Gaussian Naïve Bayes Classifier

Vikrant Bhati

1. Approach(es)

Naïve Bayes

The Naïve Bayes classifier is a family of classifiers based on Bayes' Theorem. It makes an assumption that features are independent of each other given the class label, which is known as the naïve independence assumption. Despite this simplifying assumption, Naïve Bayes performs well in many real-world applications, particularly in text classification, spam detection, and medical diagnosis. Naïve Bayes can be represented as:

Bayes Theorem:

$$P(C \mid X) = \frac{P(X \mid C).P(C)}{P(X)}$$

Where:

- P(C|X) = Posterior probability (the probability of class C given X)
- P(X|C) = Likelihood (the probability of X given class C)
- P(C) = Prior probability of class C
- P(X) = Evidence (the total probability of X)

Where Naïve Bayes after naïve assumption:

$$P(X \mid C) = P(X1 \mid C) \cdot P(X2 \mid C) \cdot \cdots \cdot P(Xn \mid C)$$

Where:

- P(X|C) = Probability of X given class C
- P(Xi|C) = Probability of Xi given class C

However, there are many scenarios where our input columns are continuous, and therefore for continuous features we use Gaussian Naïve Bayes.

Gaussian Naïve Bayes

The Gaussian Naïve Bayes classifier is a probabilistic model based on Bayes' Theorem, assuming that the features are conditionally independent given the class label and follow a Gaussian (normal)

distribution. It is used for classification tasks, especially when the input features are continuous. The probability density function for a feature given a class is modeled as:

$$P(x \mid C) = (1 / sqrt(2\pi\sigma\_C^2)) * \exp(-((x - \mu\_C)^2 / (2\sigma\_C^2)))$$

Where:

- μ_C = Mean of the feature for class C
- σ_C = Standard deviation for class C

Dataset

For the experiments, I used the Breast Cancer Wisconsin dataset from scikit-learn where the output column is a binary classification based on 30 input columns.

Output Column:

The target variable (output) is a binary classification:

- **0**: Benign (non-cancerous)
- **1**: Malignant (Cancerous)

Input Features:

- The dataset consists of 30 real, positive features that are derived from the Breast Cancer Wisconsin dataset.

Table 1: A summary of the breast cancer dataset

Classes	2
Samples per class	212(M),357(B)
Samples total	569
Dimensionality	30
Features	real, positive

Fig1: Summary of the breast cancer dataset

Further, the data was split into the training test data using *train_test_split* with 80% data being used for training and the remaining 20% for testing.

Adding Gaussian Noise

Zero mean Gaussian noise was added to see the impact on the accuracy score and confusion matrix. I used 6 noise variances of values: 50, 1000, 20000, 400000, 8000000 and 100000000.

Evaluation

In this experiment, the performance of the Gaussian Naïve Bayes classifier was evaluated using two key metrics:

Accuracy Score:

Accuracy is the proportion of correctly classified instances out of the total number of instances.

$$Accuracy = (True\ Positives + True\ Negatives)/Total\ Number\ of\ Predictions$$

Confusion Matrix:

A confusion matrix summarizes the classification results by showing the counts of true positives, true negatives, false positives, and false negatives.

$$[[True\ Positives \;\; False\ Negatives]$$

$$[False\ Positives \;\; True\ Negatives]]$$

2. Experimental Results

For the first experiment, I used the original data without any Gaussian noise and the output was:

- Accuracy Score: 0.9210526315789473
- Confusion Matrix: [[33 6] [3 72]]

After adding the noise, a zero mean Gaussian noise with a variance of 50 was:

- Accuracy Score: 0.8859649122807017
- Confusion Matrix: [[31 8] [5 70]]

After adding the noise, a zero mean Gaussian noise with a variance of 1000 was:

- Accuracy Score: 0.8596491228070176
- Confusion Matrix: [[27 12] [4 71]]

After adding the noise, a zero mean Gaussian noise with a variance of 20000 was:

- Accuracy Score: 0.8859649122807017
- Confusion Matrix: [[29 10] [3 72]]

After adding the noise, a zero mean Gaussian noise with a variance of 400000 was:

- Accuracy Score: 0.7807017543859649
- Confusion Matrix: [[23 16] [9 66]]

After adding the noise, a zero mean Gaussian noise with a variance of 8000000 was:

- Accuracy Score: 0.5526315789473685
- Confusion Matrix: [[12 27] [24 51]]

After adding the noise, a zero mean Gaussian noise with a variance of 100000000 was:

- Accuracy Score: 0.6403508771929824
- Confusion Matrix: [[10 29] [12 63]]

3. Discussion

The Gaussian classifier performed well on the clean dataset, achieving an accuracy of 92.11%. The confusion matrix for this setup was [[33 6] [3 72]]. However, as noise variance increased, the accuracy steadily declined, reflecting the classifier's sensitivity to noise. Initially, the classifier maintained a high accuracy at lower noise levels, but the performance degraded significantly with higher noise variance.

Trend Observations

Noise Variance: 50 to 20,000 – Accuracy fluctuated between 85.96% and 88.60%, indicating the classifier's robustness to moderate noise.

Noise Variance: 400,000 – Accuracy dropped to 78.07%, showing the beginning of a significant decline.

Noise Variance: 8,000,000 and 100,000,000 – The accuracy dropped to 55.26% and 64.04%, respectively. The high misclassification rate at these noise levels indicates that the noise overwhelmed the signal in the data.

This trend is evident in blow graph that showcase the accuracy vs noise.

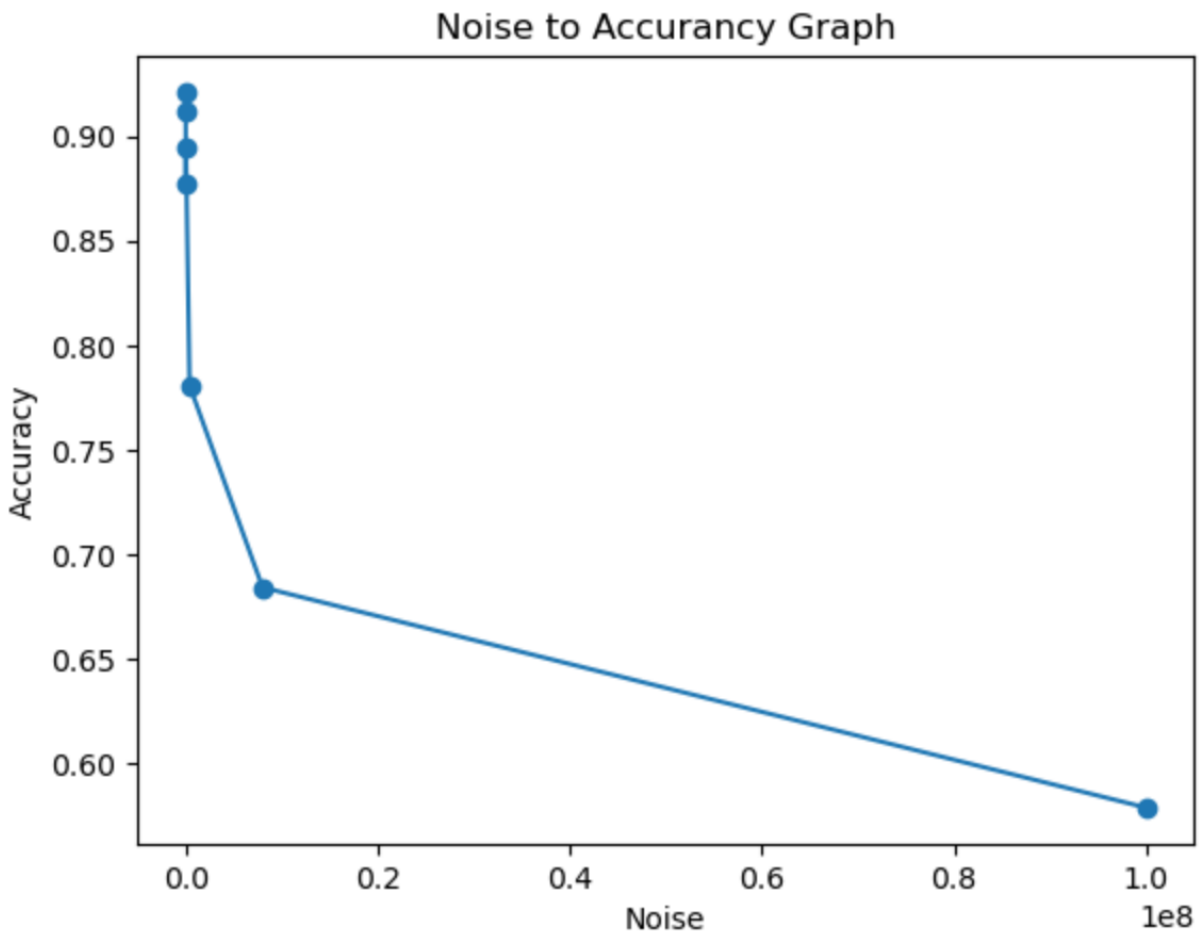

Fig2. Graph that showcases the accuracy vs noise

The confusion matrices at higher noise levels showed a marked increase in false positives and false negatives, indicating that the classifier struggled to maintain decision boundaries.

The results demonstrate a clear negative correlation between noise variance and accuracy. While the Gaussian Naïve Bayes classifier is reasonably robust to small amounts of noise, its performance deteriorates as noise becomes more dominant in the feature space.