Part A

Wednesday, February 7, 2024 10:33 PM

Step 1

class labels are either +1 or −1
and there are 5 features with 10 values / data points

$$p(y=1) = \frac{\text{number of 1}}{\text{total number of data points}} = \frac{4}{10} = \underline{\underline{0.4}}$$

$$p(y=-1) = \frac{6}{10} = \underline{\underline{0.6}}$$

Step 2

→ $\underline{\underline{x_1}}$

Frequency Table

x_1 \ y	-1	1
0	3	1
1	3	3

Likelihood Table

x_1 \ y	-1	1	$p(x_1)$
0	3/6	1/4	4/10
1	3/6	3/4	6/10
$p(y)$ →	6/10	4/10	

→ $\underline{\underline{x_2}}$

Frequency Table

x_2 \ y	-1	1
0	1	4
1	5	0

Likelihood Table

x_2 \ y	-1	1	$p(x_2)$
0	1/6	4/4	5/10
1	5/6	0/4	5/10
$p(y)$ →	6/10	4/10	

This is the case for zero-probability

⇒ Usind "<u>Laplace smoothing</u>" and adding 1 to numerator & 2 to denomenator for "y = 1", we get

★ ⇒ $P(x_2=0 \mid y=1) = 5/6 = 0.833$ $\quad P(x_2=1 / y=1) = 1/6 = 0.1667$

→ x_3

Frequency Table

x_3 \ y	-1	1
0	2	1
1	4	3

Likelyhood Table

x_3 \ y	-1	1	$p(x_3)$
0	2/6	1/4	3/10
1	4/6	3/4	7/10
$p(y)$	6/10	4/10	

→ x_4

Frequency Table

x_4 \ y	-1	1
0	1	2
1	5	2

Likelyhood Table

x_4 \ y	-1	1	$p(x_4)$
0	1/6	2/4	3/10
1	5/6	2/4	7/10
$p(y)$	6/10	4/10	

→ x_5

Frequency Table

x_5 \ y	-1	1
0	4	3
1	2	1

Likelyhood Table

x_5 \ y	-1	1	$p(x_5)$
0	4/6	3/4	7/10
1	2/6	1/4	3/10
$p(y)$	6/10	4/10	

For $x(1)$

$p(x_1=0/y=-1)=0.5$ $p(x_1=0|y=1)=0.25$

$p(x_1=1|y=-1)=0.5$ $p(x_1=1/y=1)=0.75$

For $x(2)$

$p(x_2=0|y=-1)=0.1667$ $p(x_2=0|y=1)=0.833$

$p(x_2=1|y=-1)=0.833$ $p(x_2=1|y=1)=0.1667$

For $x(3)$

$p(x_3=0|y=-1)=0.333$ $p(x_3=0|y=1)=0.25$

$p(x_3=1|y=-1)=0.667$ $p(x_3=1|y=1)=0.75$

For $x(4)$

$p(x_4 = 0 \mid y = -1) = 0.1667$ $\quad$ $p(x_4 = 0 \mid y = 1) = 0.5$

$p(x_4 = 1 \mid y = -1) = 0.833$ $\quad$ $p(x_4 = 1 \mid y = 1) = 0.5$

For $x(5)$

$p(x_5 = 0 \mid y = -1) = 0.667$ $\quad$ $p(x_5 = 0 \mid y = 1) = 0.75$

$p(x_5 = 1 \mid y = -1) = 0.333$ $\quad$ $p(x_5 = 1 \mid y = 1) = 0.25$

Step 3

Using Bayes Rule, we know that

$$P(A/B) = \frac{P(B|A) \times P(A)}{P(B)}$$

For $x(1)$

$x(1) = 0$

$$P(y=1 \mid x_1 = 0) = \frac{P(x_1 = 0 \mid y = 1) \times P(y=1)}{P(x_1 = 0)} = \frac{0.25 \times 0.4}{0.4} = 0.25$$

$$P(y=-1 \mid x_1 = 0) = \frac{P(x_1 = 0 \mid y = -1)\ P(y=-1)}{P(x_1 = 0)} = \frac{0.5 \times 0.6}{0.4} = 0.75$$

$x(1) = 1$

$$P(y=1 \mid x_1 = 1) = \frac{P(x_1 = 1 \mid y = 1) \times P(y=1)}{P(x_1 = 1)} = \frac{0.75 \times 0.4}{0.6} = 0.5$$

$$P(y=-1 \mid x_1 = 1) = \frac{P(x_1 = 1 \mid y = -1)\ P(y=-1)}{P(x_1 = 1)} = \frac{0.5 \times 0.6}{0.6} = 0.5$$

For $x(2)$

$x_2 = 0$

$$P(y=1 \mid x_2=0) = \frac{P(x_2=0 \mid y=1)\ P(y=1)}{P(x_2=0)} = \frac{0.833\times 0.4}{0.5} = 0.666$$

$$P(y=-1 \mid x_2=0) = \frac{P(x_2=0 \mid y=-1)\ P(y=-1)}{P(x_2=0)} = \frac{0.1667\times 0.6}{0.5} = 0.2$$

$x_2 = 1$

$$P(y=1 \mid x_2=1) = \frac{P(x_2=1 \mid y=1)\ P(y=1)}{P(x_2=1)} = \frac{0.1667\times 0.4}{0.5} = 0.1333$$

$$P(y=-1 \mid x_2=1) = \frac{P(x_2=1 \mid y=-1)\ P(y=-1)}{P(x_2=1)} = \frac{0.833\times 0.6}{0.5} = 0.9996$$

For $x(3)$

$x_3 = 0$

$$P(y=1 \mid x_3=0) = \frac{p(x_3=0 \mid y=1)\ P(y=1)}{P(x_3=0)} = \frac{0.25\times 0.4}{0.3} = 0.333$$

$$P(y=-1 \mid x_3=0) = \frac{p(x_3=0 \mid y=-1)\ p(y=-1)}{P(x_3=0)} = \frac{0.333\times 0.6}{0.3} = 0.666$$

$x_3 = 1$

$$P(y=1 \mid x_3=1) = \frac{P(x_3=1 \mid y=1)\ P(y=1)}{P(x_3=1)} = \frac{0.75\times 0.4}{0.7} = 0.428$$

$$P(y=-1 \mid x_3=1) = \frac{p(x_3=1 \mid y=-1)\ P(y=-1)}{P(x_3=1)} = \frac{0.667\times 0.6}{0.7} = 0.5717$$

For $x(4)$

$x_4 = 0$

$$P(y=1 / x_4=0) = \frac{P(x_4=0 | y=1)\ P(y=1)}{P(x_4=0)} = \frac{0.5 \times 0.4}{0.3} = 0.666$$

$$P(y=-1 / x_4=0) = \frac{P(x_4=0 | y=-1)\ P(y=-1)}{P(x_4=0)} = \frac{0.1667 \times 0.6}{0.3} = 0.3334$$

$x_4 = 1$

$$P(y=1 / x_4=1) = \frac{P(x_4=1 / y=1)\ P(y=1)}{P(x_4=1)} = \frac{0.5 \times 0.4}{0.7} = 0.2857$$

$$P(y=-1 | x_4=1) = \frac{P(x_4=1 | y=-1)\ P(y=-1)}{P(x_4=1)} = \frac{0.833 \times 0.6}{0.7} = 0.714$$

For $x(5)$

$x_5 = 0$

$$P(y=1 | x_5=0) = \frac{P(x_5=0 | y=1)\ P(y=1)}{P(x_5=0)} = \frac{0.75 \times 0.4}{0.7} = 0.4285$$

$$P(y=-1 | x_5=0) = \frac{P(x_5=0 | y=-1)\ P(y=-1)}{P(x_5=0)} = \frac{0.667 \times 0.6}{0.7} = 0.5717$$

$x_5 = 1$

$$P(y=1 | x_5=1) = \frac{P(x_5=1 | y=1)\ P(y=1)}{P(x_5=1)} = \frac{0.25 \times 0.4}{0.3} = 0.333$$

$$P(y=-1 | x_5=1) = \frac{P(x_5=1 | y=-1)\ P(y=-1)}{P(x_5=1)} = \frac{0.333 \times 0.6}{0.3} = 0.666$$

Step 4

For $x = (1, 1, 0, 1, 0)$

$\boxed{y=1}$

$$P(y=1 \mid x_1=1, x_2=1, x_3=0, x_4=1, x_5=0)$$

$$\propto P(x_1=1 \mid y=1) \times P(x_2=1 \mid y=1) \times P(x_3=0 \mid y=1) \times P(x_4=1 \mid y=1) \times P(x_5=0 \mid y=1) \quad P(y=1)$$

$$\propto 0.75 \times 0.1333 \times 0.25 \times 0.5 \times 0.75 \times 0.4$$

$$\propto 0.00375$$

$\boxed{y=-1}$

$$P(y=-1 \mid x_1=1, x_2=1, x_3=0, x_4=1, x_5=0)$$

$$\propto P(x_1=1 \mid y=-1)\, P(x_2=1 \mid y=-1)\, P(x_3=0 \mid y=-1)\, P(x_4=1 \mid y=-1)\, P(x_5=0 \mid y=-1)\, P(y=-1)$$

$$\propto 0.5 \times 0.833 \times 0.333 \times 0.833 \times 0.667 \times 0.6$$

$$\propto 0.0462$$

Since for $x' = (1, 1, 0, 1, 0)$

$$P(y=-1 \mid x') > P(y=1 \mid x')$$

$\Rightarrow$ The class is $\underline{\underline{-1}}$

For $x = (0, 0, 0, 0, 0)$

$\boxed{y=1}$ $P(y=1 \mid x_1=0, x_2=0, x_3=0, x_4=0, x_5=0)$

$$= \frac{P(x_1=0 \mid y=1)\, P(x_2=0 \mid y=1)\, P(x_3=0 \mid y=1)\, P(x_4=0 \mid y=1)\, P(x_5=0 \mid y=1)\, P(y=1)}{P(x_1=0)\, P(x_2=0)\, P(x_3=0)\, P(x_4=0)\, P(x_5=0)}$$

$$= \frac{0.25 \times 0.833 \times 0.25 \times 0.5 \times 0.75 \times 0.4}{0.4 \times 0.5 \times 0.3 \times 0.3 \times 0.7} = \frac{0.0078}{0.0126} = 0.619$$

$\boxed{y=-1}$ $P(y=-1 / x_1=0, x_2=0, x_3=0, x_4=0, x_5=0)$

$$= \frac{P(x_1=0|y=-1)\ p(x_2=0|y=-1)\ p(x_3=0|y=-1)\ p(x_4=0|y=-1)\ p(x_5=0|y=-1)\ p(y=-1)}{p(x_1=0)\ p(x_2=0)\ p(x_3=0)\ p(x_4=0)\ p(x_5=0)}$$

$$= \frac{0.5 \times 0.1667 \times 0.333 \times 0.1667 \times 0.667 \times 0.6}{0.4 \times 0.5 \times 0.3 \times 0.3 \times 0.7} = \frac{0.00185}{0.0126} = 0.1469$$

Since $p(y=1 / (0,0,0,0,0)) > P(y=-1 | (0,0,0,0,0))$

$\Rightarrow x = (0,0,0,0,0)$ is classified as $\underline{\underline{y=+1}}$

Part B

1) Gaussian Naive Bayes:

The Bayes rule, we go from P (X | Y) that can be found from the training dataset to find P (Y | X).
What we know from training data: P (X | Y) = P (X ∩ Y) / P(Y)

Bayes Rule = P (Y | X) = P (X | Y) * P (Y) / P (X)

Naive Bayes: Bayes rule provides the formula for the probability of Y given condition X, which in the real world, may not have multiple X variables. In this case, when you have independent features, the bayes rule is extended to the Naive Bayes rule where X's are independent of each other. One drawback is this only works for categorical variables.

For working on continuous variables, we use the Gaussian Naive Bayes. If we assume that X follows a particular distribution, you can use the probability density function of that distribution to calculate the probability of likelihoods.

Say X's follow a Gaussian or normal distribution, we substitute the probability density of the normal distribution, and this is called Gaussian Naïve Bayes.

$$P(X|Y=c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{\frac{-(x-\mu_c)^2}{2\sigma_c^2}}$$

The formula determines the likelihood of input values for each category by utilizing frequency calculations. The mean and standard deviation of x's for each class for the entire distribution can be calculated as

mean(x) = 1/n * sum(x)

where n is the number of instances and x is the value of the input variable.

And,

standard deviation(x) = sqrt(1/n * sum(xi-mean(x)^2))

where n is the number of instances and xi is a specific x value.

We can use this Gaussian probability density function to make predictions by substituting the parameters with the new input value of the variables. The Gaussian function as a result will be give an estimate for the new input value's probability.

We need to ensure the new input we test on should also follow Gaussian distribution.

2) Using GaussianNB() on the breast cancer dataset:

Part (a)

1. **Loading the Data and Understanding**
   a. After Loading the Data, we can see that the data has 30 different input features.
   b. The data has 569 rows of data
   c. It is a binary classification problem with two class 0 and 1 (Malignant and Benign)
2. **Splitting the Data**
   a. The data is split into a 75-25 train test split with a random seed of 42.
   b. Selected the split at random since no split was mentioned
   c. The seed is selected so that the data split is the same for any number of runs.
3. Plotting the Data distribution
   a. Since Gaussian NB works on normally/Gaussian distribution of data.
   b. We see from the plots that the almost all the features follow nearly gaussian distribution, hence we can use Gaussian NB without much feature engineering required.
4. **Model Definition**
   a. We define the model next and fit the model on the train set.
   b. Evaluations on the train set reveal the model accuracy as
      i. Prediction accuracy: 93.66%
      ii. Confusion Matrix:
```
[[139  19]
 [  8 260]]
```
5. Results on Test/Validation set
   a. Prediction accuracy: 95.80%
   b. Confusion Matrix:
```
[[51  3]
 [ 3 86]]
```
6. Trying **Cross-Validation**
   a. Running the model on 5 folds.
   b. Cross-Validation Experiment reveals that the first fold has the least accuracy score.
      i. Results:
```
[0.87719298 0.92105263 0.95614035 0.97368421 0.95575221]
```
      ii. This could maybe because of having some outliers in the first fold.

Discussion:

The classifier's performance on the original dataset gives us a good baseline for comparison. The classifier demonstrated high accuracy, showcasing its ability to effectively model the distribution of the dataset's features and make accurate predictions based on the learned patterns. The prediction accuracy in such controlled conditions usually reflects the classifier's potential when the assumption of feature independence and the presumed distributions align closely with the dataset's characteristics.

Part (b)

1. We follow the same procedure for part B as in part A.
2. The algorithm remains the same.
3. We add random gaussian noise to the data with 5 different variances [50, 100, 200, 400, 800].
4. The plots after adding random gaussian noise, show that the **data distribution follows a better Gaussian distribution** than in the original dataset and most of the variables follow similar distributions.
5. To compute the random gaussian noise, we need the std deviation which we get by sqrt(variance) and keep the dimensionality same as the input (30 features) and for all the data 569 data points.
6. The results are:

    a. Variance: 50
    Prediction Accuracy: 95.80%
    Confusion Matrix:
    ```
    [[49  5]
     [ 1 88]]
    ```

    b. Variance: 100
    Prediction Accuracy: 93.71%
    Confusion Matrix:
    ```
    [[47  7]
     [ 2 87]]
    ```

    c. Variance: 200
    Prediction Accuracy: 93.71%
    Confusion Matrix:
    ```
    [[47  7]
     [ 2 87]]
    ```

    d. Variance: 400
    Prediction Accuracy: 93.01%
    Confusion Matrix:
    ```
    [[46  8]
     [ 2 87]]
    ```

    e. Variance: 800
    Prediction Accuracy: 93.01%
    Confusion Matrix:
    ```
    [[46  8]
     [ 2 87]]
    ```

Discussion:

After adding random gaussian noise, we see the change in data distribution, where some features were skewed a bit before, adding random noise makes the data follow a better gaussian distribution.

These results show how the classifier's performance varies with the introduction of noise to the dataset. Interestingly, the classifier maintains relatively high accuracy even as noise variance increases, which suggests robustness of the Gaussian Naïve Bayes classifier to some extent of feature noise. However, as expected, there's a general trend of decreased performance (both in terms of accuracy and confusion matrix outcomes) as the variance of the noise increases.

CODE ATTACHED as mridul-khurana_naive-bayes.ipynb file.

3) Advantages and Disadvantages of Naive Bayes:

Advantages:

1. Naive Bayes classifiers are easy to implement and are fast to predict the class of the test data. It is a simple algorithm to understand.
2. It works well for multi-class problems.
3. Works well in higher dimensionality of input spaces as well. For ex. Text classification
4. Requires less training data and can fit with small amounts of data
5. Sometimes beats logistic regression or decision trees too when the assumption of independence holds
6. Performs good in categorical input variables.

Disadvantages:

1. It has a strong dependence on feature independence assumption. Fails in cases in which it doesn't hold true which is the case for most real-world applications.
2. If a categorical variable has a category in the test data, which was not observed in the training, then the model assigns zero probability. Also known as the zero probability / zero-frequency issue.
3. It is not the best estimator of probabilities. The probability outputs are not always reliable, better use it for classification accuracy.
4. It has difficulty in handling the continuous features as it always assumes the features to be either categorical or follow a particular distribution like gaussian if they are continuous.

Solving the "zero probability" (or "zero frequency") problem.

1. Laplace Smoothing: This technique is used for smoothing out the categorical data particularly to handle the zero-probability problem.
    a. This is done by adding a small positive value like 1 to the count for each class label at every feature. And the denominator is added by the small positive value * the number of features.
    b. This adjustment is done to avoid the probability never being zero.

2. Another way is ignoring the feature which leads to zero-probability, but this can possibly lead to loss of some valuable information.