Report - EALG

Introduction

This report summarizes the implementation and evaluation of a logistic regression classifier to predict employee attrition based on key factors such as education level, environment satisfaction, job involvement, job satisfaction, performance rating, relationship satisfaction, and work-life balance. The goal is to provide actionable insights for HR analytics, helping organizations understand and address workforce attrition effectively. A custom logistic regression model is built from scratch using only numpy and pandas to verify whether these factors can predict employee attrition (stay/leave).

Dataset

The IBM HR Analytics dataset containing employee information such as demographic details, job satisfaction metrics, performance ratings, and relevant attributes is used for this project. The dataset has features like age, education level, environment satisfaction, job involvement, job satisfaction, performance rating, relationship satisfaction, work-life balance, and more. The target variable is a binary value indicating whether an employee has left the company (attrition) or is still employed.

Approach

The solution implements a custom `LogisticRegressor` class with the following key functionality:

1. Sigmoid activation function
2. Cross-entropy loss calculation
3. Fit method to train the model using mini-batch gradient descent
4. Predict method to make predictions on new data

The dataset is first preprocessed by encoding categorical features (one-hot encoding for non-binary, binary encoding for binary) and normalizing numerical features using z-score normalization. The preprocessed data is then split into training and test sets (80:20 ratio).

Hyperparameters

The `LogisticRegressor` is instantiated with the following hyperparameters:

- `learning_rate` = 0.1
- `num_iterations` = 250
- `batch_size` = 128

Learning Algorithm

The fit method performs the following steps for each iteration:

1. Shuffle the training data
2. Split data into mini-batches of specified batch_size
3. For each mini-batch:
    a. Perform forward propagation
    b. Calculate cross-entropy loss
    c. Perform backward propagation to compute gradients
    d. Update weights and bias using gradients
4. Compute and store average loss for the iteration

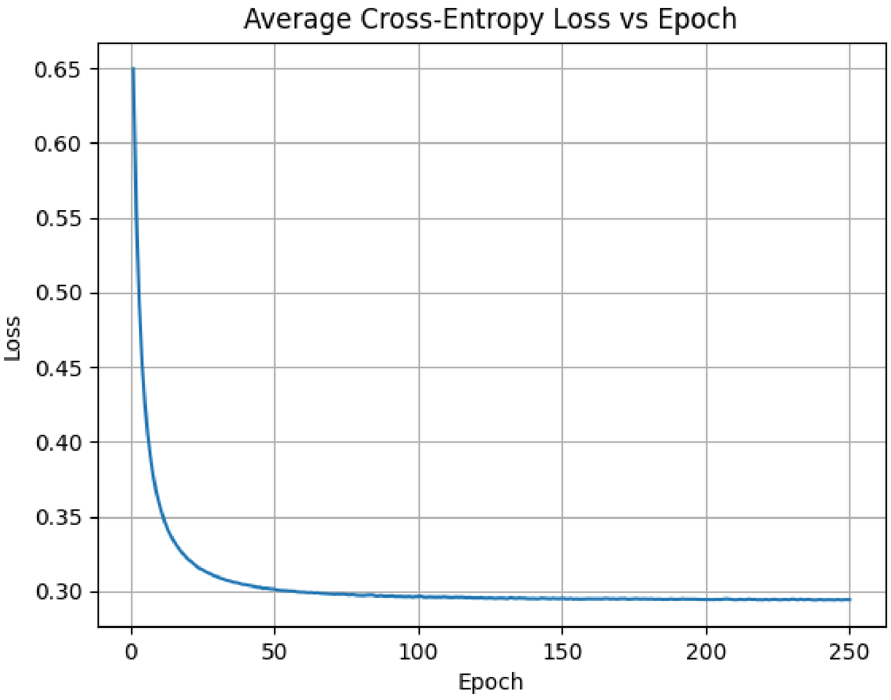
After training, the predict method uses the learned weights and bias to make predictions on the test set by applying the sigmoid activation.

Results

The classification report metrics computed on the test set are as follows:

- Precision: 0.46153846153846156
- Recall: 0.5806451612903226
- F1-score: 0.5142857142857142
- Accuracy: 0.8843537414965986

Below is the progression of average cross-entropy loss with every epoch:


Average Cross-Entropy Loss vs Epoch
Loss
Epoch