Mandar Angchekar
386916341

Intro to Machine Learning and Algorithmns

HW05 Report

- **Objective**: To classify citation data into three categories based on the ratio of citations in consecutive years (2022/2021).
- **Dataset**:
  - Utilized a dataset containing citation counts from 2017 to 2022.
  - The **ratio** column was calculated by dividing the citation count of 2022 by 2021.
  - Categories were defined based on the **ratio**:
    - Class 1: **ratio** < 1.05
    - Class 2: 1.05 ≤ **ratio** ≤ 1.15
    - Class 3: **ratio** > 1.15
- **Preprocessing**:
  - Normalized citation count data for better model performance.
  - Implemented one-hot encoding for categorical labels to facilitate the classification.
- **Model Architecture**:
  - Designed a neural network with two dense layers, using ReLU and softmax activation functions for non-linear transformation and probability distribution respectively.
- **Training Process**:
  - Split the data into 80% training and 20% testing to validate the model.
  - Employed the cross-entropy loss function suitable for multi-class classification.
  - Used ADAM optimizer for efficient and adaptive parameter updates.
- **Evaluation**:
  - Achieved an accuracy of 75% on the test data, indicating a reasonably good model for the given task.
  - However, the model displayed limitations with specific classes, having a precision and recall of 0.0 for category 2, indicating no correct predictions for this category.
- **Insights & Comments**:
  - The model performed well for category 1 and moderately for category 3 but failed to identify any instances of category 2 correctly. This is be due to:
    - Imbalanced dataset with insufficient examples of category 2.
    - Model architecture might be too simple to capture the complexity needed for accurate predictions across all categories.
  - Shuffling the data can result in different train-test splits, potentially affecting model performance and the balance of classes within each split.