Assignment 1: Linear and Logistic regression implementation and Mathematical intuition

Pravin Lamichhane

September 2025

1 Introduction

This assignment covers the impementation and mathematical intution for linear and logistic regresssion .

2 Linear Regression

Linear regression is a statistical method used to model the relationship among variables. It assumes a linear relationship between the input variables (features) and the output variable (target). The goal of linear regression is to find the best-fitting line that minimizes the difference between the predicted values and the actual values. Supervised learning algorithm linear regression calculates cost functon using mean square error(MSE), which is the average of the squared differences between the predicted value and true vale of the target variable for each data point, and further minimize using algorithm like gradient descent.

2.1 Code Implementation

```
from sklearn import datasets
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split

def main():
    diabetes = datasets.load_diabetes()
    linear_reg = LinearRegression()
    X = diabetes.data
    y = diabetes.target
    X_train , X_test, y_train ,y_test =
    train_test_split(X,y , test_size=0.2)
    linear_reg.fit(X_train,y_train)
    predictions = linear_reg.predict(X_test)
    print(predictions[:5].round(),y_test[:5])
```

```
    print(linear_reg.score(X_test, y_test))
if __name__ == "__main__":
    main()
```

2.2 Mathematical Intuition

The cost function for linear regression is given by: Mean Squared Error (MSE)

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \tag{1}$$

Here error(predicted value - actual value) is squared to avoid negative ,peanilized larger error , to derive in further optimization .to avoid cost function being bigger we take average by dividing by number of samples m .

3 Logistic Regression

Logistic regression is a statistical method used for binary classification problems, where the goal is to predict one of two possible outcomes based on input feature. An threshold value is used to classify the output into two classes. The model use sigmoid function to map predicted values to like probabilities 0-1.

3.1 Code Implementation

```
    from sklearn import datasets
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression

def main():
    breast_cancer = datasets.load_breast_cancer()
    logistic_reg = LogisticRegression()
    X = breast_cancer.data
    y = breast_cancer.target
    X_train , X_test , y_train , y_test = train_test_split(X, y , test_size = 0.
    logistic_reg.fit(X_train, y_train)
    logistic_reg.predict(X_test)
    print(logistic_reg.score(X_test, y_test))

if __name__ == "__main__":
    main()
```

3.2 Mathematical Intuition

Logistic regression uses sigmoid function to map predicted values to probabilities between 0 and 1. The sigmoid function is defined as:

$$z = \frac{1}{1 + e^{-z}} \tag{2}$$

where Z is the linear combination of features and weights (i.e., y = mx + b).when Z approraches larger positive value output approaches 1 and vice versa.

The cost function for logistic regression is given by: Log Loss (Cross-Entropy Loss) cause while using MSE the cost function is not convex as in linear regression, which causes vanishing gradient , making it difficult to optimize the model using gradient descent. Log Loss is defined as :

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left[L\left(f\theta(x^{(i)}, y^{(i)}) \right) \right] \tag{3}$$

$$L\left(f\theta(x^{(i)}, y^{(i)}) \right) = \left(-Log(f_\theta(x^{(i)})) \text{ if } y^{(i)} = 1 \right) or \left(-Log(1 - f_\theta(x^{(i)})) \text{ if } y^{(i)} = 0 \right) \tag{4}$$

Here f(x(i)) is the predicted probability that the output is 1 given input features x(i) and model parameters . The cost function penalizes incorrect predictions more heavily whhich generate a convex function that is easier to optimize using gradient descent.

4 Conclusion

In conclusion, linear regression is used for predicting continuous values, while logistic regression is used for binary classification .