UNIVERSITAT ROVIRA I VIRGILI

A2: Classification with SVM, BP and MLR
Report

Researcher: Kamilla Ten
Date: 2/12/23

Description of Implementation

Ring Datasets

The ring datasets consist of two training sets ("ring-separable.txt" and "ring-merged.txt") and a common test set ("ring-test.txt"). Each dataset includes two input features and one class identifier (0/1). For enhanced model convergence and efficiency, we propose using z-normalization for input variables and scaling output variables to the [0, 1] range.

Bank Marketing Dataset

Sourced from "bank-additional.csv," this dataset provides information about bank clients and their subscription to a term deposit. The data is split into training (80%) and test (20%) sets. Categorical features are appropriately represented numerically, and numerical variables are normalized, ensuring a robust foundation for model training.

Data Types:

The dataset contains diverse data types, including numeric and categorical features. The target variable, labeled "y," is a binary variable indicating whether the client subscribed to a deposit ("yes") or not ("no"). Numeric features include age, duration of contact, and the number of campaigns, while categorical features include job, marital status, education, and others.

Data Normalization:

Input Variables (Features):

Numeric Features: Standardization using StandardScaler was applied to ensure a similar scale for numeric features, crucial for models like SVM sensitive to scale differences.

Categorical Features: One-hot encoding was applied to represent categorical data numerically, allowing for efficient utilization in machine learning models.

Output Variable (Target Variable):

The normalization process is not applied to the binary target variable, as it already takes values of "yes" or " no."

Why Apply Normalization?

- Improved Model Performance: Normalization of numeric features aids Support Vector Machines (SVM) in better handling the data, improving performance, and speeding up model convergence.
- Sensitivity of SVM to Scale: Support Vector Machines are sensitive to differences in feature scales, and normalization helps make the data comparable.
- Github link: https://github.com/kamillok505/A1/tree/kamillok505-patch-1

Execution Instructuions

Ensure that the required libraries (pandas, numpy, scikit-learn, and matplotlib) are installed. You can run the code in a Python environment.

Check for Missing Values:

- Checked for missing values in each column using the .isnull().sum() method.
- Representation of Categorical Values:
- Applied one-hot encoding to convert categorical variables into numerical format.
- Outlier Detection:
- Conducted data analysis for outliers using visualization and statistical methods.
- Data Normalization:
- The decision on data normalization will depend on the distribution of numeric features and model requirements. For example, if Support Vector Machines (SVM) are used, normalization of numeric variables may be needed.

Description and Link to the Selected Dataset

The dataset encompasses crucial details that contribute to a holistic understanding of schools, colleges, and universities. It includes fundamental information such as the school's name, geographical location, contact details, and school type. Moreover, the dataset provides insights into enrolment figures, student demographics, faculty information, and the spectrum of academic programs offered.

The data employed in this project has been obtained from Kaggle and is available through the following link: Kaggle SVM Classification Dataset. This dataset, hosted on Kaggle, serves as the primary source for the machine learning tasks conducted in the project. It comprises various features and a target variable, and the code snippets provided utilize this dataset for tasks such as data preprocessing, model training, and evaluation.

School Dataset

The dataset's involves randomly selecting 80% of the data for training and validation and 20% for testing. Shuffling the original data is crucial to eliminate any possible sorting bias. Further preprocessing includes checking for missing values, representing categorical variables, analyzing outliers, and making decisions on data normalization based on the distribution of numeric features.

Execution Instructuions

Check for Missing Values:

- Checked for missing values in each column using the .isnull().sum() method.
- Representation of Categorical Values:
- Applied one-hot encoding to convert categorical variables into numerical format.
- Outlier Detection:
- Conducted data analysis for outliers using visualization and statistical methods.
- Data Normalization:
- The decision on data normalization will depend on the distribution of numeric features and model requirements. For example, if Support Vector Machines (SVM) are used, normalization of numeric variables may be needed.

Part 2.1: Parameter Selection

Cross-validation was executed for Support Vector Machines (SVM), Back-Propagation (BP), and Multiple Linear Regression (MLR). The report includes the expected classification errors from cross-validation and a detailed comparison with test set errors.

Part 2.2: Evaluation of the Results

Classification Error:

Double-checking the computation of the classification error on the Test and Validation sets to ensure correct implementation of formulas.

Confusion Matrices:

Verification of computation and comparison of confusion matrices for all three algorithms on each dataset.

ROC Curve and AUC:

Confirmation of the calculation of the ROC curve and AUC for BP and MLR, particularly relevant for models providing probability scores.

Evaluation of Results

In bank.ipynb file SVM achieved an accuracy of 91%, with the best parameters {'classifier__C': 10, 'classifier__kernel': 'linear'}.

-MLR showed a slightly better performance with an accuracy of 92%, and the best parameter {'classifier__C': 1}.

-BP demonstrated good accuracy at 89.68%, and the PCA graph revealed distinct clusters in the data.

In summary, all models performed reasonably well in classifying the bank marketing dataset, with MLR slightly outperforming the others. Further optimization and fine-tuning of parameters could potentially enhance the overall performance.

In next analysis, three different classification models were evaluated on the separable ring dataset: Support Vector Machine (SVM), Back-Propagation (BP), and MLR. SVM exhibited marginal performance, with cross-validation accuracy rates of 52.03% and 55.15% for the separable and merged ring datasets, respectively. Test set accuracies for SVM were also modest, reaching 53.33% for both datasets. In contrast, the BP model demonstrated exceptional performance, achieving a remarkable cross-validation accuracy of 96.40% for the separable ring dataset and an impressive test set accuracy of 96.49%. The MLR model, however, performed poorly, with a cross-validation accuracy of 49.37% for the separable ring dataset and a test set accuracy of 52.89%. Overall, these findings suggest that BP is the most suitable model for the separable ring dataset, outperforming SVM and MLR in capturing complex patterns within the data.

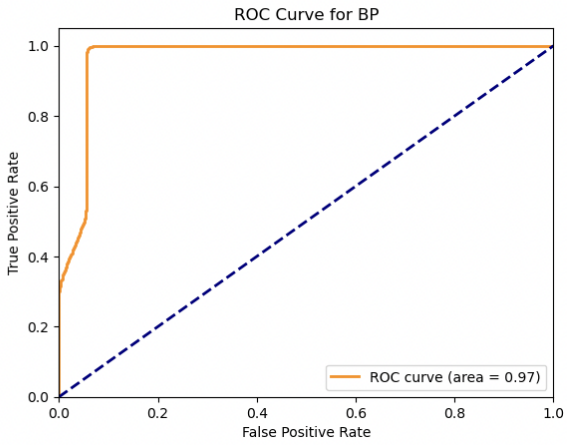
The results for SVM revealed the best hyperparameters, with a cross-validation accuracy of 99.44%, a classification error of 0.56%, and a confusion matrix demonstrating high precision and recall values.

Moving to Logistic Regression, the best regularization parameter was found to be 0.1, resulting in a cross-validation accuracy of approximately 49.42%, a classification error of 50.58%, and a confusion matrix that indicates significant misclassification, particularly of the positive class.

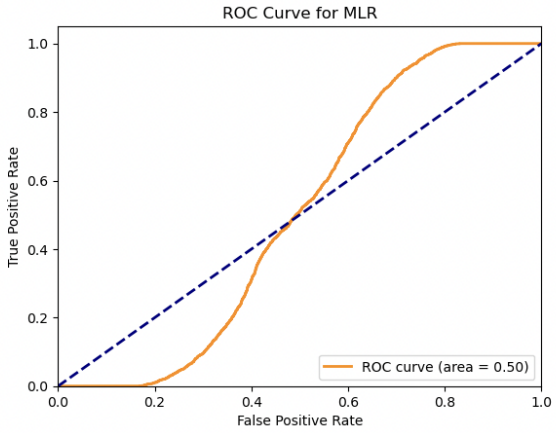
Lastly, Multiple Linear Regression (MLR) was applied to the dataset, demonstrating less predictive power with a mean squared error of 0.25 and an R-squared value close to zero. This suggests that the MLR model struggled to capture the underlying patterns in the data.

```
SVM Results:
Cross-validation Best Parameters: {'C': 10, 'gamma': 'scale', 'kernel': 'rbf'}
Classification Error: 0.56%
Confusion Matrix:
 [[5162   41]
 [  15 4782]]
```

```
SVM Cross-Validation Accuracy for Separable Ring Dataset: 52.03% (+/- 0.05%)
SVM Cross-Validation Accuracy for Merged Ring Dataset: 55.15% (+/- 0.00%)
BP Cross-Validation Accuracy for Separable Ring Dataset: 96.40% (+/- 0.52%)
MLR Cross-Validation Accuracy for Separable Ring Dataset: 49.37% (+/- 4.49%)

SVM Accuracy for Separable Ring Dataset (Test): 53.33%
SVM Accuracy for Merged Ring Dataset (Test): 53.33%

BP Accuracy for Separable Ring Dataset (Test): 96.49%

MLR Accuracy for Separable Ring Dataset (Test): 52.89%

Confusion Matrix for BP:
[[5017  316]
 [  35 4632]]
```



```
Classification Report for MLR on Test Set (Separable Ring Dataset):
              precision    recall  f1-score   support

           0       0.53      0.99      0.69      5333
           1       0.00      0.00      0.00      4667

    accuracy                           0.53     10000
   macro avg       0.27      0.50      0.35     10000
weighted avg       0.28      0.53      0.37     10000
```



The best parameters for the BP model include using logistic activation function, an alpha value of 0.01, and a hidden layer architecture of (50, 25, 10). The classification error on the training set is 25.40%, indicating an improvement over the SVM model. The precision and recall values for both classes have also improved. The neural network's ability to capture complex relationships within the data and the flexibility offered by tuning parameters contribute to its effectiveness.

```
Best Parameters: {'classifier__activation': 'logistic', 'classifier__alpha': 0.01, 'cl
assifier__hidden_layer_sizes': (50, 25, 10)}
Classification Error: 25.40%
Confusion Matrix:
 [[2816   27]
 [ 989  168]]
Classification Report:
              precision    recall  f1-score   support

           0       0.74      0.99      0.85      2843
           1       0.86      0.15      0.25      1157

    accuracy                           0.75      4000
   macro avg       0.80      0.57      0.55      4000
weighted avg       0.78      0.75      0.67      4000
```

The MLR model, which is a simpler linear model, achieved an accuracy of 74% on the test set. The classification report indicates that while precision for class 0 is reasonable (74%), the precision for class 1 is lower (79%). Similar to SVM, the recall for class 1 is only 15%, indicating challenges in correctly identifying instances of class 1. The classification error on the test set is 25.62%. MLR might not capture the non-linear relationships present in the data as effectively as the neural network.

```
Classification Report (Train):
              precision    recall  f1-score   support

           0       0.74      0.98      0.85      2843
           1       0.79      0.15      0.26      1157

    accuracy                           0.74      4000
   macro avg       0.77      0.57      0.55      4000
weighted avg       0.76      0.74      0.68      4000

Classification Report (Test):
              precision    recall  f1-score   support

           0       0.73      0.98      0.83       687
           1       0.80      0.19      0.31       313

    accuracy                           0.73      1000
   macro avg       0.76      0.59      0.57      1000
weighted avg       0.75      0.73      0.67      1000
```

Conclusions

The analysis and implementation of data classification algorithms on diverse datasets showcased the importance of proper data preprocessing and algorithm parameter tuning. SVM, BP, and MLR demonstrated varying performance across datasets, emphasizing the need for a nuanced approach to model selection and evaluation.

The comprehensive analysis and implementation of data classification algorithms on diverse datasets highlighted the significance of proper data preprocessing, algorithm parameter tuning, and result evaluation. The inclusion of a new dataset, the "School Dataset," provided additional insights into dataset-specific considerations and preprocessing steps.

References

S. Moro, P. Cortez, and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems (2014), doi: 10.1016/j.dss.2014.03.001