**CIS 662: Intro to Machine Learning and Algorithms**

**HW-6**

1. **Problem Statement:**

Use the same data set for

**1. Linear regression:** Fit a line to go very near the 2017-2021 citation columns, minimizing MSE.  
Use that line to predict the 2022 citation numbers, as in HW4.

**2. Logistic regression:** Classify individuals into 3 categories, as in HW5.

How do the results compare with those of HW4 and HW5, respectively?

Your submission should include:  
1. code   
2. output files (predictions in CSV format)  
3. report\_HW6.pdf (comparison of results and your conclusion)

1. **Approach:**

**Linear Regression:**

* The linear regression model is trained using the citation numbers from 2017 to 2021 as features and 2022 citations as the target variable.
* The model is evaluated using Mean Absolute Error (MAE), a measure of the average absolute errors between the predicted and actual values.
* Predictions are made for the 2022 citation numbers.

**Logistic Regression:**

* The logistic regression model is trained to classify individuals into three categories based on the given criteria.
* The features include citation numbers from 2017 to 2021.
* The model is evaluated using accuracy, precision, recall, and F1-score.
* Predictions are made for the category of each individual.

1. **Results:**

**A close up of numbers

Description automatically generated**

**Linear Regression:**

* MAE for Linear Regression: 81.12.
* The model provides predictions for 2022 citation numbers.

A table with numbers and letters

Description automatically generated

**Logistic Regression:**

* Accuracy for Logistic Regression: 60.00%.
* The model classifies individuals into three categories: Low, Medium, and High based on the given criteria.

A table of numbers and numbers

Description automatically generated

**Comparison with Previous Results:**

**Linear Regression (HW4):**

* In HW4, various optimizers (SGD, Adam, RMSprop) were tested with different learning rates and batch sizes.
* The best-performing combination was using the Adam optimizer with a learning rate of 0.1 and a batch size of 4.
* The MAE obtained in HW4 was 158.74.

**Logistic Regression (HW5):**

* In HW5, a 6-6-3 neural network was used for classification into Low, Medium, and High categories.
* The model achieved an accuracy of 80% on the test set.
* The classification report provided insights into precision, recall, and F1-score for each category.

1. **Conclusion:**

**Linear Regression:**

* The linear regression model yielded a significantly lower MAE (81.12) compared to the best result obtained in HW4 (158.74).
* This indicates that linear regression provides a more accurate prediction of 2022 citation numbers compared to the neural network approach used in HW4.

**Logistic Regression:**

* The logistic regression model achieved an accuracy of 60.00%.
* While the accuracy is lower than the neural network in HW5 (80%), logistic regression provides a simpler and interpretable model for classification.

**Overall:**

* The choice between linear regression and logistic regression depends on the nature of the problem and the specific goals. Linear regression is suitable for predicting numerical values, while logistic regression is appropriate for classification tasks.
* The results highlight the trade-off between complexity and interpretability in choosing a modelling approach.

**References:**

<https://chat.openai.com/>,

<https://keras.io>

<https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html>

<https://towardsdatascience.com/machine-learning-classification-52241849468a>

<https://scikit-learn.org/stable/modules/model_evaluation.html>