Lab Course Machine Learning

Exercise Sheet 7

Prof. Dr. Dr. Lars Schmidt-Thieme
Jung Min Choi

HiWi: Harish Malik

Submission deadline : December 8, 2024 - Extended due to the SRP presentation

General Instructions

1. Data should be normalized.
2. Train to Test split should be 80-20 / with Validaiton 70-15-15
3. Convert any non-numeric values to numeric values. For example you can replace a country name with an integer value or more appropriately use one-hot encoding.

1 SVM with Submanifold Minimization (10 points)

- Class +1:

$$X_{\text{pos}} = \begin{bmatrix} 2.0 & 2.2 \\ 2.7 & 2.5 \\ 2.3 & 2.0 \\ 3.1 & 2.3 \\ 2.5 & 2.4 \\ 2.8 & 2.7 \end{bmatrix}, \quad y_{\text{pos}} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

- Class -1:

$$X_{\text{neg}} = \begin{bmatrix} 1.6 & 1.5 \\ 2.0 & 1.9 \\ 2.1 & 1.8 \\ 1.7 & 1.6 \\ 1.8 & 1.7 \\ 2.0 & 1.6 \end{bmatrix}, \quad y_{\text{neg}} = \begin{bmatrix} -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \end{bmatrix}$$

1. **[8 Points]** Implement a Support Vector Machine (SVM) using the **Submanifold Minimization Algorithm** using the dataset above. Refer to the slides from **Machine Learning 1** for the algorithm. You may use `numpy.linalg.solve` where needed. You can set a `max_iterations` parameter for the while loops.
2. **[2 Points]** Compute and report the following metrics using `sklearn`:
   - Accuracy
   - F1 Score
   - Recall

2 Imbalanced Classification with Sampling Techniques and MLP (10 points)

1. **[2 Points] Dataset Preprocessing and Analysis:**
   a) Download the provided dataset from the following link: **Credit Card Fraud Dataset**.
   b) Split the dataset into training and testing sets using sklearn. No validation set is required.
   c) Analyze the dataset to understand the distribution of classes.
   d) Apply the following resampling techniques to address the class imbalance:

- **SMOTE Oversampling:** Synthesize new samples for the minority class using SMOTE.
- **Random Undersampling:** Reduce the number of majority class samples to match the minority class.

2. **[5 Points] Model Implementation:**
   a) Build a **three-layer fully connected neural network (MLP)** using PyTorch with the following specifications:
      - Use the **ReLU activation function** between layers.
      - For the output layer, use an appropriate activation function for this dataset classification.
      - Select a suitable loss function for this dataset classification.
   b) Use the Adam optimizer with a learning rate of your choice. Note that dropout or normalization layers are not required.
   c) Train the model on the datasets:
      - Original Dataset
      - SMOTE Oversampling
      - Random Undersampling
3. **[3 Points] Evaluation and Analysis:**
   a) Report the following metrics for each model. You can use sklearn.
      - Accuracy
      - Recall
      - F1-score
   b) Compare and analyze the results between the different datasets and sampling techniques. Discuss the trade-offs in performance for each resampling method and their impact on the classification results.

3 **Bonus: SVM with Pegasos (8 points)

You can studyh the basic Pegasos Algorithm from the paper *Pegasos: Primal Estimated sub-GrAdient SOlver for SVM*. Refer to Figure 1 on page 5 of the paper.

1. **[8 Points]** Implement the basic Pegasos Algorithm and report a final accuracy on the test set. You can use the Credit Card Fraud Dataset from the second question (No Sklearn allowed).