Breast Cancer Classification Using Naive Bayes and KNN

By Sukanya De

COMPSCIX415.2-009

Introduction To Data Science- Fall 2020

Table Of Contents

Introduction

As per CDC, Breast cancer is the second most common cancer among women in the United States, most common being, the skin cancer. The average risk of developing breast cancer in adult women in the United States is 13%. The American Cancer Society's estimates for breast cancer in the United States for 2020 are:

- About 276,480 new cases of invasive breast cancer will be diagnosed in women.
- About 48,530 new cases of carcinoma in situ (CIS) will be diagnosed (CIS is non-invasive and is the earliest form of breast cancer).
- About 42,170 women will die from breast cancer.

In recent years, incidence rates have increased slightly (by 0.3% per year)
It is also the second leading cause of cancer death in women, only lung cancer kills more women each year. The chance that a woman will die from breast cancer is about 1 in 38 (about 2.6%).

Early Diagnosis and Treatment of Breast Cancer:

Breast cancer is sometimes found after symptoms appear, but many women with breast cancer have no symptoms. This is why regular breast cancer screening is so important.Different tests can be used to look for and diagnose breast cancer. Most common of them being:

- Mammograms
- Breast Ultrasound
- Breast MRI
- Newer and Experimental Breast Imaging Tests

According to the American Cancer Society, when breast cancer is detected early, and is in the localized stage, the 5-year relative survival rate is 99%. The early diagnosis and prognosis of cancer have become a necessity in cancer research, as it can facilitate the subsequent clinical management of patients. The ability of machine learning tools to extract key features from a complex dataset along with the speed of classifying and characterizing large volumes of data within a short span of time has led to a huge interest in the application of Machine Learning algorithms in the field of Biomedical research. Even though it is evident that the use of ML methods can improve our understanding of cancer detection and progression, an appropriate level of validation is needed in order for these methods to be considered in the everyday clinical practice. In the present work, I have analyzed Naive Bayes and KNN algorithms to accurately classify a breast tumor as a Benign or Malignant.

The current dataset is an extract from Kaggle
https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

Attribute Information:

The attributes in the dataset are as follows:

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

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

- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness (perimeter^2 / area - 1.0)
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- Symmetry
- fractal dimension ("coastline approximation" - 1)

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

All feature values are recorded with four significant digits.

Preliminary Data Analysis:

Metadata:

- Dataset Characteristics: Multivariate
- Attribute Characteristics: Real
- Attribute Characteristics: Classification
- Number of Instances: 569
- Number of Attributes: 32
- Missing Values: No

First few rows of data:

```
data.head()
```

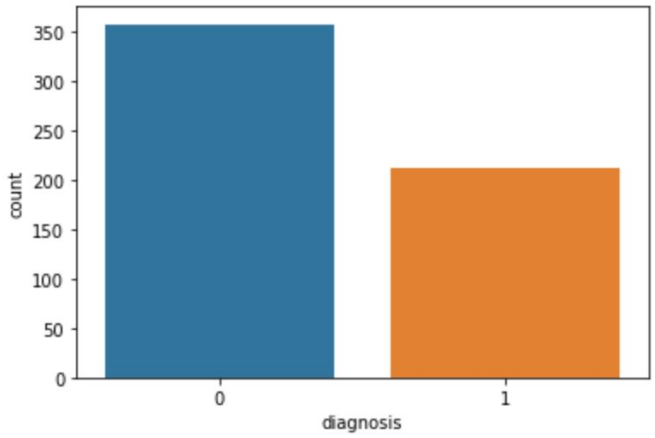
	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	te
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...	
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...	
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...	
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...	
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...	

5 rows × 33 columns

The dataset contains 350 Benign and 219 malignant records

```
sns.countplot(data['diagnosis'],label="Sum")

plt.show()
```



Objective and rationale of using Naive Bayes Classifier

In statistics, **Naive Bayes classifiers** are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naïve) independence assumptions between the features. They are among the simplest Bayesian network models.Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. A naive Bayes classifier considers each of these features to contribute independently to the probability, regardless of any possible correlations between the various features.

Steps of implementing the algorithm

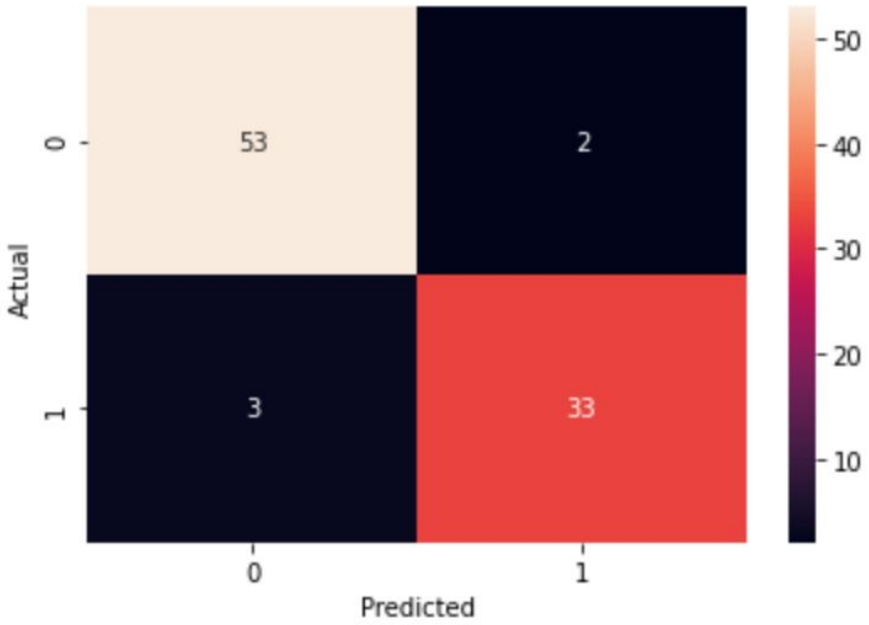
Following steps were performed for implementing Naive Bayes algorithm on the dataset:

- Data Preprocessing - Dataset has been analyzed to get an idea about the different fields available and probable correlation between the fields.
- It has been observed that the dataset does not contain any Null values.
- Features with a larger range of values can dominate the distance metric relative to features that have a smaller range, so feature scaling is important. The dataset has been normalized for easier comparison of various features.
- The fields ID and Unnamed has been removed since they were not adding values to the dataset
- The value of the categorical variable ‘diagnosis’ has been converted from String to Integers for easier application of the Naive Bayes and KNN algorithm
- The following features have been chosen for classification of a set of features into Malignant or Benign based on manual analysis:
    - Radius_worst
    - Symmetry_worst
    - Concavity_worst
    - fractal_dimension_worst
    - concave points_worst
    - Concavity_worst
    - Radius_worst
    - compactness_worst
    - Texture_worst
    - Perimeter_worst
    - Area_worst
    - Smoothness_worst

Data has been split into training and testing sets
Gaussian Naive Bayes have been applied on the dataset and the following accuracy has been achieved:

```
Accuracy: 0.945054945054945
Recall Score  :  0.9166666666666666
Precision Score:  0.9428571428571428

Text(0.5, 15.0, 'Predicted')
```



Interpretation of the results and prediction accuracy

Prediction accuracy has been very high at about 94%
A recall score of .91 indicates 9 percent of the malignant cases are misclassified as Benign. Our aim would be to minimize recall score since we do not want to misclassify any of the Malignant cases.

Performance improvement techniques and improved accuracy achieved

Following steps are performed for performance improvement:

1. Implement KNN classifier using the same features.
2. Use GridSearch CV to optimize KNN parameters to achieve better results.
3. Feature selection using Random Forest Regressor
4. Feature selection using Lasso Regressor

The accuracy obtained after implementing the above mentioned steps are as follows:

```
]: compare
```

]:

	Accuracy	Recall	Precision
KNN_ini	0.938596	0.893617	0.954545
KNN_opt	0.991228	0.978723	1
KNN_rf	0.956044	0.944444	0.944444
KNN_lasso	0.978022	0.944444	1
Naive Bayes	0.938596	0.93617	0.916667

Overall insights

KNN algorithms provide better results compared to Naive Bayes for the dataset analyzed.

K-NN classifier is a **supervised lazy classifier** which has local heuristics. Being a lazy classifier, it is difficult to use this for prediction in real time. The decision boundaries achieved with K-NN are much more complex than any decision trees, thus obtaining a nice classification. When solving a problem which directly focuses on finding **similarity** between observations, K-NN does better because of its inherent nature to optimize locally.

Naive Bayes is an **eager learning** classifier and it is much **faster** than K-NN. Thus, it could be used for prediction in real time.It assumes **conditional independence** between the features and uses a maximum likelihood hypothesis.

The current dataset being small, the execution time difference between KNN and Naive Bayes was not significant.

KNN model with features selected by LAsso Regression gives us the best result in this case. With Accuracy of 98%, which means that 98% of the data are categorised correctly. A recall percentage of 94 indicates that 6 out of 100 malignant cases are misdiagnosed as Benign by the model whereas a precision of 100 percentage indicates that all the benign cases are correctly classified by the model.

References

https://www.cdc.gov/cancer/breast/statistics/index.htm
https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html
https://en.wikipedia.org/wiki/Naive_Bayes_classifier
https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.302.8448&rep=rep1&type=pdf
https://medium.com/data-science-analytics/naive-bayes-or-k-nn-for-classification-60a4d92e7bab