Mohamad Nael Ayoubi

150120997

Introduction to Machine Learning

CSE4288

Assignment #1

1. Introduction

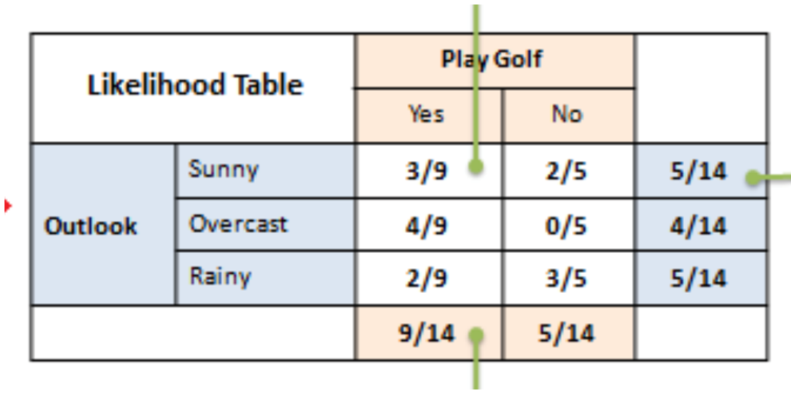
The Naive Bayes algorithm is a probabilistic machine learning classifier based on Bayes' theorem, assuming that the features used for classification are independent.

The goal of this assignment is to implement a Naive Bayes classifier to predict whether a person plays tennis based on four features: Outlook, Temperature, Humidity, and Wind. The dataset used consists of these features with binary classification labels ("Yes" or "No") for the target variable, PlayTennis.

2. Methodology

- The dataset used in this assignment is stored in a CSV file named dataset.csv. It contains several instances of daily weather conditions along with whether a person plays tennis. The data was read using the pandas library, and its structure was printed for verification. The relevant columns in the dataset are: Outlook, Temperature, Humidity, Wind, PlayTennis (target variable).
- To implement the Naive Bayes classifier, I followed these steps:
  - Prior Probabilities: The first step was to calculate the prior probabilities of the two possible classes, "Yes" and "No", based on the frequency of each class in the dataset.
  - Likelihood Probabilities (with Laplace Smoothing): For each feature, I computed the conditional probabilities (likelihoods) of each possible value given each class label. Since some feature values may not appear for certain classes, I used Laplace smoothing (adding 1 to the count of each feature value) to avoid zero probabilities. These conditional probabilities were calculated

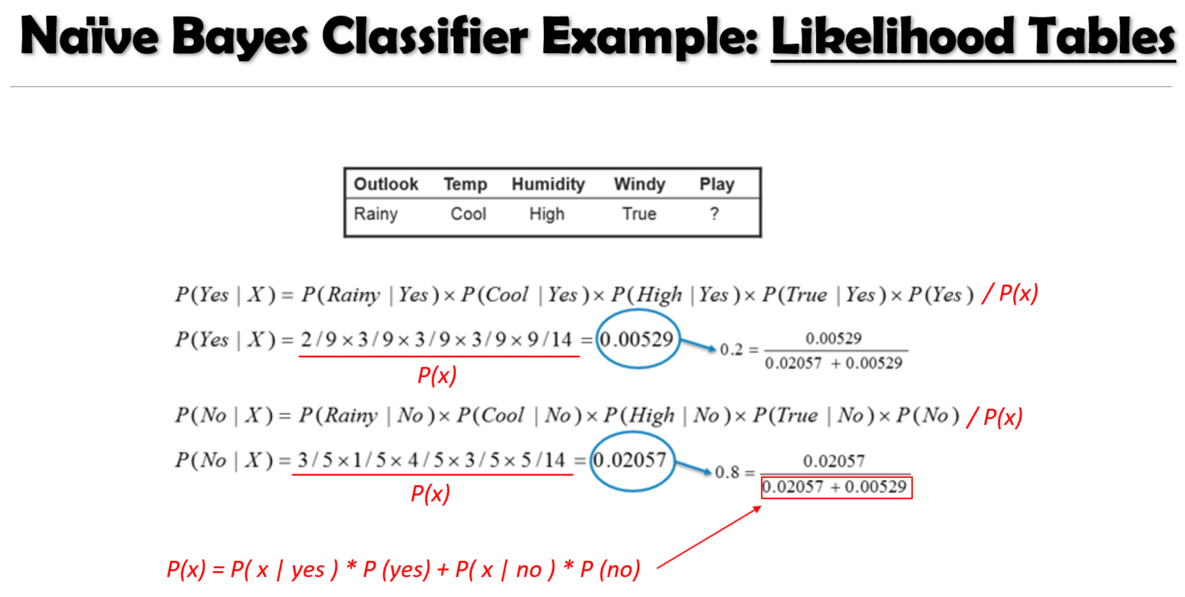
separately for each class ("Yes" and "No") and stored for later use.



Likelihood Table		Play Golf		
		Yes	No	
Outlook	Sunny	3/9	2/5	5/14
	Overcast	4/9	0/5	4/14
	Rainy	2/9	3/5	5/14
		9/14	5/14	

Figure 1: Example likelihood table [1]

- Prediction: For each test instance, the log of the prior probability for each class ("Yes" and "No") was added to the logarithm of the corresponding likelihood probabilities for each feature value in the instance. I used log because Probabilities in Naive Bayes calculations can be very small, especially when you multiply many probabilities together. The class with the higher posterior probability was predicted as the output.



Naïve Bayes Classifier Example: Likelihood Tables

Outlook	Temp	Humidity	Windy	Play
Rainy	Cool	High	True	?

$$P(Yes \mid X) = P(Rainy \mid Yes) \times P(Cool \mid Yes) \times P(High \mid Yes) \times P(True \mid Yes) \times P(Yes) \ / \ P(x)$$

$$P(Yes \mid X) = \underbrace{2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14}_{P(x)} = \boxed{0.00529} \rightarrow 0.2 = \frac{0.00529}{0.02057 + 0.00529}$$

$$P(No \mid X) = P(Rainy \mid No) \times P(Cool \mid No) \times P(High \mid No) \times P(True \mid No) \times P(No) \ / \ P(x)$$

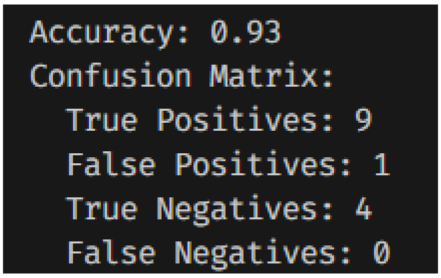
$$P(No \mid X) = \underbrace{3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14}_{P(x)} = \boxed{0.02057} \rightarrow 0.8 = \frac{0.02057}{\boxed{0.02057 + 0.00529}}$$

P(x) = P(x | yes) * P (yes) + P(x | no) * P (no)

Figure 2: Above is shown how result is produced [1].

3. Results

The classifier was tested on the entire dataset. Below are the key performance metrics: Accuracy: The classifier achieved an accuracy of approximately 0.93 (calculated after testing on the entire dataset). Confusion Matrix is shown below:



```
Accuracy: 0.93
Confusion Matrix:
  True Positives: 9
  False Positives: 1
  True Negatives: 4
  False Negatives: 0
```

4. Discussion

The Naive Bayes classifier performed well with a reasonable accuracy, indicating that it was able to predict whether a person plays tennis based on weather conditions.

There are several possible reasons why the classifier may make incorrect predictions:

- Insufficient Data: If the dataset is not large enough or does not cover a wide range of possible weather conditions, the classifier may not generalize well.
- Feature Independence Assumption: Naive Bayes assumes that all features are independent, which may not hold true in reality. This could lead to suboptimal predictions in some cases.
- Imbalanced Classes: If the dataset contains significantly more instances of one class (e.g., more "Yes" than "No"), the classifier may be biased toward predicting that class, leading to misclassifications.

5. Conclusion

The implementation demonstrates the effectiveness of Naive Bayes, particularly for small datasets and situations where feature independence assumptions hold approximately true.

6. References

[1] Slides of Chap6-ClassificationBasic which was given in lectures.