Breast Cancer Detection through Supervised Classification Statistical Learning Methods

James Bui
Honors Business Administration, Wilfrid Laurier University, Waterloo, Ontario
Email: buix0840@mylaurier.ca

Received Spring 2021

ABSTRACT

This report presents a method for attempting to identify the presence of breast cancer using collected patient data. The set used is the UCI (Wisconsin) Breast Cancer Dataset holding records which contains 569 patient records from hospital cases. Each instance contains 10 quantitative features to describe the patient with a total of 30 features. We first prepared and pruned the data using LASSO regularization to identify nine relevant features. Next, we evaluated two classification models: Logistic Regression and Support Vector Machine (SVM) learning. The logistic regression achieved an F1 score of 0.9701 compared with SVM which only scored 0.9630. This suggests that linear models provide a better fit and more accurate predictions for breast cancer prediction tasks. However, with consideration towards evaluation metrics such as recall, the SVM model provided a better recall score of 0.9701 and may be better moving forward.

1. Introduction

Cancer is a globally important disease which affects the lives of nearly everyone and is the second highest cause of death worldwide. Breast cancer specifically has over 1.3 million cases and kills 450,000 people each year (Banin Hirata, B. K., Oda, J. M., Losi Guembarovski, R., Ariza, C. B., de Oliveira, C. E., & Watanabe, M. A.(2014)). Furthermore, early identification of breast cancer presence in patients significantly improves the likelihood of a successful recovery. A system which can automatically identify the presence of cancer in potentially vulnerable patients is valuable to healthcare professionals.

Classification is a common statistical learning method employed in the disease detection area of the health services industry. When brainstorming ways to apply statistical learning to practical settings this became a starting point from which to integrate the course knowledge learned over the semester. The main purpose of this project is to measure the extent to which 30 dimensions have an effect on Breast Cancer detection. Using the (Wisconsin) Breast Cancer dataset found from the UCI machine learning repository will train a classification model to predict a patient's tumor condition. It is through the statistical learning methods learned in class, namely logistic regression and SVMs that this report will propose a novel model for tumor diagnosis and determine features that best predict malignant breast cancer tumors. This research is valuable for the general public, of whom may be unaware of what to look for or lack the required capital to access their healthcare providers, benefitting 1 in 8 women who are diagnosed with this disease yearly (American Cancer Society, 2021).

2. Methods

This methods section will outline the processes used in this research project including the theory behind what was done, why methods were chosen, and why they work.

2.1 Data Cleaning

Data cleaning is the process of detecting inaccurate records and correcting them within a dataset. Most often this involves finding NA values, investigating the distributions of individual columns, and determining if there are similar clusters or patterns. Following this process improved decision making can take place and save researchers time and resources.

2.2 Column Distributions

Column Distributions are used to gain insight into the distribution of variables. This is typically understood through the use of histograms. Looking into these variables can help detect outliers which can severely effect results. Or rather, ensure normality of data and assess its suitability for regression.

2.3 One-Hot Encoding

One-Hot encoding is the process of converting a categorical variable into dummy variables denoted by 0's and 1's. 0 values represent the non-existence of a category, whereas 1 values represent their existence. This method allows categorical data to be expressed seamlessly through machine learning algorithms as some statistical learning methods cannot work directly with categorical data.

2.4 Box-Cox Transformation

Box-Cox transformations are used to transform skewed variables to a normal distribution. A Box-Cox Transformation otherwise referred to as a *power normal transformation* is an example of such a transformation applied on exponentially distributed data. This transformation is represented by the following formula mathematically:

$$f(y) = \frac{1}{(1 - I(f < 0) - \operatorname{sgn}(f)\Phi(0, m, \sqrt{s}))\sqrt{2\pi s^2}} \exp\left\{ -\frac{1}{2s^2}\left(\frac{y^f}{f} - m\right)^2 \right\}$$

Logistic regression models operate on assumptions such as (1) Linearity, (2)Independence, (3) Heteroskedasticity and (4) Normality. It is for this reason, when columns are exponentially skewed, this method is chosen to meet the regression assumption of normality required. Otherwise results would not be accurate.

2.5 Normalizing Features

Data normalization is a standard process through which data records are transformed to read similarly across all records. This is done to improve data redundancy and data integrity prior to being used in a machine algorithm. This process can be conducted in two ways (1) mean normalization (2) Min-Max normalization. Mean normalization normalizes data to make features have approximately zero mean. Min-Max normalization transforms the lowest value to 0 and the highest to 1. Values in between are thereby converted into decimals between 0-1

2.6 Logistic Regression

Logistic regression models are used to estimate the probability of a class existing such as pass/fail or malignant/benign. In this case, a logistic regression was used to predict whether a tumor diagnosis was benign or malignant. This model is mathematically represented by the following equation:

$$P = \frac{e^{a+bX}}{1 + e^{a+bX}}$$

Where P represents the target variable which varies between 0 and 1; *a* represents the y-intercept; *b* represents the slope coefficient; *X* represents the independent variable.

The reason why this method is heavily employed is because it is easy to implement, interpret and train data. Similarly, a disadvantage to this method is that it may lead to overfitting. To solve this feature selection should be used to select key variables.

2.7 Lasso Regression

Least Absolute Shrinkage and Selection Operator otherwise referred to as LASSO regularization was used to prune the data for variable selection. In LASSO regularization a factor of sum of absolute value of coefficients is added in the optimization objective function:

Objective = RSS + ***α * (sum of absolute value of coefficients)***

where *RSS* represents Residual Sum of Squares; *α (alpha)* represents different values that can balance RSS and a variety of coefficients

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

The following function is formally represented above and reads as for every value of α (alpha) there is some ‘s’ such that old and new cost functions will give the same coefficient estimates.

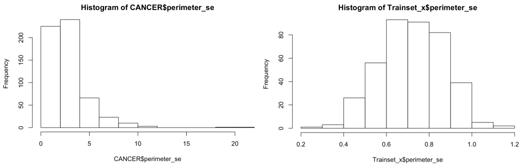
2.8 SVM

SVM otherwise referred to as Support Vector Machine Learning is a supervised learning method for classification. In this machine learning algorithm, a line is created called the maximum margin classifier that gives the most space between the nearest observation and line. The datapoints are subsequently classified depending on whether they fall above or below the line (Jakkula, V. (2006)).

3. Results

3.1 Data Preparation

The results from data cleaning showed no NA values to remove however, upon further inspection showed variables were mostly skewed exponentially. For example, this was most apparent for 4 variables which were later transformed using a box-cox transformation: radius_se, perimeter_se, area_se and fractal_dimension_se.



The following exhibit above shows an example of a column prior and post transformation. This box cox transformation successfully transformed the variable to assume a normal distribution. Following this a correlational list was generated and 8 highly correlated variables were removed: area_mean,radius_mean,area_worst,compactness_mean,perimeter_worst,compactness_se,concavity_worst and fractal_dimension_worst.

3.2 LASSO

Initially, a logistic regression model was built including all variables. The results showed that the model could not converge. The next step was to run the model without the highly correlated variables. This time the model was able to converge however, there very few significant variables. This meant the next step was to use a LASSO regression to select the most significant features and prune the data.

The LASSO regression pruned the data to identify **9** features that best predict whether a tumor is benign or malignant:: perimeter_mean, concavity_mean, concavity_se, concave points_se, fractal_dimension_se, radius_worst, texture_worst, smoothness_worst.

3.3 Logistic Regression

When running the logistic model without highly correlated variables against the test set, this confusion matrix was produced and showed high accuracy as well as recall:

Confusion Matrix:

	True	False
0	102	1
1	4	64

As observed in this confusion matrix there is an accuracy score of .9707 , recall score of 0.9412 and F1 score of 0.9624. In some cases, especially with cancer diagnosis recall can be an better evaluation metric than precision because the cost of false negatives can be life threatening.

Following the LASSO regression the model was pruned down to 9 features. This made the model less complex and as observed from the confusion matrix yielded stronger results with variables that were statistically significant.

Confusion Matrix:

	True	False
0	102	1
1	3	65

This confusion matrix shows an improved accuracy score of 0.9766, recall score of 0.9559 and an F1 score of 0.9701.

As such the final model proposed is:

Final_model <-glm(Trainset_y~perimeter_mean + concavity_mean + concavity_se + `concave points_se` + fractal_dimension_se + radius_worst + texture_worst + smoothness_worst, family = binomial, data = Trainset_x)

3.4 SVM

To compare the results of the logistic regression an SVM model was built and used to compare the results. In this case, a radial kernel was used which allowed us to compare the results between a linear method against a non-linear method.

Confusion Matrix:

	True	False
0	101	3
1	2	65

Summarized in this table is an Accuracy score : 0.9708, Recall score: 0.9701 and F1 Score: 0.9630.

4. Discussion

4.1 Linear vs. Non-Linear Methods

As observed from the final results the linear method provided the highest accuracy score of 0.9766 and highest F1 score of 0.9701. Since we are dealing with breast cancer detection even though the SVM had a lower performance as seen in its accuracy score of 0.9708 and F1 score of 0.9630 it may be better due to a lower false negative rate and higher recall score. In breast cancer detection recall may be an even more important metric as false negatives can cost patients their lives.

5. Conclusion

To conclude, we tried to predict breast cancer tumor diagnosis using patient data from the UCI (Wisconsin) dataset. We used a linear and nonlinear method – a logistic regression and SVM learning. The logistic regression provided the highest accuracy with an overall F1 score of 0.9701. The SVM model provided a F1 score of 0.9630 which was less accurate however provided better recall scores. Moving forward in for future work since we showed that non-linear methods work future models such as neural networks or tree-based models can be tested.

REFERENCES

American Cancer Society. (2021, April 26). *About Breast Cancer*. Retrieved from Cancer.org: https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html

Banin Hirata, B. K., Oda, J. M., Losi Guembarovski, R., Ariza, C. B., de Oliveira, C. E., & Watanabe, M. A. (2014). Molecular markers for breast cancer: prediction on tumor behavior. *Disease markers*, *2014*, 513158. https://doi.org/10.1155/2014/513158

Jakkula, V. (2006). Tutorial on support vector machine (svm). *School of EECS, Washington State University*, *37*.