Machine Learning

Classification for Breast Cancer Diagnosis

Aditya Lahiri[1,*]

[1]Department of Electrical and Computer Engineering, 188 Bizzell Street, College Station, TX 77843.

Abstract

Motivation: Cancer is a broad term used to classify a spectrum of disease caused by the loss of cell cycle control in the body. One such manifestation of this phenomena is breast cancer. Over the recent years, the number women affected by breast cancer in the United States has risen along with the cost of healthcare. This trend calls for fast, affordable and reliable healthcare solutions that can be integrated with the currently available technology. With the technological advancements over the years, there is an abundance of omics data that can be integrated with modern technology to curb the cost of cancer diagnosis and make them more reliable. Machine learning is the perfect tool to bridge the gap between data and engineering. In this paper, we apply machine learning algorithms on real world data for breast cancer tumors to classify them as either malignant or benign. We furth/er cross validate our model with the real diagnosis of these tumors to evaluate the reliability of our model.
Results: Linear SVM outperformed all the classifiers proposed in this paper.
Availability: The datasets used are available at the UCI Machine Learning Repository, Kaggle and through University of Wisconsin, Madison http://ftp.cs.wisc.edu/math-prog/cpo-dataset/machine-learn/cancer/WDBC

Contact: alahiri2@tamu.edu

1 Introduction

In the U.S alone every one out of eight women will develop invasive breast cancer. That is about 12% of the women in the U.S [1]. It is estimated that 255,180 new cases of breast cancer will be diagnosed in women in the U.S [1]. Among women in the U.S. it was found that African American women and low income women have lower breast cancer screening and higher rate of late stage disease diagnosis than their counterparts [2]. The current methods of detecting breast cancer is through mammography screenings and MRI [3]. Though MRIs are much more accurate than mammography, they (MRI) can be quite expensive [3]. Therefore, there is a need to provide affordable, quick and reliable methods to diagnose breast cancer. With advancements in genetics and computer science a large amount of omics data is currently available and it will be sensible to incorporate this data in improving the accuracy of current diagnostic methods and bringing down the cost of diagnosis. Machine Learning, a subgroup within Artificial Intelligence uses the theory of statistics, probability and optimization to extract useful information from large, complex and noisy datasets [4]. Due to these qualities of Machine Learning it has been heavily applied to the area of cancer diagnostics and it is proven by PubMed statistics, which indicate more than 1500 papers have been published on the subject of cancer and Machine Learning. PubMed has further reported that a vast majority of these papers use Machine Learning as a tool to identify, classify, detect and differentiate between malignant and benign tumors. In this paper, we are interested in the classification of tumors as malignant or benign, therefore it will be very prudent to use Machine Learning algorithms to do so.

The data used for analysis in this project was provided by the University of Wisconsin, Madison. The data contains thirty features (measurements) regarding a tumor such as its radius, texture, perimeter etc. These features were computed from digitized images of a breast mass obtained through a fine needle aspirate (FNA). Our objective is to use this data in a machine learning algorithm that can predict the diagnosis with high accuracy. We will measure the accuracy of our algorithms using the actual diagnosis results that have been provided to us. Since we are given the actual results of the diagnosis we have a way to relate the ten features of our data to their respective diagnosis, this will help our algorithm to learn the trend of the data. Therefore, we will be using supervised learning algorithms like SVM, K Nearest Neighbors, DLDA and Logistic Regression to address the task of classification and our computational requirement will depend on the algorithm we choose to implement. In order to measure the performance of our classifier, we will use Kaggle's recommended metric of computing the area under the ROC curve. With the results we obtain, we will be able to recommend whether our algorithm is well suited and safe for the use of breast cancer diagnosis.

2 Methods

2.1 Summarizing the data

To begin the process of classification of the breast tumor data, we begin by summarizing the data and learning its properties. From the data, we were able to determine that there were 569 total data points among which 62% of the tumors were diagnosed as benign and the rest 48% were diagnosed as benign. This distribution of the entire data set has been present in fig 1. To understand the degree of correlation between each of the features in the dataset we created a correlation map. From this map, we were able to tell that some of the features such as the perimeter and radius, worst perimeter and the standard error in smoothness are highly correlated. Due to such instances of correlation among the given feature sets and the relative small size of our data we concluded that using regularization over feature selection will be a more prudent way to accommodate this aspect of our data set. The correlation map of the features has been presented in figure 2. Furthermore, we divided our data set into an 80:20 ratio of training and testing set respectively. The data set was partitioned randomly using MATLAB packages.

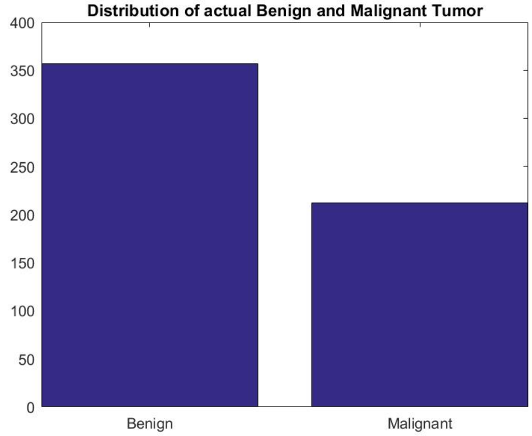

Fig. 1. : Distribution of entire dataset

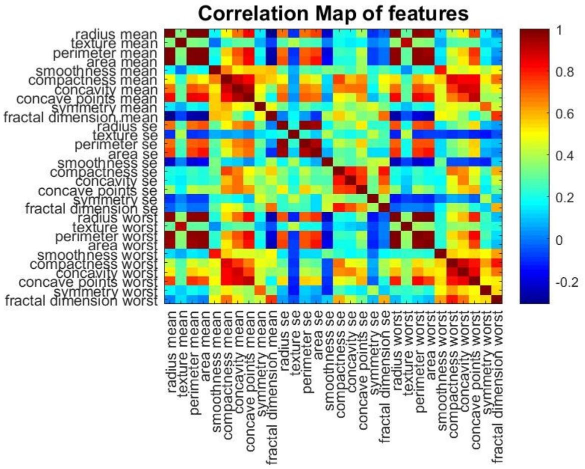

Fig 2.: Plot of correlation map of features

Since the data provided to us is being used for predicting the diagnosis of breast tumors, we wanted our classifier not to just evaluate a statistical score and return its class label (Malignant or Benign); but, also provide a probabilistic outcome so that we can evaluate to likelihood of the data belonging to either of the classes. Such a probabilistic treatment will provide us much deeper insight into the way our model behaves with the data. Additionally, we wanted a way to control the number of false positive outcomes and false negative outcomes in our model by adjusting the thresholding probability of our classifier rather than a hyperparameter associated with our classifier. The thresholding probability can be set at 0.5 in absence of expert knowledge. Such a probabilistic framework is provided by a very simple classifying algorithm known as logistic regression and we will use this classifier along with ridge regularization to accommodate for correlation in the feature set. Finally, we will compare the performance of our logistic regression classifier with other common classification algorithm using ROC curves.

2.2 Logistic Regression with Ridge Regression

Once the data set has been partitioned into the training and testing sets it is now ready for use in our classifiers. In logistic regression, for each data tumor in our data set, the class conditional probability of its label (Y = Malignant or Benign) is given by :

$$\mathrm{P(Y|X) = P =} \ \frac{1}{1+ exp(-(b_0+b_1X_1+b_2X_2+\cdots+b_nX_n)} \tag{1}$$

where X_1 ,..., X_n represent the feature vectors associated with the tumor and b_0 ,....., b_n represents the coefficients associated with each of the features. In order to find the coefficients associated with each of the features, we need to solve for the log odds and this is done as follows:

$$\mathrm{Odds} = \frac{\mathrm{P}}{1-\mathrm{P}} \tag{2}$$

$$\mathrm{logit(Odds)} = \ln(\frac{\mathrm{P}}{1-\mathrm{P}}) \tag{3}$$

$$\mathrm{logit(Odds)} = \mathrm{b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n} \tag{4}$$

Now that we have solved for the log odds, we can see that eq. 4 is exactly like the line of best from linear regression and we can solve the coefficients by minimizing the log odds cost function given by (5):

$$\mathrm{J(b)} = -\sum_{i=1}^{n} \mathrm{Y}_i \log(\mathrm{P}) + (1-\mathrm{Y_i})\log(1-\mathrm{p}) + \frac{\lambda}{2}\sum_{i=1}^{n} \mathrm{b_i} \tag{5}$$

The above equation represents the log odds function with the ridge penalties (represented by the second summation) that we need to minimize to obtain the coefficients for computing the conditional probabilities. When λ in the above equation is zero it represents the case when we have no ridge penalties or just the normal logistic regression case. However, when we have $\lambda > 0$, we try to minimize the log likelihood as much as possible while making sure that the coefficients *bi* do not become very large. After obtaining the coefficients we can compute the conditional probability for each of the tumors in the data set and threshold it at 0.5 to assign it to a particular class.

2.3 Diagonal Linear Discriminant Analysis (DLDA)

The diagonal linear discriminant classifier is a special case of the linear discriminant classifier where the pooled sample covariance matrix is considered to be diagonal. In general, DLDA performs better than LDA in small sample sizes. The DLDA classifier is obtained the same way as the LDA classifier, where the algorithm searches for best linear combination of features to separate the two classes, hence performing its own feature selection routine. The DLDA classifier can be derived as follows:

$$a_n = C^{-1}(\mu_1 - \mu_0)^T \tag{6}$$

$$b_n = -\frac{1}{2}\left((\mu_1 - \mu_0)C^{-1}(\mu_1 + \mu_0)^T\right) \tag{7}$$

$$C_{ii} = \frac{1}{n_1+n_0-2}\{(n_1-1)C_1 + (n_0-1)C_0\} \tag{8}$$

$$C_{ij} = 0 \tag{9}$$

In the above set of equations C represents the covariance matrix and C_0, C_1 represents the covariance in class 0 and class 1 respectively. The final classifier is given by:

$$\varphi_n(x) = \begin{cases} 1, & a_n X + b_n \geq 0 \\ 0, & otherwise \end{cases} \tag{10}$$

2.4 Linear Support Vector Machines (SVM)

The SVM classifier constructs a hyperplane by determining the best coefficients that separate the two classes the most in a binary classification problem. The distance between the data points closest to the hyperplane is defined as the margin. These points form the hyperplane and are called the support vectors. When the data is not linearly separable slack variables are optimized to obtained the maximal margin. It should be noted that while slack variables need to be made small to prevent overfitting but they should not be made too small as that can make the classifier underfit the data. The SVM classifier for a SVM classifier with general kernel is given as follow [5]:

$$\varphi_n(x) = \begin{cases} 1, & \sum_{i \in S} \lambda_i^* y_i k(x_i, x) + a^* > 0 \\ 0, & otherwise \end{cases}$$

where

$$a^* = -\frac{1}{n_m} \sum_{i \in S} \sum_{j \in s_m} \lambda_i^* y_i k(x_i, x) + \frac{1}{n_m} \sum_{i \in s_m} y_i$$

2.5 K Nearest Neighbor Rule (KNN)

The K nearest neighbor algorithm classifies a test point by finding K training points that are closest to the test point in consideration. The test point is then assigned the majority class label among these K training points. K is considered to be an odd integer in the application of KNN as it prevents the classifier from running into a tie. Our model KNN model was constructed with 10 fold cross validation with and the value of K was set to 5. The KNN classifier is given as follows:

$$\varphi_n(x) = \begin{cases} 1, & \sum_{i=1}^{K} \mathrm{I}_{\{Y=1\}} > \sum_{i=1}^{K} \mathrm{I}_{\{Y=0\}} \\ 0, & \text{otherwise} \end{cases}$$

3 Results

The four classifying algorithms, logistic regression with ridge penalties, DLDA, linear SVM and K (=5) nearest neighbor were applied on the training data to develop our classification models. These models were then used on the testing to obtain the predicted class labels. Figure 3 shows the ROC curves for each of the classifiers along with their respective AUC values.

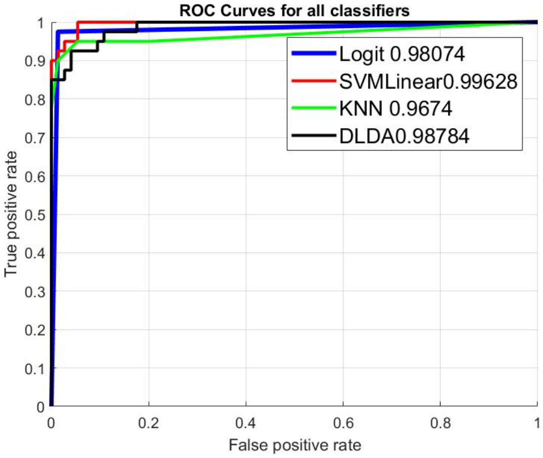

Fig 3.: ROC Curves for the Classifier and their AUC Values

From the above chart for ROC of different classifiers, it is evident that the linear SVM outperforms every other classifier, including and our proposed classifier i.e. logistic regression with ridge penalties. Although we can see that logistic regression came has similar performance. The high AUC values in all the classifiers tells us that the classifiers are sufficiently able to discriminate between the distribution of the two classes in the testing data.

Future Work

Now that we have concluded that linear SVM outperforms ridge penalized logistic regression, we would like to introduce L1 and L2 regularization for the Linear SVM model and also compare it with elastic net regularized logistic regression model. It is in our interest to improve our logistic regression model as it not only provides the class label but also provides a probabilistic framework for each of the class label decided. By evaluating the performance of these classifiers, we will be able to provide a more robust classification model for breast cancer diagnosis.

References

[1] Breastcancer, "U.S. Breast cancer statistics," in www.breastcancer.org, Breastcancer.org, 2017.

[2] A. S. O'Malley, C. B. Forrest, and J. Mandelblatt, "Adherence of low-income women to cancer screening recommendations," vol. 17, no. 2, Feb. 2002.

[3] L. Eggertson, "MRIs more accurate than mammograms but expensive," vol. 171, no. 8, Oct. 2004.

[4] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," vol. 2, Feb. 2007.

[5] U. Braga Neto, Lecture 10 SVM .