Predict whether the cancer is benign or malignant using Naive Bayes Classification

Md Ashraful Alam Hridoy

Computer Science and Engineering

East West University

2018-1-60-174@std.ewubd.edu

Md. Ferdous

Computer Science and Engineering

East West University

2018-1-60-098@std.ewubd.edu

Abstract

Breast cancer forms in the cells of breast. This cancer mainly seen in women, and very few in men. There are lots of sign and symptoms in breast cancer. For example- Size and shape changes, lumps feel different, color of the skin also changes. Cancer of the breast is a too common disease and its incidence is increasing.[1]

When breast cancer occurs few cells in the breast starts to grow very rapidly. These cells rapidly divide and started to form lump. Then it might start to spread other body parts.

At the beginning it starts to effect the Milk-producing ducts which is also known as invasive ductal carcinoma. Other than this the glandular tissue called lobules (invasive lobular carcinoma) can be affected at very beginning.

The only way to know the result of breast cancer is by pathology examination. Then we can be certain if it is benign or malignant. Data from the pathology examination can be classified using Naïve Bayes to predict whether the patient is benign or malignant. We achieved 94.7% accurate result from our experiment where we used Naïve Bayes classification.

Keywords: Breast Cancer features: Naïve Bayes classification;

Introduction

Breast cancer develop in the breast tissue. This disease can cause many effect such as change in breast shape, red skin on breast, bone pain, shortage of breath or yellow skin. Lack of physical exercise, obesity, alcoholism or ionizing radiation. There are number of risk factors such as sex, aging, estrogen, family history, gene mutations and unhealthy lifestyle, which can increase the possibility of developing breast cancer. [2]

Primary risk factor is become a female and old, other factor is genetic or taking alcohol, or excessive hormone or eating habit and obesity. Smoking also increase the risk for breast cancer. Long term smoker has 30% to 50% higher risk.

In the earlier time breast cancer was detected by mammogram which is X ray that detect the masses of sodium salt which was scattered around the mammary gland. now in modern days' cancer is diagnose by microscopic analysis. In the diagnosis process various things are measured for each cell nucleus.

Ten real-valued features are computed for each cell nucleus:

1. Radius (mean of distances from center to points on the perimeter)

2. Texture (standard deviation of gray-scale values)
3. Perimeter
4. Area
5. Smoothness (local variation in radius lengths)
6. Compactness (perimeter^2 / area - 1.0)
7.concavity (severity of concave portions of the contour)
8. Concave points (number of concave portions of the contour)
9. Symmetry
10. Fractal dimension ("coastline approximation" - 1)

Analyzing all these data of various patients' machine learning algorithm can answer weather the patient has M = malignant, B = benign.

In this study, Naïve Bayes method is proposed to classify the model. From the dataset 90% of the data will be used for training and rest 10% for test purpose.

Related work

Naïve Bayes main used for classification and prediction. There are a lots of research paper on the Naïve Bayes classification. This classification used in so many field for example in weather forecasting, many kind of data modeling, medical diagnosis, filtering out spam etc.

Method

Naive Bayes is a simple technique for constructing classifiers. It can be trained very efficiently in a supervised learning setting. Naïve Bayes classifiers is an algorithm that analyze data then process and categorize the data. it analyzes big data and based on that predict for any new analyzed data. [3]

Reverend Thomas Bayes research on the probability parameter of a binomial distribution and tried to find a way to compute a distribution for this. The Bayes theorem was based on this name and legacy. Bayes classifiers were introduce in 1960 from then on it was so popular for sort data based on the textual content such as email filtering. [4]

Naïve Bayes classifier algorithm is based on Bayes theorem. This algorithm predict the probability of any giver data based on the data it already studied on using Bayes theorem. It is mostly used for classification tasks. It is so fast and eager classifier for this reason it can be used to predicted in real time. Naïve Bayes classification algorithm is known because it use to predict multi class. If a target variable is given this algorithm can predict the probability of multiple classes. It is also use in sentiment analysis, spam filtering and text classification because it has higher success rate compare to other algorithms. Naïve Bayes classifier and collaborative filtering builds a recommendation system together it works based on data mining and machine learning to filter any unseen information and also predict if a user will like the recourse that given to them or not.

Naïve Bayes classifiers predict on an occurred situation based on the subject and the event. For example, if we have two coin that has head and tail. Then we toss it the chances of two tails is 1/4 same goes for two head at the same time. And the chance of getting at least one tail or one head is 3/4 but if we know the 1st coin is head then the chance of getting two head will be increased to 1/2. [5]

In Bayes theorem we have a conditional probability of event *c*, where *x* event has already occurred. For our example, 1st coin tossing will be event *x* and 2nd coin tossing is event *c*. Bayes theorem provides a way of calculating posterior probability *P(c|x)* from *P(c), P(x)* and *P(x|c).* Look at the equation below:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \ldots \times P(x_n \mid c) \times P(c)$$

$$P(X|c) = \prod_{i=1}^{n} P(X_i \mid c) \qquad [6], [4]$$

Above,

- P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

This is the formula used by naïve Bayes classifiers for giving probability from a larger data and build an accuracy from the data. [7]

Algorithm

For this report we used python as our coding language. Because it is one of the most convenient coding language for AI and machine learning. We used the build in naïve Bayes function for our report.

Firstly we imported numpy, pandas and sklearn. Numpy helps us using big array data and pandas use to work with read or assign data list with the code. Sklearn is use for importing other needed functions and naïve Bayes is one of the function that we need to be imported from sklearn. After that we called the data and started working on the data. 1st we checked whether the data is executable or not. After confirming that we put all the columns inside a value name features but hence we have one column for x we excluded the column from features. In out code the X column is diagnosis so we excluded that column from features.

For the testing and train we divided the data into 10% and 90%. To train we used 90% data from the full data sheet. And then we test other 10% for the accuracy. For this we took two variable x_train and entrain and insert train [features] and train ["diagnosis"] respectively and to test we took two variable x_test and y_test and did the same with rest. Now for the actual testing in Nb variable we called naïve Bayes function and then we use fit function to train the Nb with the chosen train data's. After that we use naïve Bayes classifier to predict the x_test and insert the values of that in a variable name y_pred.

```
x_train = train[features]
y_train = train["diagnosis"]

x_test = test[features]
y_test = test["diagnosis"]
```

```
Nb=GaussianNB()
Nb= Nb.fit(x_train,y_train)
```

```
y_pred=Nb.predict(x_test)
```

After all this we checked the accuracy of our prediction and checked the results from the data sheet. In order to do so we import accuracy_score from sklearn. Then we insert the value of accuracy to a variable name score and displayed the round value of our codes prediction.

```
from sklearn.metrics import accuracy_score
score = accuracy_score(y_test, y_pred) * 100
print("Accuracy using Naive Bayes: ", round(score, 1), "%" )
```

We have successfully execute our code and got a promising outcome from our code.

Experiment & result

Because we are studying how accurate the prediction of naïve Bayes classifier on a given database for any topic in this report. We had to find a suitable database for this report and since cancer is a big subject and it a dangerous topic especially some specific type of cancers. We choose to act on predicting breast cancer for out report. While looking for the database we had to think and find a proper database with all the possible factors that somehow can be related to causing the cancer. We had search for an efficient database and we found the used database from kaggle.com which we think is a suitable and good database for our research. It has 569 of total data rows and about 31 columns in total.

For the result we 1st executed non needed column from and started our program. We divided the full data into two part one is to train the program and other one part is to compare the data. For the test part the data studied all the columns data from test part except diagnosis which make sure weather the person has cancer or not. And so that we can compare the result of our code with the data of diagnosis. After studying the 90 percent of the data our code was able to predict with 94.7% accuracy. The accuracy of 94.7 percent shown from the program is good in all over. It was a promising result for the program.

```
In [28]: y_pred=Nb.predict(x_test)

In [29]: #Diplay
         y_pred

Out[29]: array(['B', 'M', 'B', 'M', 'B', 'B', 'B', 'B', 'B', 'B', 'B', 'M',
                'M', 'B', 'B', 'M', 'M', 'M', 'B', 'M', 'B', 'M', 'M', 'B',
                'B', 'B', 'B', 'B', 'M', 'B', 'B', 'B', 'B', 'B', 'M', 'B',
                'B', 'M', 'M', 'M', 'B', 'B', 'B', 'B', 'M', 'B', 'M', 'M',
                'M', 'M', 'M', 'B', 'B'], dtype='<U1')

In [30]: from sklearn.metrics import accuracy_score
         score = accuracy_score(y_test, y_pred) * 100
         print("Accuracy using Naive Bayes: ", round(score, 1), "%" )

         Accuracy using Naive Bayes:  94.7 %
```

Although our program of naïve Bayes preform the accuracy of 94.7%. It can be much better and much more efficient with more studied of train case. A program is more efficient and much more valuable when the prediction accuracy of that program is near to 100%. Normally when a program is 95 or more accurate we can say that this is a very good program. We can improve the result of our code with larger database and more training of the naïve Bayes. The larger the database and the more train the code gets it will be able to give more and more efficient and accurate result for new test case database. [8] It has been shown that with more and large train database a code become more accurate at predicting the result.

Conclusion and Future work

In this paper, a learning based early detection of breast cancer method has been proposed for any single data. Although the Naïve Bayes model that we worked on this paper can detect the breast cancer 94.7% accurately whether it is benign or malignant, but there is still room for further improvement in this paper

In the future we will do the research work with same database but with much more cleaning and filtering. We will also split the data into other combination for training and testing.

Reference:

[1] "Breast Cancer." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1543267/pdf/procrsmed00092-0009.pdf?fbclid=IwAR0Gr14nALncVjKaLwSfy05gge62D9BZKOud4xu-J_A5XODvL-l-B3KCE_8 (accessed May 24, 2021).

[2] Y.-S. Sun *et al.*, "Risk Factors and Preventions of Breast Cancer," *Int. J. Biol. Sci*, vol. 13, no. 11, pp. 1387–1397, 2017, doi: 10.7150/ijbs.21635.

[3] Y. H. Chang and H. Y. Huang, "An automatic document classifier system based on Naïve Bayes classifier and ontology," in *Proceedings of the 7th International Conference on Machine Learning and Cybernetics, ICMLC*, 2008, vol. 6, pp. 3144–3149, doi: 10.1109/ICMLC.2008.4620948.

[4] "Naive Bayes Classifiers Definition | DeepAI." https://deepai.org/machine-learning-glossary-and-terms/naive-bayes-classifier (accessed May 24, 2021).

[5] "Naive Bayes Classifier - Machine Learning Simplilearn." https://www.simplilearn.com/tutorials/machine-learning-tutorial/naive-bayes-

classifier (accessed May 24, 2021).

[6] I. Rish, “An empirical study of the naive Bayes classifier.”

[7] S. Ray, “Learn Naive Bayes Algorithm | Naive Bayes Classifier Examples.” https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/ (accessed May 24, 2021).

[8] R. Kohavi, “Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid.” https://www.aaai.org/Papers/KDD/1996/KDD96-033.pdf (accessed May 24, 2021).