Noah Saffie

Final Proposal

The problem I will be solving is the classification of breast cancer into either benign or malignant. Breast cancer is a major concern in this world. As with all forms of cancer it is a serious condition that while often treatable if detected early can become deadly very quickly. Hopefully, through this project more insight could be gained about the classification of breast cancer, that may be through finding a highly accurate algorithm for the classification. Also, it could be through some data exploration of the dataset that might find some useful information to consider about the attributes.

The dataset I will be using is “Breast Cancer Wisconsin (Original) Data Set” by Dr. William H. Wolberg, published on the UCI Machine Learning Repository website. The dataset has nearly 700 instances, with 10 different attributes as follows: Sample code number, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses, Class. Although there is never such a thing as enough data, until you encapsulate the entire population, I believe there is enough data here to make meaningful insights possible.

Since the dataset has 700 instances of data, which can be a considerable amount for medical data, and the data has already been preprocessed, in that we don’t have raw images, but are given rankings (1-10) of all attributes but Sample code number, and Class. This ranking does provide some issues with the dataset as well, since the data values are very discrete, and a small range it does take away some of the ability for correlations to exist, and thus the ability of machine learning algorithms to have more refined and accurate results. Typically, you would want to have data that is either binary, categorical, or most preferably real-valued which a high degree of precision to allow nuances to be seen. This dataset has two branch datasets, one from diagnostic purposes, and one for prognostic purposes, those datasets although smaller in size, are real-valued, and have a much higher attribute count.

The plan is to use many different forms of visualizations on the dataset, to see what insights I might be able to glean from it. One such visualization will be a pair-wise correlation table that should the correlation values for each pair of attributes. This can be helpful for seeing what kind of relationships exist between the attributes, and can allow you to potentially remove certain attribute if they correlate highly with other non-target (non-Class attributes in our case), since if one can reasonably predict the other it can be better to remove redundant attributes, and simply the model (often significantly). This can also help you get an idea about what attributes most closely affect the target class, however this data should be assumed to be all-knowing, since there can and often is more complicated relationships that require complex algorithm to discover. Following that I will try to apply one or more machine learning algorithms to the dataset and see what kind of accuracy I can come up with. I am not sure what algorithm(s) I may attempt, but most likely one of them will include Naïve Bayes’ a simplistic Bayesian model used for classification tasks.