Predicting the presence of breast cancer using classification methods based on metabolic and anthropometric data

Farnaz Mohammadi
BE 188 final project report

March 20, 2019

1 Introduction

Breast cancer is one of the most common cancers in women, early and reliable detection could be a huge help in treatment of it. Hence, having a robust and reliable model based on non-invasive, cheap and easy-to-get data could be very valuable [1]. In this study, given some blood-measurements we intend to use some Machine Learning classification algorithms to predict whether a person has breast cancer or not. This model could provide a good biomarker of breast cancer patients to help diagnosis. Here, we use a data set from UCI repository which is measurements of some metabolic parameters that can be obtained from patient's blood, including concentration of Insulin, Resistin, HOMA, Leptin, Adiponectin and MCP-1, along with age and BMI as other features. These have been acquired from 116 people, including 52 healthy and 64 breast cancer patients. Three types of binary classification will be used and ROC curves, specificity and sensitivity as quality indicators will be reported.

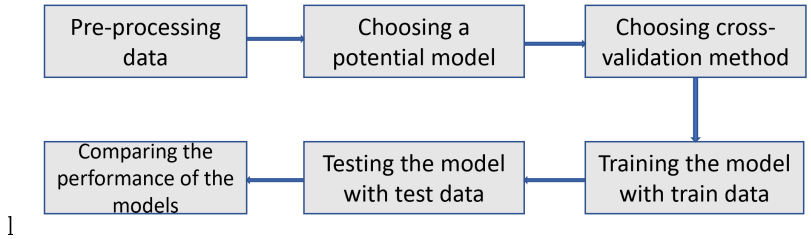
2 Problem definition

Given this set of data, we are interested to come up with a model that if we measure the aformentioned features of a new person that walks in to the clinic, the model could help us predict whether the person is most likely to have breast cancer or not with high confidence. To do that, we will use classification methods on the data we have to assign labels to a weighted combinations of the features. We need to identify and select for the best model which requires cross validating over a number of potential models and pick the one with the best prediction outcome. Using classification for prediction purposes is a widely used approach in predicting many many types of outcomes via data, especially in medical grounds. In our case, if we have more relevant features which could be acquired non-invasively and be cheap, could improve accuracy of prediction, and we may be able to increase the sensitivity and specificity of the model.

3 Methods

The binary classification methods we use here are: 1) Logistic Regression, 2) PLS-DA, and 3) Support Vector Machine (SVM). For each of those we will find the best parameters and find the sensitivity and specificity of the model to select for the models with highest sensitivity and specificity, and will plot the ROC curve (Receiver Output Characteristic) to evaluate the quality of classifier output on cross-validation – the paper has used Monte Carlo Cross Validation. AUC is the area under the ROC curve which is another measure of model evaluation and is desirable to have greater value for the AUC, because it means higher sensitivity and higher specificity. For all of our methods, we need to split the data into train and test sets to be able to evaluate prediction accuracy of the model, we could use a variety of cross-validation methods like K-fold, leave one out, random split based on test size, etc here 4-fold cross-validation has been used for Logistic Regression and SVM, and for PLS-DA randomly 15% of the data has been put aside for testing. Also, we need to normalize the data set after splitting the train and test data, since mostly we need to assume that the data is zero-mean and unit variance to use the functions.

I have used Python 3 in Jupyter Notebook for the programming. For most of the parts built-in function from sklearn package has been used, including SVM, LogisticRegression, and PLSRegression. For model evaluation, I have written a function "ROC" which calculates the sensitivity, specificity, and AUC based on the test data, and another function to plot the ROC curve with considering standard deviation. This function is not applicable to PLS-DA model, so I just plotted the predicted output test versus true output test to show its performance. In order to get the confusion matrix–which shows the number of false positive, false negative, true positive and true negatives– I used the most optimum threshold in the ROC curve, and set that threshold to build a predicted binary output, and used a built-in function $confusion_matrix$ in sklearn.metrics package.



l

Figure 1: flowchart of the process of modeling

3.1 Logistic Regression

This kind of model is employed when we have a binary output or just 2 labels, for here being healthy or having breast cancer that uses a logistic (or sigmoid) function to model a binary variable. We could change the number of variables to include in the model and see which ones are more influential than others and select them for our model. In python there is a ready function in sklearn package named "LogisticRegression" which takes in the data and the labels, and fits the model. It also returns a measure of performance namely "score", but I have used ROC and AUC to evaluate the model.

3.2 PLS-DA

Basically PLS-DA is PLSR for a binary output which takes the covariance of X and label data into consideration, and is appropriate for prediction. In this case the output vector is either "cancerous" as being 0 or “healthy" being 1. To find the best number of components to use, we could calculate Q2Y for each number of components, and pick the one with highest value. The formula for Q2Y that has been used here, is the one we had in homework 4 which is:

$$R^2 = 1 - \frac{\sum(y_{predicted} - y_{real})^2}{\sum(y_{predicted})^2 - \frac{(\sum y_{predicted})^2}{n}}$$

3.3 SVM

For Support Vector Machine model, we could try different kernels each with different parameters to see which one results in the best prediction. Here, three types of kernels have been applied to the model, including linear kernel, RBF kernel with varying gamma parameter, and polynomial kernel with varying degree, for C = 1000 in all cases. For each type, I got the AUC for every parameter and picked the highest one to plot the ROC, as shown in Figures 8 ,9 ,10.

4 Results

The results for the three types of models are as follows.

4.1 Logistic Regression

Logistic Regression results in more accuracy if it is given all of the variables in the data. Here the model is built first with all 9 variables, and then with 4 mentioned variables in the paper (Age,

BMI, Resistin, Glucose). It turned out that with 4 variables we get a more accurate result than all of the variables. As shown in the following figures, the AUC is 0.76 for all variables-model, and 0.83 for the model with 4 variables. I tried to vary the optimizer function, but if the random-state = 0 in train/test selection, then all the solvers give the same AUC. To better evaluate the model, ROC curve has been plotted and confusion matrix has been calculated. We could see in the notebook that the results are not that great, for instance the confusion matrix for all 9 variables and 4 variables are as follows respectively:

$$9 \text{ variable}: \begin{pmatrix} 44 & 20 \\ 14 & 38 \end{pmatrix} \quad 4 \text{ variable}: \begin{pmatrix} 49 & 15 \\ 13 & 39 \end{pmatrix}$$

We could see the performance of the model has improved with reducing the variables that are not very relevant, but still not very useful for clinical purposes. The following figures show the ROC of the model with 9 and 4 variables.

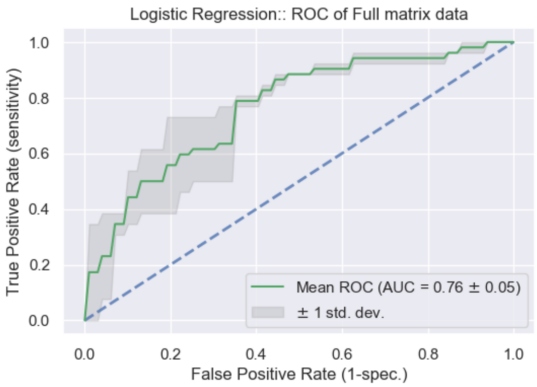


Figure 2: Logistic Regression with all 9 variables

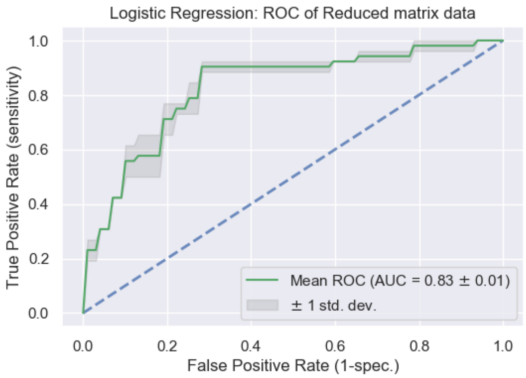
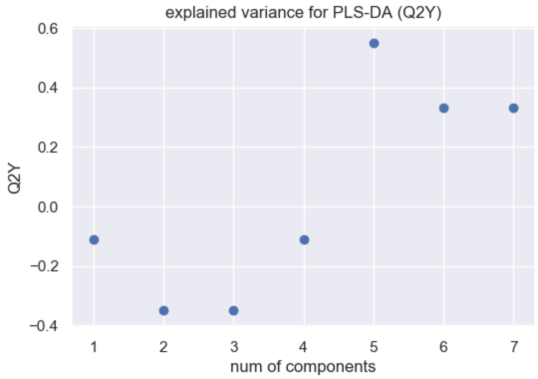


Figure 3: Logistic Regression with 4 variables

4.2 PLS-DA

Since there is a lot of overlap between classes with this type of classification –as shown in scores plot, we could see the model cannot predict the cancerous patients from healthy people on the test data with very high accuracy, as we see from the Q2Y plot, the maximum explained covariance is around 60% for 5 components. So we pick 5 as the best number of components, and build the model with that. The following are the scores and loadings plot for 5 components. To visualize the performance of the model, figure 7 shows the test data, the true and predicted labels, we could see there are a few points that they do not match, and those have been predicted wrongly.



l

Figure 4: explained covariance for PLS-DA

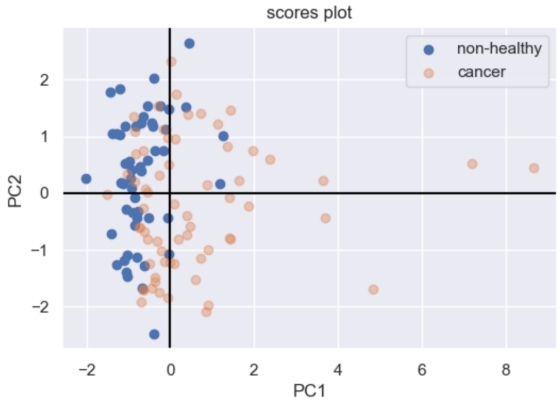


Figure 5: Scores plot for 5 principal components

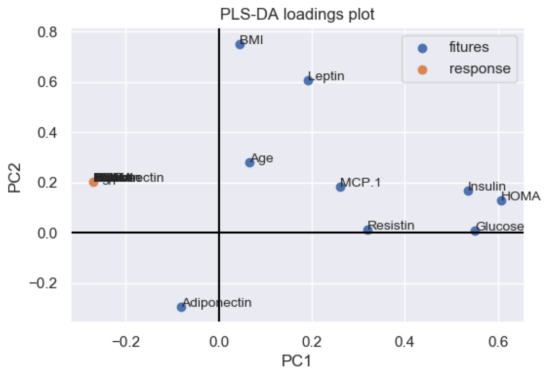
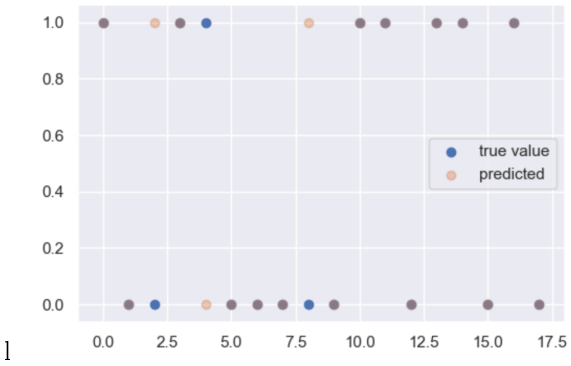


Figure 6: Loadings plot for 5 principal components



l

Figure 7: predicted vs. true test data for PLS-DA, 5 PC

4.3 SVM

This model has been applied to the 4-variable data, as used in the paper, but the sensitivity and specificity that the paper has got, is higher than what I got here. The best result was gained with RBF kernel in terms of AUC = 0.87 with $\gamma = 0.002$, and the confusion matrix:

$$\begin{pmatrix} 46 & 18 \\ 16 & 36 \end{pmatrix}$$

The AUC reached with linear kernel is 0.82 and and for the polynomial kernel the best degree was 3, and the AUC become 0.8.

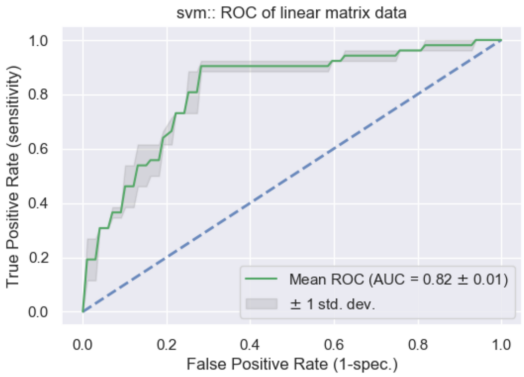


Figure 8: Linear SVM

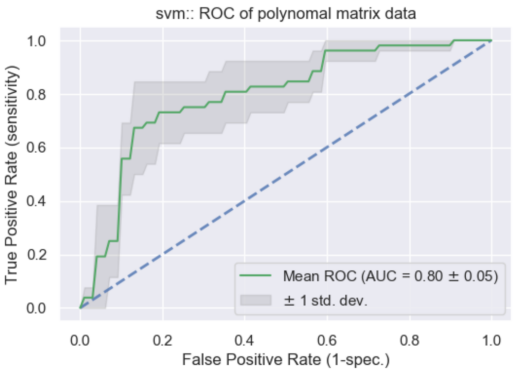


Figure 9: polynomial SVM with degree 3

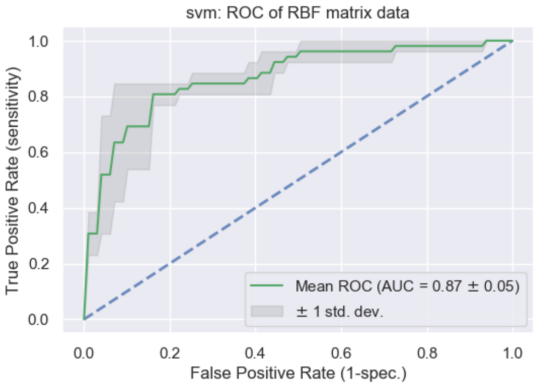


Figure 10: RBF SVM with gamma = 0.002

5 Discussion

In this study, we have used a dataset of some measurements of healthy and patients having breast cancer, and tried to find a data-driven model to help us predict if a person is likely to have breast cancer given those measurements or not. Since in rare disease like cancer is very important to provide high reliable test results, we need a test with very high sensitivity and specificity. In our analysis, the SVM model worked better that others, but models we got here, are not that reliable to be used in clinical applications, but if we could gather more relevant data and use other methods to confidently identify the variables that contribute more in getting a good prediction, we could improve our model.

References

[1] M. Patrício, J. Pereira, J. Crisóstomo, P. Matafome, M. Gomes, R. Seiça, and F. Caramelo, "Using resistin, glucose, age and bmi to predict the presence of breast cancer," *BMC cancer*, vol. 18, no. 1, p. 29, 2018.