**Machine Learning Engineer Nanodegree**

**Capstone Project**

Bora Pajo, PhD

April 6, 2017

**I. Definition**

***Project Overview***

According to the Centers for Disease Control and Prevention (CDC) breast cancer is the most common type of cancer for women regardless of race and ethnicity (CDC, 2016). Around 220,000 women are diagnosed with breast cancer each year in the United States (CDC, 2016). Although we may not be aware of all the factors contributing in developing breast cancer, certain attributes such as family history, age, obesity, alcohol and tobacco use have been identified from research studies on this topic (DeSantis, Ma, Bryan, & Jemal, 2014).

***Problem Statement***

This project focuses in investigating the probability of predicting the type of breast cancer (malignant or benign) from the given characteristics of breast mass computed from digitized images. The cases provided, are cases diagnosed with some type of tumor, but only some of them (approximately 37%) are malignant. This project will examine the data available and attempt to predict the possibility that a breast cancer diagnosis is malignant or benign based on the attributes collected from the breast mass.

***Dataset and Inputs***

The characteristics of the cell nuclei have been captured in the images and a classification methods which uses linear programming to construct a decision line. The dataset is published by Kaggle and taken from the University of California Irvine (UCI) machine learning repository. The data is taken from the Breast Cancer Wisconsin Center. It includes ten (10) attributes taken from each cell nucleus as well as ID and the diagnosis (M=malignant, B=benign). The dataset has 570 cases and 31 variables.

***Evaluation metric***

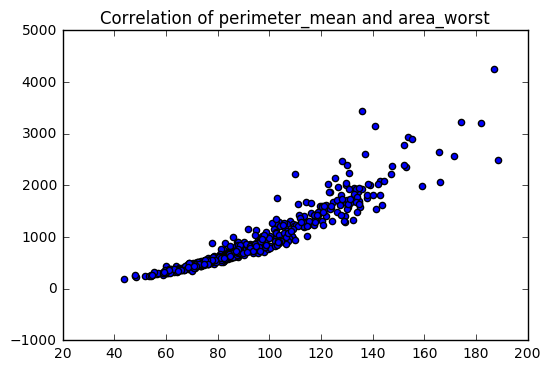
F1 score is a measure of accuracy or the ratio of the data that was accurately predicted. The closer the F-score is to 1 the best the prediction is and the closer to 0 it is, the worse the prediction. F-score considers the true positives and the true negatives, and is best used when comparing various classifiers as I am proposing to do in this dataset. From the literature, I reached the conclusion that F-score is the best evaluation metric to be used for this type of classification problem. The formula for the F1 score from the sklearn documentation is F1 = 2\* (precision \* recall) / (precision + recall).

**II. Analysis**

***Data Exploration and Visualization***

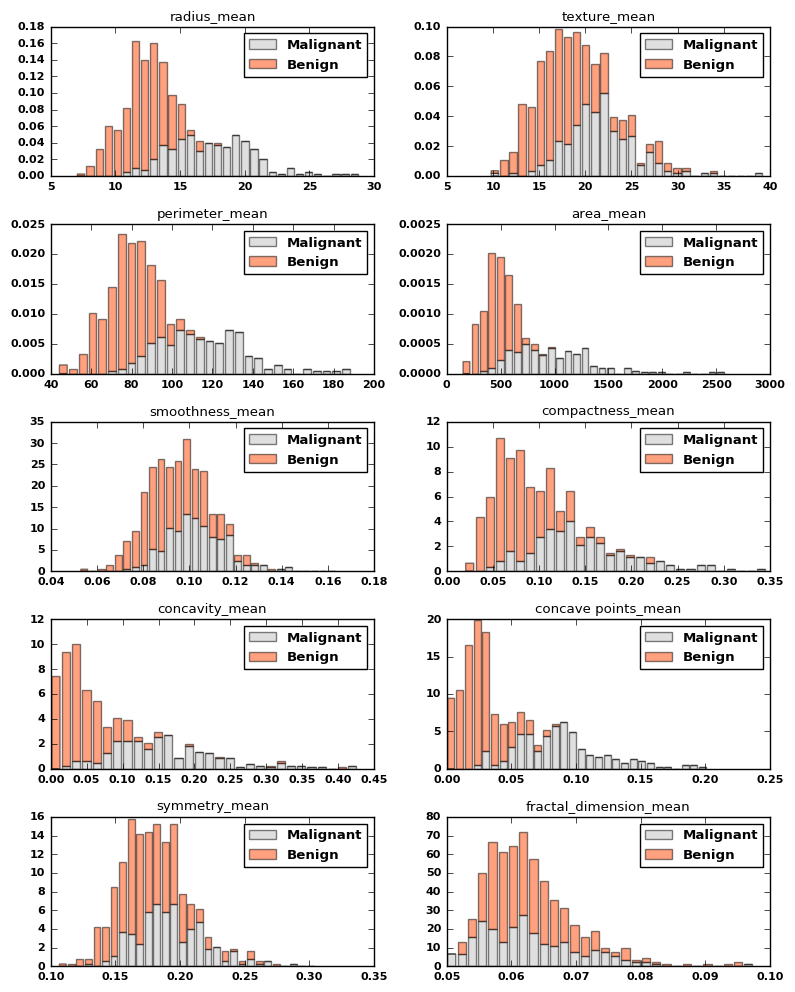
The dataset is taken from the UCI machine learning repository. It includes 569 cases of women diagnosed with cancer. There are 32 variables in the dataset (31 variables if we exclude the ID variable). Each variable includes information of the cell nucleus in the images captured. There are only 10 such attributes expressed in these 31 variables. These 10 attributes are: 1) radius, 2) texture, 3) perimeter, 4) area, 5) smoothness, 6) compactness, 7) concavity, 8) symmetry, 9) concave points, and 10) fractal dimensions. The dataset includes 212 cases diagnosed as malignant, and 357 cases diagnosed as benign. The ratio of malignant cases is 37.25%. Variables that relate to the same attribute are highly correlated with each other (as expected). Figure 1 below shows one of these examples where the perimeter mean and radius worst are highly correlated with a Pearsons’ r =.97) which means that the correlation of determination is r2=.94. This correlation of determination indicates that 94% of the changes in perimeter mean are explained by changes in the radius worst.

Figure 1: Correlation of perimeter mean and area worst



In addition, we can see that the malignant cases have higher values in almost every attribute when compared to benign cases. The averages of radius, perimeter, smoothness, texture, area, concavity, symmetry, concave points, and fractal dimensions of almost every attribute depict a visible difference in values where the malignant cases have much higher values. It is perhaps a sign of how spread out the tumor is compared to benign cases. Figure 2 show all these averages.

Figure 2: Stacked charts of the attributes for both cases



***Benchmark Model***

The highest accuracy the better the model will be predicting whether a person diagnosed with breast cancer has a benign or malignant tumor. To form a solid idea on the benchmark models, I perused the research literature on the same topic – predicting the type of cancer from computer-generated images. One study that uses mammogram images of 200 cases and utilizes both k-nearest and SVM methods to classify the type of cancer achieved specificity of 92.10% with standard deviation of 2.75 and *accuracy levels of 92.16%* with standard deviation of 3.60 (Zhang, Wang & Yang, 2016). Another study that suggests the use of magnetic resonance imaging (MRI) instead of ultrasound to detect the type of breast cancer among diagnosed women, utilized a sample of 110 lymph nodes from pre-operative MRIs. They *achieved an accuracy level of 79.1%,* and was considered much superior to the ultrasound (Chung, Hyun, Kim, Gweon, Kim, Ryu, & Son, 2014). A third study that also compares the accuracy levels of ultrasounds and mammography, found that mammography images are a better predictor than ultrasounds at accuracy levels of 90.7% (Tozaki & Fukuma, 2011). Judging from these three different studies on this similar topic, I calculated an average of accuracy levels of 87.32%. Therefore, my benchmark levels will be 87% or higher.

**III. Methodology**

***Data Preprocessing***