Assignment – 3A

Prerak Patel - 000825410

# Data Description

Data set name:

Wisconsin Breast Cancer

Source: <http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29>

Features:

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. Below mentioned are the features in the data set used which has feature’s title and range of that feature.

|  |  |
| --- | --- |
| 1. Radius | 1. Texture |
| 1. Perimeter | 1. Area |
| 1. Smoothness | 1. Compactness |
| 1. Concavity | 1. Concave points |
| 1. Symmetry | 1. Fractal dimension |
| 1. Radius error | 1. Texture error |
| 1. Perimeter error | 1. Area error |
| 1. Smoothness error | 1. Compactness error |
| 1. Concavity error | 1. Concave points error |
| 1. Symmetry error | 1. Fractal dimension error |
| 1. Worst radius | 1. Worst texture |
| 1. Worst perimeter | 1. Worst area |
| 1. Worst smoothness | 1. Worst compactness |
| 1. Worst concavity | 1. Worst concave points |
| 1. Worst symmetry | 1. Worst fractal dimension |

Breast Cancer Data set is classified into two parts.

1. Benign
2. Malignant

There are total 569 data sets.

Out of which 357 are classified as “BENIGN” and rest of the 212 are “MALIGNANT”

# **Test** Results

## **Average accuracy and the average probability score for both correct and incorrect predictions.**

Below test results are computed running the Gaussian Naïve Bayes Classifier for 500 times with different training and testing split.

Training and testing split are shuffled every single time before the run and then separated as 80% training data and 20% testing data.

After running the algorithms for 500 times we derived this result.

Text

Description automatically generated

**Average Accuracy: 93.83%**

**Average Probability Score for Correct Predictions: 1.0**

**Average Accuracy Score for Incorrect Predictions: 0.94**

## **Naïve Bayes vs kNN vs Decision Trees**

Below test results are computed by running the all the algorithms for 500 times with different training and testing split.

Best Results for each algorithm,

**K-Nearest Algorithm**

A picture containing text

Description automatically generated

**Decision Trees**



**Gaussian Naïve Bayes Classifier**



# Discussion

1. **Why Naïve Bayes might have done better or worse than the K-Nearest or Decision Trees algorithms?**

* Above I have described the best results from all of the three algorithms and the clear winner is K-Nearest algorithm as the dataset that we are working with has very few outlier points and so K-Nearest Neighbour algorithm is able to produce the best results.
* Whereas, Decision Tree was able to perform well, but it would have performed extraordinarily if it would have given more dataset for training.
* **Naïve Bayes** was able to produce accuracy of 93.94% which doesn’t make this algorithm a worst performer among all three. **Naïve Bayes algorithm** is based on the estimation of probability that whether the testing data is classified as “Malignant” or “Benign”, so it was able to achieve this score where the probability was higher but when there was a low confidence, we found most of the mislabeled data.

1. **Discuss the average probability scores and give an example of how you might use the probability score to provide more useful information than simply a classification label.**

* Average Probability score is copied from above explanation.
* **Average Probability Score for Correct Predictions: 1.0**

**Average Accuracy Score for Incorrect Predictions: 0.94**

* Probability scores could be used more than just labelling the data.
  + It is used to compute the confidence in the labelling the data which helps us understand how important each class is in classifying the data.
  + It can help generate reliability array using which we can weigh classes differently to generate reliable data. For example, in our dataset we know that “Texture” feature is not very helpful instead “Worst Texture” feature is producing the higher probability score we can give “Worst Texture” feature more weight while training and testing the data.