Machine Learning and Breast Cancer Detection

Data Domain

I come from a healthcare background, so I sought to use machine learning algorithms to aid in prediction for various healthcare diseases and diagnoses. Breast cancer continues to be a significant culprit of morbidity and mortality, even today with our current medical advances. The data that this dataset consists of was actually published in a medical journal article and data was collected at the University of Wisconsin (Wolberg, M.D. *et al.*¸1995). While this source is somewhat dated, my main goal with this project is to demonstrate that machine learning algorithms could be used to assist in cancer detection as adjunct to physician expertise.

It is no secret that our healthcare is cumbersome, overpriced, and our outcomes are unsatisfactory when compared to our other peer countries. In a data-driven world, with the vast amounts of personal health information and data available in electronic health records, the sky is the limit. Could machine learning algorithms be used to improve diagnosis, save lives, and prevent suffering? Up until this point, we have only been limited by data availability. Doctors are fallible just as humans and mistakes and misdiagnoses do happen. There are multiple ways that healthcare can use machine learning from customer service to cancer detection (Sarkar, 2020). Some are using various machine learning algorithms to detect DNA changes that could predict cancer (Ippolito, 2019). Breast cancer detection is yet another potential extension of this methodology. In fact, deep learning is being used to detect breast cancers through mammogram images (Shen *et al.*, 2019).

According to the CDC, breast cancer is the second most common cause of death due to cancer in women (CDC, 2020). Even with modern advances, approximately 39,000 women die from this disease annually (Esserman *et al.*, 2020). Due to increasing adoption of screening mammograms for the appropriate populations, we have reduced mortality from breast cancer by almost 14% (de Gelder *et al.*, 2015). Breast masses are detected using mammography/ultrasound and if suspicious, many women will get a fine needle aspiration or core needle biopsy. This involves using a needle to sample the tissue in question and observing it under a microscope to determine if it is has benign or malignant features. Even with a fine needle aspiration, this still cannot detect invasive cancers and further surgical care may be necessary (Joe *et al.*, 2020). Pathology determination can be subjective and one study estimates that sensitivity of approximately 70% in determining tumors and is heavily dependent on training (Ljung *et al.*, 2001). Therefore, the question remains, can machine learning help us predict malignant tumors better?

Dataset Description

The dataset that I will be examining comes from the University of California – Irvine Machine Learning Repository. Many of these datasets have been made public for the use of machine learning algorithms to use machine learning algorithms for different types of data science problems.

The link can be found here and a detailed listing of the variables can be found below: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

This data was collected at the University of Wisconsin in 1995. The file is in a csv format in which microscopic images of Fine Needle Aspirates of suspicious breast tissue was digitized. There are a total of 32 variables with 570 subjects. They were examining suspicious masses in those without evidence of metastasis (distant spread of cancer to other parts of the body). The target variable is categorical and is coded as M for malignant and B for benign. There are no missing data in the CSV file. There are case identifiers but other demographic information such as age, co-morbidities, and family history are not available.

The variables examined include measurements of digitized images on both cell size and shape. These were already identified as either benign or malignant. The data aims to determine if these specific characteristics can be used to predict whether a tumor was malignant or benign.

Research Questions and Rationale

Looking at the dataset, the variables are measures that describe cell size and/or cell shape. On a basic level, cancer is detected under a microscope by looking at cell size, cell shape, as well as various features of the cell that could be predictive of malignant characteristics. In this example dataset, they used the same features and computed the mean, standard error, and the worst measurement from each of these variables to determine if a tumor was malignant or benign. Following is a list of questions I seek to answer:

- Are any of the variables highly correlated and redundant? Can we simplify the variables without reducing the predictive accuracy?
- Are measures of cell size or cell shape more important in determining malignancy? Is it a combination of the two?
- Are bigger cells more likely to be malignant?
- How does the cell shape like smoothness or compactness influence the prediction?
- Is the dataset unbalanced or significant outliers that need to be addressed?

Methods

Fundamentally, this is a classification problem. Given the different characteristics of each of our variables, can machine learning predict whether a tumor was malignant or benign? I plan to trial three different methods of classification including logistic regression, K-nearest neighbors, as well as a Random Forest method and compare the accuracy, precision, recall, and F1 scores between the three methods.

Logistic regression may be a good model in that it is simpler and determines probabilities of the predicted event occurring. In the healthcare setting, many medical diagnoses and treatment modalities are based on probabilities.

K nearest neighbors predicts a class of the target variable by looking at the geometric distance between the individual data points to arrive at the true classification. This is also a simpler method and may be more widely amenable to widescale implementation.

Finally, we will try a Random Forest model. A random forest model will be useful in that it breaks down data classification into individual decision trees and then uses a majority voting option with all of the decision trees considered to render a final object. These are more computationally intensive so may not be amenable to widescale implementation.

I will import the data into Python within a Dataframe, perform exploratory data analysis of the variables to look for unbalanced classes and covariance between the predictor variables, and then use the three methods previously discussed.

Potential Issues

One potential issue is that some the data are mostly mathematical measurements, so it is difficult to tell if they are all on the same scale. I will need to determine this in order to run the algorithms appropriately. The Random Forest model may not be as heavily affected by data measurements on different scale though KNN and Logistic Regression will require that the units are standardized. Without this standardization, the results would be erroneous.

Further, since the target variable is categorical, this will need to be One-Hot Encoded to a dummy variable in order to perform the mathematical analysis required for both KNN and Logistic Regression. Random forest models account for categorical variables appropriately.

One factor that must be considered is if the three methods above do not get high enough accuracy, precision, recall, or F1 scores. In that case, it will be necessary to rethink a strategy on what other potential machine learning algorithms could be used to accomplish the task.

Concluding Remarks

Data science has traditionally been used in corporations and technology-predominant organizations. Healthcare generates a staggering amount of data that can be used in machine learning algorithms. The fact of the matter is that even with excellent care, mistakes and bad outcomes still do occur. Breast cancer is a major cause of morbidity and mortality for women worldwide. Even with advanced screening techniques, 39,000 women still die each year from this disease. If an abnormality is detected on physical exam or mammography, typically the lesion is biopsied using a Fine Needle Aspirate (FNA) or some other form of biopsy. The sample is examined microscopically to determine if the cells are indicative of malignancy. The diagnosis can still be subjective, and mistakes do happen (both false positives diagnoses and false negative diagnoses). The question remains, can machine learning algorithms assist healthcare providers to diagnose tumors more accurately?

This dataset collected by digitizing images of FNA biopsies at the University of Wisconsin will be examined using three different machine learning algorithms to determine the predictive accuracy on whether a biopsy was benign or malignant. Predictive accuracy will be determined using accuracy, precision, recall, and F1 scores to rate which machine learning algorithm resulted in the best scores.

References:

Centers for Disease Control. (2020, June 08). Breast Cancer Statistics. Retrieved September 03, 2020, from https://www.cdc.gov/cancer/breast/statistics/index.htm

De Gelder, R., Heijnsdijk, E. A., Fracheboud, J., Draisma, G., & Koning, H. J. (2014). The effects of population-based mammography screening starting between age 40 and 50 in the presence of adjuvant systemic therapy. *International Journal of Cancer, 137*(1), 165-172. doi:10.1002/ijc.29364

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Esserman, L. J., MD, MBA, & Joe, B. N., MD, PhD. (2019, May 30). Diagnostic evaluation of women with suspected breast cancer. Retrieved September 04, 2020, from https://www.uptodate.com/contents/diagnostic-evaluation-of-women-with-suspected-breast-cancer?search=breast cancer diagnosis

Ippolito, P. (2019, September 12). Understanding Cancer using Machine Learning. Retrieved September 04, 2020, from https://towardsdatascience.com/understanding-cancer-using-machine-learning-84087258ee18

Joe, B. N., MD, PhD, & Esserman, L. J., MD, MBA. (2019, May 16). Breast biopsy. Retrieved September 04, 2020, from https://www.uptodate.com/contents/breast-biopsy?search=fna breast

Ljung, B., Drejet, A., Chiampi, N., Jeffrey, J., Goodson, W. H., Chew, K., . . . Miller, T. R. (2001). Diagnostic accuracy of fine-needle aspiration biopsy is determined by physician training in sampling technique. *Cancer, 93*(4), 263-268. doi:10.1002/cncr.9040

Sarkar, T. (2020, April 30). AI and machine learning for healthcare. Retrieved August 30, 2020, from https://towardsdatascience.com/ai-and-machine-learning-for-healthcare-7a70fb3acb67

Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., Mcbride, R., & Sieh, W. (2019). Deep Learning to Improve Breast Cancer Detection on Screening Mammography. *Scientific Reports, 9*(1). doi:10.1038/s41598-019-48995-4

Wolberg, W. H., M.D., Street, W. N., Ph.D., Heisey, D. M., Ph.D., & Mangasarian, O. L., Ph.D. (1995). Computerized Breast Cancer Diagnosis and Prognosis From Fine-Needle Aspirates. *Archives of Surgery, 130*(5), 511. doi:10.1001/archsurg.1995.01430050061010

Appendix A: Detailed Code Book for Dataset

Variables:

Id: Number signifying unique samples (Integer)

Diagnosis: M for Malignant, B for Benign (Categorical) – the Target

Radius_mean: Mean of distance from center to points on the perimeter of tumor cell, cell size measure (Float)

Texture_mean: Mean of grey-scale values of image (Float)

Perimeter_mean: Expression of both cell size and shape (Float)

Area_mean: Mean area of cell size (Float)

Smoothness: Mean of cell smoothness and shape (Float)

Compactness: Mean of cell compactness and shape (Float)

Concavity_mean: Mean of cell concavity of image (Float)

Concave points_mean: Mean of concave points (Float)

Symmetry_mean: Mean of cell symmetry (Float)

Fractal_dimension_mean: Mean of fractal dimension, measure of cell shape (Float)

Radius_se: Standard error of distance from center to points on the perimeter of tumor cell, cell size measure (Float)

Texture_se: Standard error of grey-scale values of image (Float)

Perimeter_se: Standard error of both cell size and shape (Float)

Area_se: Standard error of cell size area (Float)

Smoothness_se: Standard error of cell smoothness (Float)

Compactness_se: Standard error of cell compactness (Float)

Concavity_se: Standard error of cell concavity of image (Float)

Concave points_se: Standard error of concave points (Float)

Symmetry_se: Standard error of cell symmetry (Float)

Fractal_dimension_se: Standard error of fractal dimension (Float)

Radius_worst: Worst measurement of distance from center to points on the perimeter of tumor cell, cell size measure (Float)

Texture_worst: Worst measurement of grey-scale values of image (Float)

Perimeter_worst: Worst measurement of both cell size and shape (Float)

Area_worst: Worst measurement of cell size area (Float)

Smoothness_worst: Worst measurement of cell smoothness (Float)

Compactness_worst: Worst measurement of cell compactness (Float)

Concavity_worst: Worst measurement of cell concavity of image (Float)

Concave points_worst: Worst measurement of concave points (Float)

Symmetry_worst: Worst measurement of cell symmetry (Float)

Fractal_dimension_worst: Worst measurement of fractal dimension (Float)