DS552 – Generative AI

Assignment 2 – Generative Vs Discriminative Models

smadhavankutty@wpi.edu
2/2/2025

Penguins DataSet

Performance of Naive Bayes (Generative Model) and Logistic Regression (Discriminative Model)

GitHub Repo Link : https://github.com/smadhavanwpi/DS552_GenAI

Python Source File : genai_assignment1_penguin.py

Python Notebook : GenAI_Assignment1_Penguin.ipynb

Explanation of the Code:

1. Data Preparation:
    - The Penguins dataset is loaded and filtered to include only two species: Adelie and Gentoo.
    - Missing values are dropped, and species labels are encoded numerically.
2. Model Training:
    - Naive Bayes (GaussianNB) and Logistic Regression models are trained on the training dataset.
3. Accuracy Evaluation:
    - The accuracy of both models is calculated on the training and test datasets.
4. AUC Calculation:
    - The AUC (Area Under the Curve) is computed for both models to evaluate their ability to discriminate between the two species.
5. Lift and Gain Charts:
    - Lift and Gain charts are generated for both models to visualize their performance in ranking predicted probabilities.
6. Model Comparison:
    - The accuracy and AUC metrics are compared to determine which model performs better.

Insights:

- Accuracy: Logistic Regression often outperforms Naive Bayes in terms of accuracy, especially when the data is not perfectly Gaussian.
- AUC: A higher AUC indicates better discrimination between the two species. Logistic Regression typically has a higher AUC due to its discriminative nature.
- Lift and Gain Charts: These charts help visualize how well the models prioritize the classification of the two species. Logistic Regression usually shows better lift and gain due to its ability to model complex decision boundaries.

Conclusion:

Based on the results, Logistic Regression is likely to perform better than Naive Bayes in classifying the two penguin species, as it generally achieves higher accuracy and AUC values. However, the final choice of model should also consider other factors like interpretability, computational efficiency, and domain-specific requirements.

MNIST (handwritten digits)

Performance of Naive Bayes (Generative Model) and Logistic Regression (Discriminative Model)

GitHub Repo Link : https://github.com/smadhavanwpi/DS552_GenAI

Python Source File : genai_assignment1_mnist.py

Python Notebook : GenAI_Assignment1_MNIST.ipynb

Explanation of the Code:

1. Data Loading:
    - The MNIST dataset is loaded using fetch_openml. It contains 70,000 images of handwritten digits (0-9), each represented as a 28x28 pixel array (flattened into 784 features).
2. Data Preprocessing:
    - Pixel values are normalized to the range [0, 1] to improve model performance.
3. Model Training:
    - A Gaussian Naive Bayes model and a Logistic Regression model are trained on the MNIST dataset.
4. Evaluation:
    - The accuracy of both models is calculated on the training and test datasets.
    - A classification report is generated to provide detailed metrics (precision, recall, F1-score) for each class.
5. Comparison:
    - The performance of Naive Bayes and Logistic Regression is compared based on test accuracy.

Penguins Vs MNIST Datasets – Differences in performance and behavior across datasets

Performance on MNIST vs. Penguins Dataset:

- MNIST Dataset:
    - The MNIST dataset is high-dimensional (784 features) and represents complex patterns in image data.
    - Logistic Regression typically outperforms Naive Bayes on MNIST because:
        - Logistic Regression is a discriminative model that directly learns the decision boundary between classes.
        - Naive Bayes assumes independence between features, which is a strong and often incorrect assumption for image data (pixels are correlated).
    - Naive Bayes struggles with the high dimensionality and complex relationships in image data, leading to lower accuracy.
- Penguins Dataset:
    - The Penguins dataset is low-dimensional (4 features) and contains simpler, tabular data.
    - Both Naive Bayes and Logistic Regression perform well on the Penguins dataset, but Logistic Regression often has a slight edge due to its ability to model more complex relationships.
    - Naive Bayes performs better on the Penguins dataset compared to MNIST because the independence assumption is less problematic for tabular data.

Behavior of Models:

- Naive Bayes:
    - Works well for small datasets with simple relationships (e.g., Penguins dataset).
    - Struggles with high-dimensional data (e.g., MNIST) due to the independence assumption and inability to capture complex patterns.
    - Computationally efficient and requires less training data.
- Logistic Regression:
    - Excels in high-dimensional spaces and can capture complex relationships (e.g., MNIST).
    - Requires more data and computational resources compared to Naive Bayes.
    - Performs well on both simple (Penguins) and complex (MNIST) datasets but is particularly effective for image data.

Key Insights and Takeaways:

- For simple, low-dimensional datasets like the Penguins dataset, both models perform well, but Logistic Regression may have a slight advantage.
- For complex, high-dimensional datasets like MNIST, Logistic Regression is significantly better than Naive Bayes due to its ability to model complex relationships.
- Naive Bayes is more suitable for small datasets or when computational efficiency is a priority, but it may not perform well on complex data like images.

This output demonstrates that Logistic Regression significantly outperforms Naive Bayes on the MNIST dataset, highlighting the differences in their suitability for complex image data.