Aroofa Mohammad

CS 4375.001 - Introduction to Machine Learning

Final Project

Due: May 8, 2022

Analyzing Machine Learning Techniques to Detect Breast Cancer

**Abstract**

Objective:

According to the American Cancer Society, around 287,850 new cases of women in the United States in 2022 were discovered to have metastatic breast cancer. A woman that leave the United States has a 1 in 8 chance of developing breast cancer at some point in her life. Breast cancer is also the second largest cause of mortality from cancer for women. Only lung cancer claims the lives of more women every year. A woman's chances of dying from breast cancer are roughly 1 in 39. The mortality rate decreased by 1% every year from 2013 to 2018. These reductions are thought to be the consequence of earlier detection of breast cancer through screening and increased awareness, as well as improved treatment options. (cancer.org)

The application of sophisticated data processing techniques to uncover connections between data items is known as data mining or machine learning techniques. These techniques or models are a widespread tool in the health-care business. This is because data mining can detect data patterns in enormous amounts of data. Regression, classification, and clustering are the three most common machine learning techniques. With the help of these models, there is a lot of hope in detecting breast cancer early to decrease the mortality rate.

Methods:

The machine learning techniques used in this study are Logistic Regression, Support Vector Machine (SVM), Decision tree, and K-Nearest Neighbors (KNN). All of these models will be tested on the database provided by the Wisconsin Breast Cancer and will generate an accurate estimate of patients' diagnosis of breast cancer.

Result:

Both Support Vector Machine and Logistic Regression would be considered the best model to use for breast cancer in this study. However, if only one had been taken into consideration Support Vector Machine would be an excellent choosedue to its simplicity and accuracy. Which will be further explained in the evaluation.

**Introduction**

The tumor attaches to the breast tissues and explodes into the lymphatic system. As a result, the tumor cells will be able to move and cause harm in other places of the body. There's a potential chance that the tumors will grow in other parts of the body. A benign tumor cell is one that does not cause harm, whereas a malignant tumor cell does. A benign tumor may not need to be diagnosed, while a malignant tumor does. By spewing forth enzymes, malignant tumors can cause injury to normal tissues.

More machine learning approaches are utilized to construct a model for the Wisconsin Breast Cancer Dataset. However, developing a complete model is challenging due to the non-linear nature of the breast cancer classification problem. The purpose of this research is to examine multiple efficient models developed in machine learning.

**Exploratory Data Analysis**

Dr. William H. Wolberg of the University of Wisconsin Hospitals in Madison provided this breast cancer dataset. A digitized picture of a fine needle aspirate (FNA) of a breast mass is used to compute features. They define the features of the image's cell nuclei.

Loading the Dataset:

Loaded the Wisconsin Breast Cancer Dataset into df. Therefore, can use and perform algorithms on data. To indicate the number of dimensions in the dataset array use shape on df, which resulted in (569,33).

Features:

To learn more about the dataset by print the info(). The columns below are features that the Wisconsin Breast Cancer Dataset provides.

1. Id number
2. Diagnosis (M = malignant, B = benign)

3-32. Features

* radius - the average of the distances between the center and the periphery points
* texture - standard deviation of the gray-scale values
* perimeter
* area
* smoothness - radius lengths vary depending on where you are.
* compactness - perimeter^2 / area - 1
* concavity - harshness of the contour's concave parts
* concave points - The number of concave contour parts symmetry
* fractal dimension – the "coastline approximation" - 1

Table

Description automatically generatedAs seen below there are two irreverent features, “id” and “Unnamed: 32”.

To clean up the data “id” and “Unnamed: 32” were dropped from the database because they were not need.

Target:

In the Wisconsin Breast Cancer Dataset, the target is the diagnostic. Determines whether the cancer is “m” for malignant meaning a high risk of cancer and “b” for benign meaning low risk of cancer. Below show the number of benign and malignant in the dataset by using value\_counts() on the diagnosis.

Pearson Correlation:

Chart, treemap chart

Description automatically generatedUsed the pearson correlation to obtain the correlations between feature-feature and feature-labels. Normally in the pearson correlation heatmap, you want to select only those features whose correlations are above 0.5. The closer to 0 indicates a lesser correlation (precise 0 indicates no connection) . The closer to 1 indicates a higher positive correlation. The closer to -1 indicates a stronger negative correlation.

Splitting the Dataset:

The data was divided into train and test sets. The train set is used to train the algorithm to understand the behavior of the data, and the test set is used to test the efficiency of our model. Predictions will be made using the columns that have been introduced to our model, as indicated by a variable "X" (features) in the code. The prediction or labels in the code are represented by a variable "y," which is the target variable that the features will predict.

Normalization:

Using feature scaling, we will be able to observe all of the variables through the same lens (same scale), bringing all of the values into the range [0,1].

**Model Analysis**

Support Vector Machine (SVM):

The Support Vector Machine is one of the most successful supervised machine learning algorithms (SVM). It may be applied to both classification and regression. SVM selects the hyperplane that achieves the greatest degree of separation between classes. If data can be separated linearly, SVM may provide the best results (Hard Margin). When data is not linearly separable, we simply need to loosen the margin to allow for misclassifications (Soft Margin). The C parameter, which has a direct influence on the hyperplane, can be used to control the amount of misclassification mistakes. The gamma function is used to provide weight to points that are near to the support vector. Changing the value of gamma, in other words, changes the form of the hyperplane.

Optimization of SVM parameters

The cost of misclassification on Train data is controlled by the C parameter. Smaller C indicates lower variance but more bias (soft margin) and lower miss-classification costs (less penalty). Lower bias and higher variance (hard margin) are associated with a larger C, which increases the cost of misclassification (more strict).

Smaller Gamma has a higher variance, a wider range of application, and a more generalized solution. Larger Gamma has a large variance and low bias, as well as a close reach and a higher weight for nearby data points.

Table

Description automatically generatedSquare

Description automatically generated with medium confidence

This confusion matrix represents:

* There are 143 people in this test set
* 87 of which are predicted to not have breast cancer. Of those, 0 people were classified as not having when actually they had (type 1 error)
* 56 of which are predicted to have breast cancer. Of those, 6 people were classified as having breast cancer when they did not (type 2 error)
* The SVM model correctly diagnosed tumors as malignant or benign with 96 percent accuracy.

By using this model we were able to reduce the type 2 mistake.

Decision tree:

The Decision-Tree method is one of the most extensively used machine learning algorithms for regression and classification issues. In general, the algorithm mimics human thought, making it popular and simpler to analyze inputs with a better comprehension of the issue statement. A decision-tree is a tree in this algorithm, with its node referring to an attribute, its link referring to a decision rule, and its leaf node referring to an output class. The goal is to develop a tree-like structure for the input characteristics, with each leaf producing a unique output.

Depth

The complete input training data is accepted as root when utilizing the decision-tree method for classification. When employing the decision-tree method to solve a classification problem, the depth analysis of the algorithm is also taken into account. Overfitting is an issue that happens when the algorithm's depth is increased. The number of nodes in the tree is represented by the size. Because each node of the decision tree technique is utilized to produce a binary classification, its size should be at least 2^(d+1)-1, where d is the tree's depth. The work's algorithm uses a depth of 3 as the best depth. Any binary classification issue can benefit from this optimum depth. The problem of over-fitting occurs at greater depths. The level of multi-class classification issues may also vary.

**Table

Description automatically generatedA picture containing application

Description automatically generated**This confusion matrix represents:

* There are 143 people in this test set
* 87 of which are predicted to not have breast cancer. Of those 2 people were classified as not having when actually they had (type 1 error)
* 56 of which are predicted to have breast cancer. Of those, 6 people were classified as having breast cancer when they did not (type 2 error)
* The decision tree model correctly diagnosed tumors as malignant or benign with 94 percent accuracy.

K-Nearest Neighbors (KNN):

The KNN technique is one of the most fundamental and extensively used supervised machine learning algorithms. A non-parametric approach to classification issues. The KNN technique exploits information about its neighbors to categorize output labels. KNN is a pattern recognition and classification method. KNN has the ability to predict the target class more precisely and simply. KNN is a type of lazy-learning algorithm in which the function is only approximated locally and the whole calculation is postponed until the classification process is finished. The problem is solved or categorized by the majority vote of its neighbors, where K is a small positive integer which is a real value. Since the KNN method is more sensitive to the local component of the input data, it is best to use for classification problems.

The importance of k's value cannot be overstated. The value of k is determined by the input data. If the value of k is big enough, the influence of noise on the output class is lowered. In this breast cancer study, the optimal value of k is 13. Raising k beyond this number may cause under-fitting. Even with its high classification efficiency, the inclusion of irrelevant characteristics, noise, in the input dataset will weaken the accuracy of the KNN method. As a result, feature selection and pre-processing are very important in this research.

**Table

Description automatically generated with medium confidenceA picture containing square

Description automatically generated**This confusion matrix represents:

* There are 143 people in this test set.
* 87 of which are predicted to not have breast cancer. Of those 2 people were classified as not having when actually they had (type 1 error).
* 56 of which are predicted to have breast cancer. Of those 6 people were classified as having breast cancer when they did not (type 2 error).
* The KNN model correctly diagnosed tumors as malignant or benign with 95 percent accuracy.

Logistic Regression:

Modeling the likelihood of a discrete result given an input variable is known as logistic regression. Popularly, logistic regression models have a binary result, such as true/false, yes/no, and so on. Multinomial logistic regression could be used to describe events with more than two discrete outcomes. Logistic regression is highly beneficial tool when it comes to identifying a new sample to fits into the best category.

First the data is splitting into training and testing set. Sklearn package has a method called GridSearchCV. The fit training data is used to predicted Logistic Regression with GridSearchCV. Then we optimize the model performance by automatically modifying hyperparameters(tuning them) using a few parameters in the function. The challenge of selecting the best characteristics for model training is quite tough. Feature selection is an important strategy for improving any model's performance. Primary component analysis may be used to choose characteristics; however it just gives us the principal components, which are produced by a combination of features. The attributes utilized to derive the primary components are not explicitly stated.Table

Description automatically generated

A picture containing square

Description automatically generated

This confusion matrix represents:

* There are 143 people in this test set
* 87 of which are predicted to not have breast cancer. Of those 0 people were classified as not having when actually they had (type 1 error)
* 56 of which are predicted to have breast cancer. Of those 6 people were classified as having breast cancer when they did not (type 2 error)
* The logistic regression model correctly diagnosed tumors as malignant or benign with 96 percent accuracy.

Evaluation

There are many ways to evaluate which machine learning techniques is best for breast cancer prediction. In this study we have already discussed confusion matrixes and F1. Bellow is a summary of the accuracy of each model.

Model: Accuracy on Test Data F1 Accuracy

Support Vector Machine (SVM): 95.80% 96%

Logistic Regression: 95.80% 96%

K-Nearest Neighbors (KNN): 95.10% 95%

Decision tree: 92.31% 92%

Taking the information above into account we can see that the support vector machine and logical regression techniques have the highest accuracy.

Logistic regression is used to solve classification problem whereas SVM can be used for both classification and regression. Logistic Regression uses various decision boundaries with various weights that are close to the desired spot. SVM tries to find the best distance between support vector and the line that separates the classes therefore lowering the risk of error of the data. Logistic regression is used in a more statistical approach and independent variables. While SVM works with the geometrical property of the data and is better with semi-structured data. Logistic regression is also more vulnerable to overfitting than SVM. SVM is faster in the kernel space, Both models do a great job detecting breast cancer but SVM is slight better because there are less chances of overfitting and it easier to used and apply.

Word Cited:

Edgar, Thomas W., and David O. Manz. “Logistic Regression.” *Logistic Regression - an Overview | ScienceDirect Topics*, 2017, https://www.sciencedirect.com/topics/computer-science/logistic-regression#:~:text=Logistic%20regression%20is%20a%20process,%2Fno%2C%20and%20so%20on.

Mir, Mahsa. “Case Study: Breast Cancer Classification Using a Support Vector Machine.” *Medium*, Towards Data Science, 18 Sept. 2020, https://towardsdatascience.com/case-study-breast-cancer-classification-svm-2b67d668bbb7.

Shetye, Abhini. “Feature Selection with Sklearn and Pandas.” *Medium*, Towards Data Science, 12 Feb. 2019, https://towardsdatascience.com/feature-selection-with-pandas-e3690ad8504b.

Society, American Cancer. “Breast Cancer Statistics: How Common Is Breast Cancer?” *American Cancer Society*, 2022, https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html#:~:text=About%20287%2C850%20new%20cases%20of,will%20die%20from%20breast%20cancer.

Wolberg, William H. “Breast Cancer Wisconsin (Diagnostic) Data Set.” *UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set*, 1 May 2022, https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic).