Classification Using Sklearn breast cancer Data

We use the breast cancer data set of sklearn to classify malignancy and benign growth using various parameters.

The data was read in from sklearn. A quick exploration showed that the file was a dictionary with eight keys including a description and a large data portion as well as a target key. We read in the description to 569 instances with 30 classes and numeric predictive attributes including radius, smoothness, texture, area etc. We obtained the yes/no class for cancer from the target section and concatenated that to plot the correlation with different parameters. The no/yes values had approximately a 200/350 distribution. We observed that there was a large overlap of parameter values in the final two classes in many instances.

To prepare the data, we ran an isnull on it and found that the data was clean.

To find the proper training method, we ran the test and train sets on four different classifiers: support vector machines, k nearest neighbors, random tree and random forest. We ran a grid search on support vector machines and found that the optimal value of c was 100. We ran k nearest neighbors for k values between 1 and 40 and we ran random forest for tree number between 50 and 500.

Support vector machines had an accuracy of .94 for the best parameters.

The highest accuracy for k nearest neighbors was .984 at k = 10.

The highest accuracy for random forest was .974 at tree number = 170.

Decision tree only had an accuracy of .92.

Hence, k nearest neighbors appeared to be the best classifier if k is chosen to be around 10.

We also printed the confusion matrix for all the classifiers and found out that the F1 score varied from .9 to .97 and was consistently lower for no cancer, the category that had fewer data. In the context of cancer prediction, false negative can cause serious health problems. Hence, recall should be high. k nearest neighbors had a high recall value for k = 6 and for both cancer and no cancer (.95 and .99).

However, it is to be noted that many of the parameters are highly overlapping and we have only 569 instances here. I would suggest adding more data and more parameters so the disparity in statistics between cancer and no cancer data is improved to make the model more predictive.