Applied Machine Learning
Exercise 5

Prof. Dr. Amr Alanwar

November 2024

Classification Dataset

Bank Marketing Dataset: Use `bank.csv` from `https://archive.ics.uci.edu/ml/datasets/Bank+Marketing`.

Regression Dataset

Wine Quality Dataset: Available at `http://archive.ics.uci.edu/ml/datasets/Wine+Quality`.

Instructions

You are required to pre-process the given datasets as follows:

1. Convert any non-numeric values to numeric values. For example, you can replace a country name with an integer value or use one-hot encoding. *(Hint: use hashmap (dict) or pandas.get_dummies)*. Please explain your solution.
2. If required, drop rows with missing values or NA.
3. Split the data into a train (80%) and test (20%) set.
4. Normalize the data.

Exercise 1: Regularization (4 Points)

For each dataset given above:

1. Implement Ridge Regression using the mini-Batch Gradient Descent (mini-BGD) algorithm. Your algorithm should have three hyperparameters:
   - Learning rate (α)
   - Regularization constant (λ)
   - Batch size (`batchsize`)
2. You can use any strategy for selecting the learning rate, such as AdaGrad, Bold Driver, or a fixed step size.
3. Choose three values for α and λ ranging from small to large. Keep a fixed `batchsize` of 50.

4. Train your model for each combination of the selected values of α and λ. For each training epoch (one pass over all mini-batches), record the RMSE on the training and test data.

5. For each combination of α and λ, plot the RMSE for training and test sets over iterations. [*Hint: Plot $RMSE_{train}$ on the positive axis and $RMSE_{test}$ on the negative axis of the same plot*].

Hyper-parameter Tuning

Exercise 2: Hyper-parameter Tuning and Cross-Validation (6 Points)

In this section, you will implement grid search with k-fold cross-validation for model selection (choosing the best hyperparameters):

1. Pick a range of α and λ values defined on a grid. Use a fixed `batchsize` of 50.

2. Implement the k-fold cross-validation protocol for grid search. For each combination of α and λ, perform k-fold cross-validation with $k = 5$.

3. Keep track of the mean performance (i.e., RMSE value) across k folds for each set of hyperparameters. Plot α vs λ with RMSE scores for all combinations.

4. Finally, for the optimal values of α and λ, train your model on the complete training data and evaluate on the test set.

5. Plot $\text{RMSE}_{\text{train}}$ and $\text{RMSE}_{\text{test}}$ over iterations. Compare your results with those from previous plots.

Hint: If you were unable to complete Exercise 1, you can still attempt Exercise 2 by using the linear regression implementation from Exercise Sheet 3 and adding a regularization term. There will be some penalty for this.

Annex

- You can use numpy or scipy for linear algebra operations.
- You can use pandas for reading and processing data.
- You can use matplotlib for plotting.
- **Do not use any machine learning libraries** (e.g., scikit-learn). If used, you will not receive any points for the task.