Naive Bayes Classifier for Balance Scale Dataset

Group Number: 26
Roll numbers: 21CS30041, 23AT61R04

Overview

This project aims to implement a Naive Bayes classifier for the balance scale dataset, exploring the algorithm's theoretical foundations and providing a transparent, from-scratch implementation in Python. The primary objective is to develop a model capable of classifying instances into one of three categories: Left (L), Balanced (B), and Right (R). The dataset contains 625 instances, each characterized by four numeric attributes and a class label, resulting in a five-dimensional dataset.

Theory

Naive Bayes Classifier

The Naive Bayes classifier leverages Bayes' theorem, a probabilistic approach to classification. The algorithm assumes that features are conditionally independent given the class label, allowing for a simplified and computationally efficient model. By calculating class and conditional probabilities, the Naive Bayes classifier makes predictions based on the maximum posterior probability.

Bayes' theorem finds the probability of an event occurring given the probability of another event that has already occurred.

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

where A and B are events and P(B) is non-zero.

- P(A) is the prior probability i.e. probability before any evidence is known.
- P(B) is the marginal probability i.e. that of the evidence
- P(B|A) is the likelihood probability i.e. likelihood that the hypothesis will come true given evidence
- P(A|B) is the posterior probability i.e. probability of the event after the evidence is seen

Implementation Details

Training and Evaluation

The implementation focuses on a theoretical understanding of the Naive Bayes algorithm while incorporating practical considerations for model training and evaluation. Key steps include:

1. **Naive Bayes Classifier Implementation:**
   - Training involves calculating prior and conditional probabilities, and, eventually the posterior probability based on whose values across all classes the class with the highest posterior probability is chosen to be the model prediction
2. **5-Fold Cross-Validation:**
   - The model is trained and evaluated using 5-fold cross-validation.
   - This ensures robust evaluation, preventing over-fitting and providing a reliable estimate of the classifier's performance.

Dataset Exploration

Balance Scale Dataset Overview

- **Number of Instances:** 625
- **Class Distribution:**
  - Balanced: 49 instances
  - Left: 288 instances
  - Right: 288 instances

Evaluation Metrics

Apart from model accuracy for each fold, model's performance is also evaluated using the following standard metrics albeit directly with the `classification_report` function from `sklearn.metrics:

- *Precision*: measures the accuracy of positive predictions, indicating the model's ability to correctly identify instances of a specific class.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- *Recall*: gauges the model's ability to capture all instances of a specific class, providing insights into its sensitivity.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- *F1* score: represents the harmonic mean of precision and recall, providing a balanced measure of the classifier's overall performance.

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Results and Comparison

The evaluation metrics provide a comprehensive view of the Naive Bayes classifier's effectiveness in classifying balance scale instances.

The algorithm implemented by us has managed to arrive at an average accuracy of `0.9070064516129033` and the decision tree classifier used by the sample code provided has an average accuracy of just `0.7948258064516129`

For further comparison of the two models, below is the output by `classification_report` with the highest accuracy obtained in the 5-fold cross-validation method for both the model trained by us

	precision	recall	f1-score	support
B	1.00	0.00	0.00	9
L	0.92	1.00	0.96	60
R	0.93	1.00	0.97	56
accuracy			0.93	125
macro avg	0.95	0.67	0.64	125
weighted avg	0.93	0.93	0.89	125

and for the code snippet provided to us

	precision	recall	f1-score	support
B	0.00	0.00	0.00	7
L	0.95	0.95	0.95	57
R	0.89	0.90	0.89	61
accuracy			0.87	125
macro avg	0.61	0.62	0.61	125
weighted avg	0.86	0.87	0.87	125