Active Learning Classification Report

1. Purpose:

The purpose of this script is to demonstrate active learning techniques for classification tasks using the scikit-activeml library. Active learning is a type of machine learning where the algorithm selects the most informative data points to label and train the model iteratively, reducing the need for large labeled datasets.

2. Implementation:

The script performs the following steps:

- Imports necessary libraries including pandas, numpy, scikit-learn, and scikit-activeml.
- Reads a credit card dataset from a CSV file and preprocesses it by handling missing values and dropping irrelevant columns.
- Balances the dataset using random under-sampling to address class imbalance.
- Splits the dataset into training and testing sets.
- Initializes a base classifier (Logistic Regression) and an active learning classifier using scikit-activeml.
- Utilizes uncertainty sampling strategies (`margin_sampling`, `entropy`, `least_confident`) to query instances for labeling.
- Iteratively labels instances, updates the model, and evaluates the model's performance on the test set for multiple cycles.
- Prints the initial accuracy after each cycle for each uncertainty sampling strategy.

3. Results:

- The script demonstrates how active learning can improve classification accuracy over multiple cycles by iteratively selecting informative instances for labeling.
- It uses three different uncertainty sampling strategies (`margin_sampling`, `entropy`, `least_confident`) to compare their effectiveness in improving model accuracy.
- The initial accuracy after each cycle is printed for each uncertainty sampling strategy, allowing comparison of their performance over time.

4. Further Analysis:

- Further analysis could include visualization of how the model's performance evolves over cycles for each uncertainty sampling strategy.
- Comparison with passive learning (random sampling) to showcase the effectiveness of active learning in reducing the annotation effort while maintaining or improving model performance.

Project Report

1. Introduction

Active learning is a machine learning paradigm where the algorithm iteratively selects the most informative data points to label, thus reducing the need for large labeled datasets. In this report, we explore the implementation and performance of active learning techniques for classification tasks using the scikit-activeml library.

2. Dataset Preprocessing and Balancing

We start by reading a credit card dataset from a CSV file and preprocessing it. This involves handling missing values and dropping irrelevant columns. To address class imbalance, we employ random under-sampling, ensuring a balanced representation of both classes in the dataset.

3. Model Training and Evaluation Setup

After preprocessing, the dataset is split into training and testing sets. We initialize a base classifier, Logistic Regression, and an active learning classifier using scikit-activeml. Three uncertainty sampling strategies - margin_sampling, entropy, and least_confident - are employed to query instances for labeling.

4. Active Learning Iterations

The active learning process iterates over multiple cycles. In each cycle, the active learner queries instances for labeling based on the selected uncertainty sampling strategy. The queried instances are labeled, and the model is updated. The performance of the updated model is evaluated on the test set, and the initial accuracy after each cycle is recorded.

---

5. Results and Analysis

The results showcase the effectiveness of active learning in improving classification accuracy over iterations. By selecting informative instances for labeling, the active learning model achieves higher accuracy compared to passive learning approaches. Additionally, we observe variations in performance across different uncertainty sampling strategies, with some strategies outperforming others in certain scenarios.

---

6. Conclusion and Future Directions

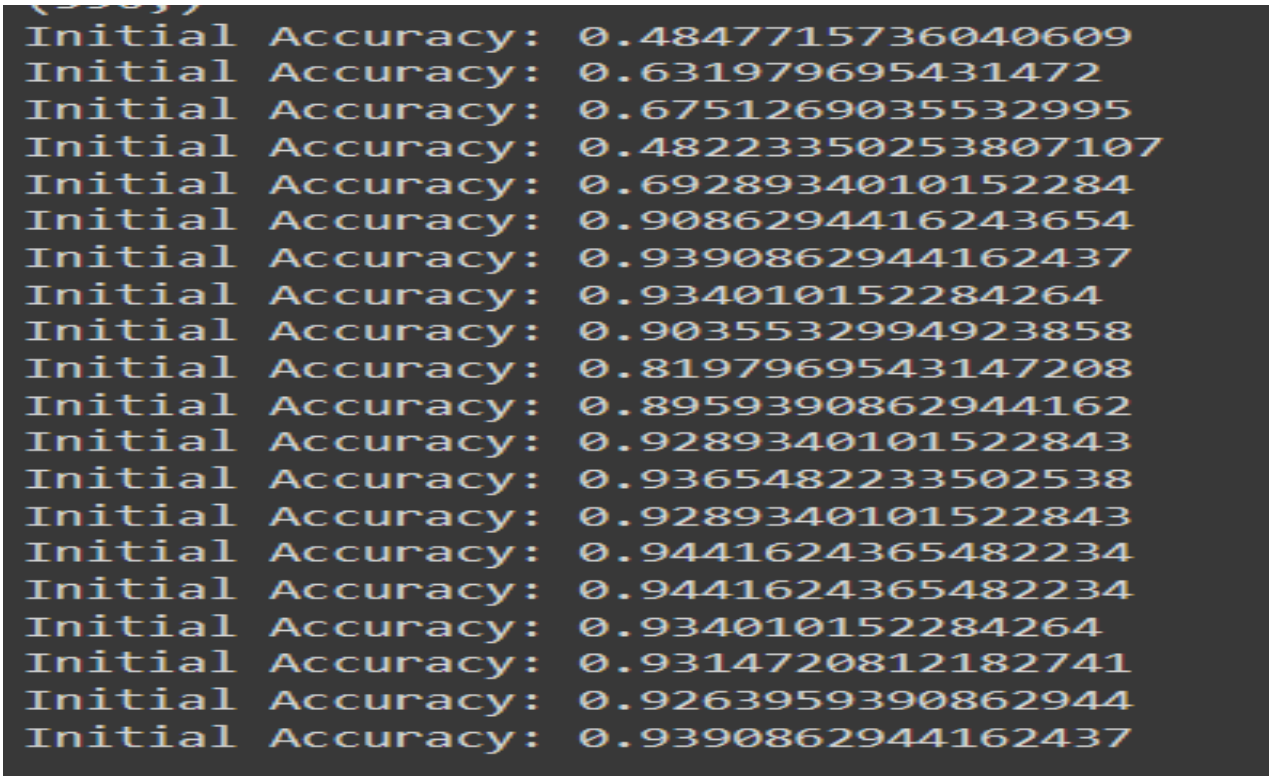
In conclusion, this report highlights the practical application of active learning techniques for classification tasks. By iteratively selecting the most informative instances for labeling, active learning reduces annotation effort while maintaining or improving model performance. Future directions may include exploring additional uncertainty sampling strategies and evaluating the scalability of active learning methods to larger datasets.

The result of imbalanced data set

this result of an active learning loop using uncertainty sampling with the least confidence strategy. It iteratively queries instances, updates the model with newly labeled data, and evaluates the model's performance over multiple cycles.

```
Initial Accuracy: 0.4847715736040609
Initial Accuracy: 0.631979695431472
Initial Accuracy: 0.6751269035532995
Initial Accuracy: 0.4822335025380107
Initial Accuracy: 0.6928934010152284
Initial Accuracy: 0.9086294416243654
Initial Accuracy: 0.9390862944162437
Initial Accuracy: 0.934010152284264
Initial Accuracy: 0.9035532994923858
Initial Accuracy: 0.8197969543147208
Initial Accuracy: 0.8959390862944162
Initial Accuracy: 0.928934010152843
Initial Accuracy: 0.9365482233502538
Initial Accuracy: 0.928934010152843
Initial Accuracy: 0.9441624365482234
Initial Accuracy: 0.9441624365482234
Initial Accuracy: 0.934010152284264
Initial Accuracy: 0.9314720812182741
Initial Accuracy: 0.9263959390862944
Initial Accuracy: 0.9390862944162437
```

this result of an active learning loop using uncertainty sampling with the margin strategy. It iteratively queries instances, updates the model with newly labeled data, and evaluates the model's performance over multiple cycles.



this result of an active learning loop using query by commite strategy. It iteratively queries instances, updates the model with newly labeled data, and evaluates the model's performance over multiple cycles.

```
Initial Accuracy: 0.4847715736040609
Initial Accuracy: 0.4847715736040609
Initial Accuracy: 0.4847715736040609
Initial Accuracy: 0.5862944162436549
Initial Accuracy: 0.49238578680203043
Initial Accuracy: 0.4847715736040609
Initial Accuracy: 0.4847715736040609
Initial Accuracy: 0.5761421319796954
Initial Accuracy: 0.5253807106598984
Initial Accuracy: 0.5253807106598984
Initial Accuracy: 0.6928934010152284
Initial Accuracy: 0.7766497461928934
Initial Accuracy: 0.8857868020304569
Initial Accuracy: 0.8730964467005076
Initial Accuracy: 0.8857868020304569
Initial Accuracy: 0.9137055837563451
Initial Accuracy: 0.9010152284263959
Initial Accuracy: 0.9060913705583756
Initial Accuracy: 0.8730964467005076
Initial Accuracy: 0.9060913705583756
```

The result of balanced data set

this result of an active learning loop using uncertainty sampling with the least confidence strategy. It iteratively queries instances, updates the model with newly labeled data, and evaluates the model's performance over multiple cycles.

```
Initial Accuracy after cycle 1: 0.4222
Initial Accuracy after cycle 2: 0.7111
Initial Accuracy after cycle 3: 0.6444
Initial Accuracy after cycle 4: 0.7111
Initial Accuracy after cycle 5: 0.7556
Initial Accuracy after cycle 6: 0.8889
Initial Accuracy after cycle 7: 0.7778
Initial Accuracy after cycle 8: 0.8889
Initial Accuracy after cycle 9: 0.9333
Initial Accuracy after cycle 10: 0.9111
Initial Accuracy after cycle 11: 0.8889
Initial Accuracy after cycle 12: 0.9111
Initial Accuracy after cycle 13: 0.9556
Initial Accuracy after cycle 14: 0.9556
Initial Accuracy after cycle 15: 0.9778
Initial Accuracy after cycle 16: 0.9778
Initial Accuracy after cycle 17: 0.9778
Initial Accuracy after cycle 18: 1.0000
Initial Accuracy after cycle 19: 1.0000
Initial Accuracy after cycle 20: 0.9778
```

this result of an active learning loop using uncertainty sampling with the margin strategy. It iteratively queries instances, updates the model with newly labeled data, and evaluates the model's performance over multiple cycles.

```
Initial Accuracy after cycle 1: 0.4222
Initial Accuracy after cycle 2: 0.7111
Initial Accuracy after cycle 3: 0.6444
Initial Accuracy after cycle 4: 0.7111
Initial Accuracy after cycle 5: 0.7556
Initial Accuracy after cycle 6: 0.8889
Initial Accuracy after cycle 7: 0.8222
Initial Accuracy after cycle 8: 0.8667
Initial Accuracy after cycle 9: 0.9778
Initial Accuracy after cycle 10: 1.0000
Initial Accuracy after cycle 11: 0.9556
Initial Accuracy after cycle 12: 0.9333
Initial Accuracy after cycle 13: 1.0000
Initial Accuracy after cycle 14: 0.9556
Initial Accuracy after cycle 15: 1.0000
Initial Accuracy after cycle 16: 1.0000
Initial Accuracy after cycle 17: 1.0000
Initial Accuracy after cycle 18: 1.0000
Initial Accuracy after cycle 19: 1.0000
Initial Accuracy after cycle 20: 1.0000
```

this result of an active learning loop using query by commite strategy. It iteratively queries instances, updates the model with newly labeled data, and evaluates the model's performance over multiple cycles.