DS 5220: Supervised Machine Learning and Learning Theory Summer, 2024
In Class Work 7

July 09, 2024

Learning Objectives:

- Modify the Naïve Bayes implementation example from today's class to handle both discrete and continuous features.
- Ensure that the classifier correctly calculates the posterior probabilities by combining the probabilities of discrete and continuous features

1 Identify and Separate Features:

I have identified and separated the features from the titanic-1.csv dataset.

- Discrete Features: These are categorical variables such as 'Pclass', 'Sex', 'SibSp', 'Parch', and 'Embarked'.
- Continuous Features: These are numeric variables such as 'Age' and 'Fare'

2 Calculate Prior Probabilities:

The Prior probabilities computed for each category in the 'Survived' column are as follows:

- Survived = 0 (Did not survive): 0.618 61.8%
- Survived = 1 (Survived): 0.382 38.2%

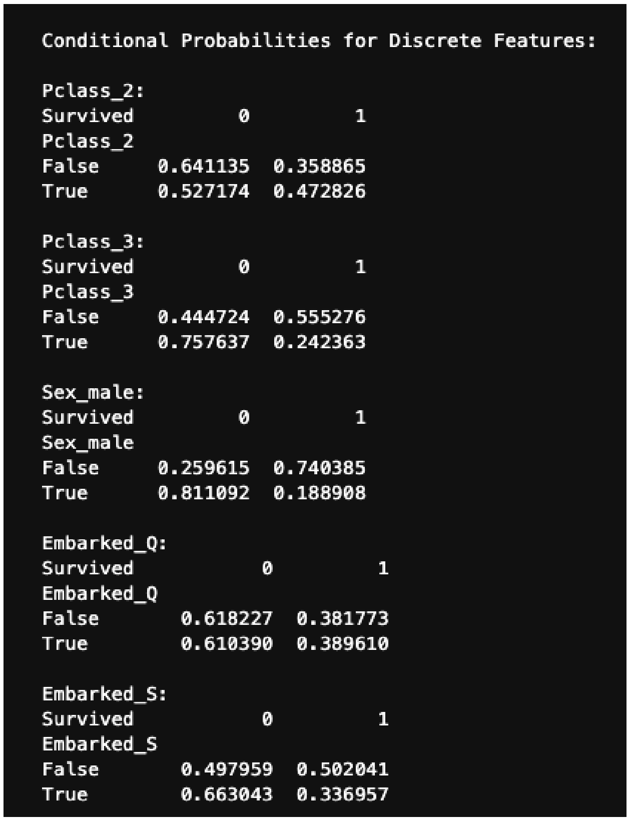
These proportions suggest that, there are more instances of individuals who did not survive compared to those who did. These prior probabilities are crucial in Bayesian inference and classification tasks, providing baseline information about the distribution of classes before any further analysis or modeling is performed.

3 Calculate Conditional Probabilities:

The tables display the correlation between various categorical features and the likelihood of survival ('Survived=1'):

- Pclass 2 and Pclass 3: Lower classes (Pclass 2=False, Pclass 3=True) exhibit higher survival rates.
- Sex male: Females (Sex male=False) have a greater chance of survival compared to males (Sex male=True).
- Embarked Q and Embarked S: Survival probabilities are influenced by the port of embarkation, with varying probabilities observed.

  These findings provide valuable insights into how categorical factors impact the chances of survival on the Titanic.



```
Conditional Probabilities for Discrete Features:

Pclass_2:
Survived         0         1
Pclass_2
False     0.641135  0.358865
True      0.527174  0.472826

Pclass_3:
Survived         0         1
Pclass_3
False     0.444724  0.555276
True      0.757637  0.242363

Sex_male:
Survived         0         1
Sex_male
False     0.259615  0.740385
True      0.811092  0.188908

Embarked_Q:
Survived           0         1
Embarked_Q
False       0.618227  0.381773
True        0.610390  0.389610

Embarked_S:
Survived           0         1
Embarked_S
False       0.497959  0.502041
True        0.663043  0.336957
```

The statistics reveal that the average age of Titanic passengers is roughly 30 years old, with a variance of 156.25 showing the spread of ages around this average. In terms of fare, the average amount paid is around 22.12 units, with a higher variance of 985.22 suggesting a wider range of fares paid by passengers.

Mean and Variance for Continuous Features:
Age - Mean: 30.03, Variance: 156.25
Fare - Mean: 22.12, Variance: 985.22

4 Implement the Classifier:

The dataset (‘X‘) is splited into training and testing sets, where the testing set comprises 20% of the total dataset and ‘random state=42‘ is set for reproducibility purposes. Subsequently, a Gaussian Naive Bayes (‘GaussianNB‘) model is implemented. The model is then trained on the training data to estimate the necessary parameters for classification based on the given training data. To summarize, this process involves training a Gaussian Naive

Bayes classifier on the training data and evaluating its accuracy on unseen test data.

5 Evaluate the Model:

- Accuracy (0.79): The model accurately predicts 79% of the outcomes in the test set.
- Precision (0.70) : 70% of the predicted positives were actually positive.
- Recall (0.78): The model recognized 78% of all actual positives in the test set.
- F1-score (0.74): A balanced measure combining precision and recall, reflecting overall model performance.