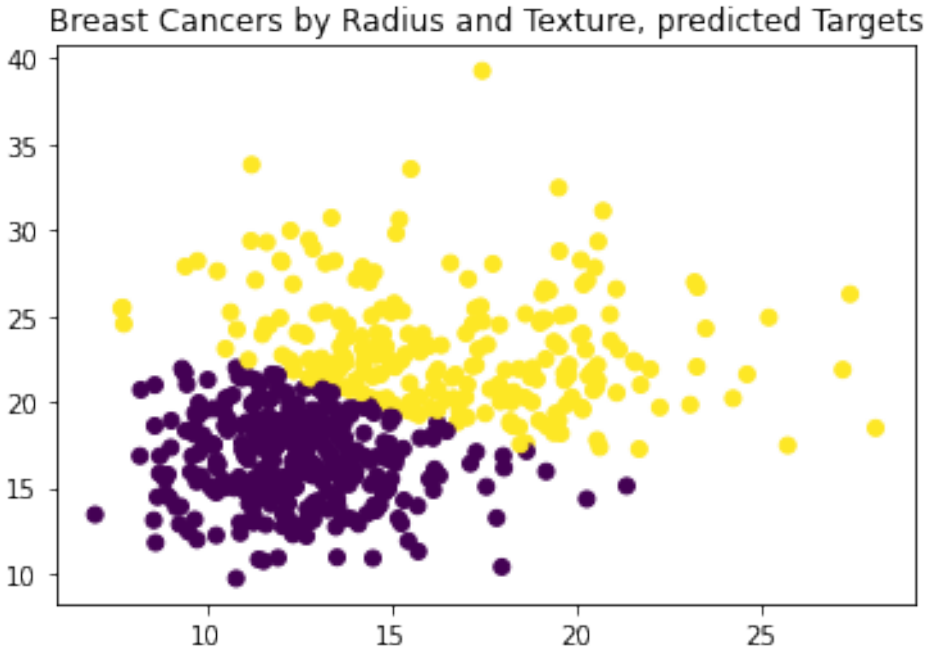

Figure 3: scatter plot of x (radius) against y (texture) of breast cancers using k-means computing 2 clusters

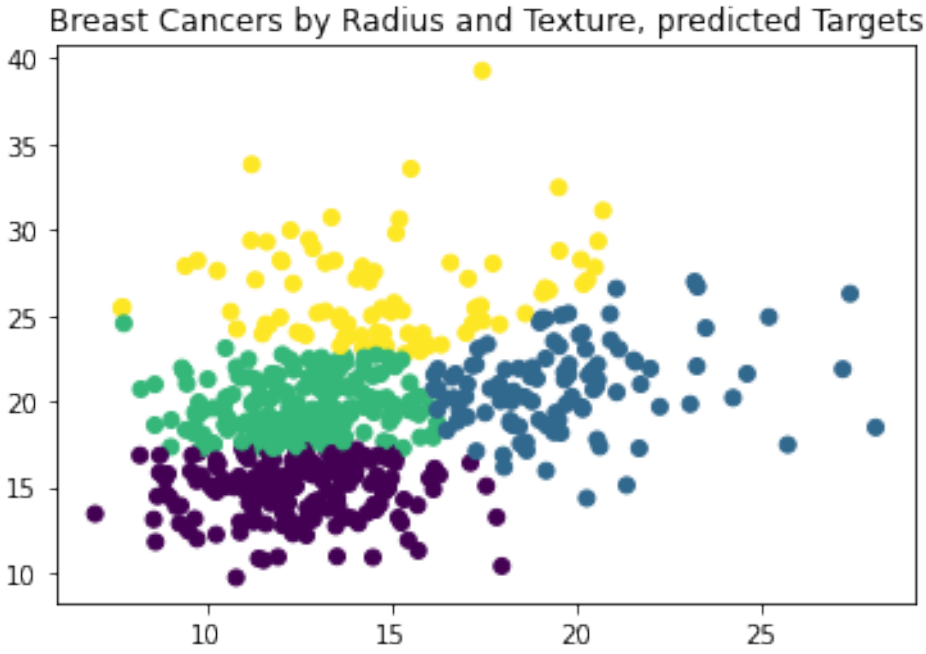

Figure 4: scatter plot of x (radius) against y (texture) of breast cancers using k-means computing 4 clusters

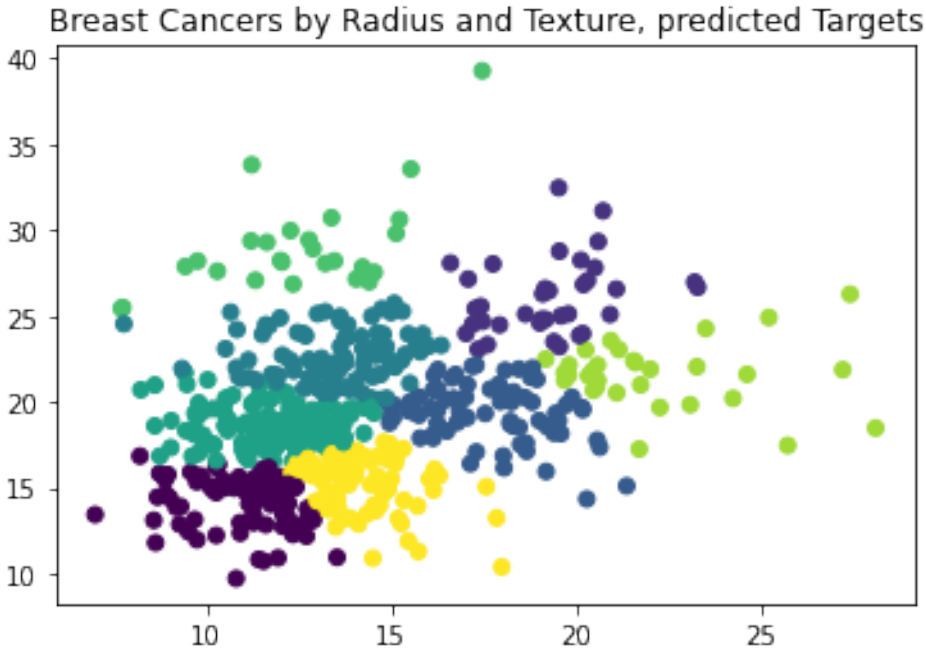

Figure 5: scatter plot of x (radius) against y (texture) of breast cancers using k-means computing 8 clusters