Convolve Epoch2 Round2

Introduction

In this report, we will delve into the process of selecting and calibrating models for binary classification tasks. Our primary goal is to predict probabilities accurately, as it can be more informative and useful than raw binary predictions. To achieve this, we followed a systematic approach that included feature selection, model selection, and probability calibration.

Methodology

1. Feature Selection

We began our analysis by inspecting the dataset and visualizing it using graphs. The initial step was to identify and eliminate bad features that could potentially hinder model performance. This process involved careful examination of feature distributions, correlation matrices, and domain knowledge to remove irrelevant or redundant features.

As the dataset had some columns which didn't have a singular datatype but where a mix of string and integer, so we decided to rely on the graphs for our feature selection.

2. Model Selection

We opted for a diverse set of five classification models to assess their performance on the binary classification task. The models we considered were:

1. Stochastic Gradient Descent (SGD)
2. Linear Support Vector Classifier (Linear SVC)
3. Support Vector Classifier (SVC)
4. Random Forest Classifier
5. k-Nearest Neighbors (KNN)

Each model was implemented and trained on the dataset to generate initial predictions.

3. Probability Estimation

For the KNN and Random Forest models, we utilized built-in functions to directly calculate the probabilities associated with each prediction. These methods provided probability estimates without additional calibration steps.

However, for the other three models (SGD, Linear SVC, and SVC), we employed `Platt Scaling` to train a `Logistic Regression` model. Platt scaling helps convert raw model scores into probability estimates. The logistic regression model was trained on the model's output scores and true labels from the training dataset.

4. Probability Averaging

To generate the final probability estimates, we implemented a `weighted averaging` approach. The weights assigned to each model were determined based on the `accuracy` of the respective models on the training dataset. Models with higher accuracy were given higher weights in the averaging process, reflecting their better performance.

Results

The final probability estimates were obtained by taking the weighted average of the probability values generated by each model. This approach ensured that the most accurate models contributed more to the final predictions.

Conclusion

In this report, we outlined our approach to feature selection, model selection, and probability calibration for binary classification tasks. Our method aimed to produce accurate probability estimates, which can be valuable for decision-making and understanding model uncertainty.

By following this systematic approach, we were able to harness the strengths of different models while mitigating their weaknesses. The use of Platt scaling for probability calibration and weighted averaging for combining model predictions yielded reliable probability estimates.

For in-depth functioning , refer to the comments and notes in the python notebook.
