1. 과제 설명

이번 과제는 직접 날씨를 예측하는 머신러닝 모델을 만드는 것이다. 최고 기온을 예측할 때는 Linear Regression을 사용하고, 날씨를 분류할 때는 Logistic Regression을 사용한다. 데이터 세트에는 측정 날짜, 강수량, 최고 기온, 최저 기온, 바람 세기, 날씨 총 6개의 관련 데이터값이 들어있다. 그림, 표 등을 포함하며 각 알고리즘을 분석하고 평가하는 내용을 보고서로 작성해 제출해야 한다.

2. 각 알고리즘 분석과 결과

2-1. Linear Regression

Linear Regression을 사용하여 최고 기온을 예측하는 모델을 만들었다. X값에 들어가는 칼럼은 [강수량, 최저 기온, 바람]이고 Y 값에 들어가는 칼럼은 [최고 기온]이다. 위에서 주어진 데이터 세트를 train data 80%, test data 20%로 나누고 test data에 대한 성능을 평가한다.

우선 Linear Regression은 연속적인 수치형 값을 예측하는 모델이다. 해당 모델은 수많은 데이터를 통해 값을 예측하고 이를 기반으로 실제 값을 나타내는 선 또는 면과 얼마큼 차이가 있는지를 바탕으로 모델의 성능을 평가한다.

제출한 코드에는 MSE와 RMSE가 있다. 먼저 MSE는 Mean Squared Error이고, "예측값과 실제 값의 오차 제곱 평균값"을 뜻한다.

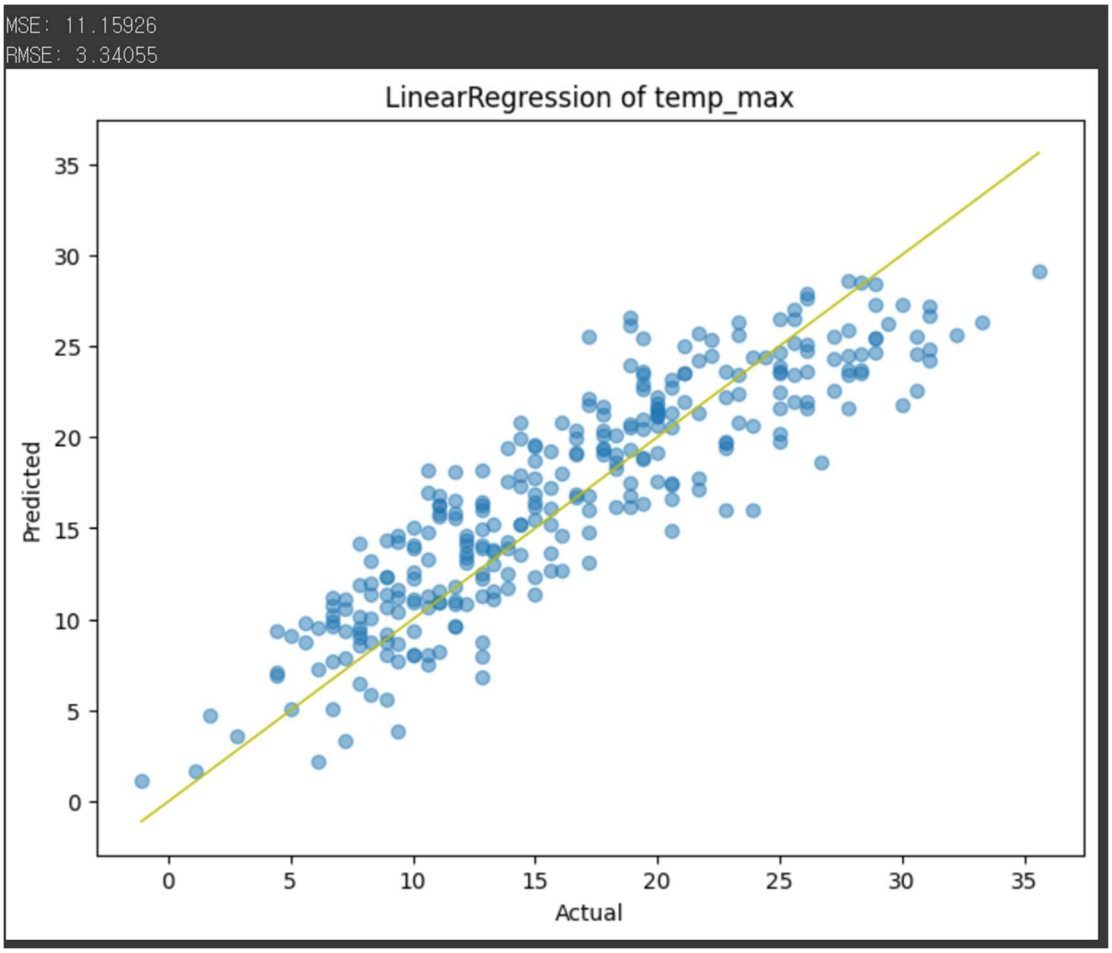
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y}_u)^2$$

이때, MSE 값이 작을수록 예측이 잘 맞는다고 할 수 있다. 반면에 단위가 y값의 제곱이므로 직관적인 해석이 어렵다.

다음으로, RMSE는 Root Mean Squared Error이고, MSE에 루트를 씌운 것이다.

$$RMSE = \sqrt{MSE}$$

RMSE 값이 작을수록 좋은 모델이고, 실제로 에러의 평균 크기가 얼마인지 알 수 있다. 또한 MSE보다 더 직관적이다는 장점이 있다.



내가 구현한 알고리즘을 통해 실행해 본 결과 RMSE : 3.34055로 적당한 예측 성능을 보여준다. 위의 그래프를 보면, 파란색 점과 노란색 선이 있다. 노란색 선은 정확한 값을 의미한다. 파란색 점은 모델에서 데이터를 기반으로 예측했을 때 실제 값을 나타낸다. 이 점들이 노란색 선에 가까울수록 잘 예측한 경우이고 멀어질수록 오차가 크다. 시각화된 그래프를 보면 점들이 선에 가깝게 배치된 것을 알 수 있다.

2-2. Logistic Regression

Logistic Regression을 사용하여 날씨를 분류하는 모델을 만들었다. X값에 들어가는 칼럼은 [강수량, 최고 기온, 최저 기온, 바람]이고 Y 값에 들어가는 칼럼은 [날씨]다. 이때 [날씨]가 sun인 경우 1, 그 외에는 0으로 변환한다. 위에서 주어진 데이터 세트를 train data 80%, test data 20%로 나누고 test data에 대한 성능을 평가한다.

Logistic regression은 클래스 값을 분류하는 모델을 뜻한다. 해당 모델은 얼마나 정확하게 각 예측 데이터를 알맞은 클래스로 분류했는지로 성능을 판단한다. 0~1 사이의 값을 출력한다. 이를 위해 아래와 같은 시그모이드 함수를 사용한다. 나온 값으로 이진 분류 또는 3개 이상의 클래스를 분류한다.

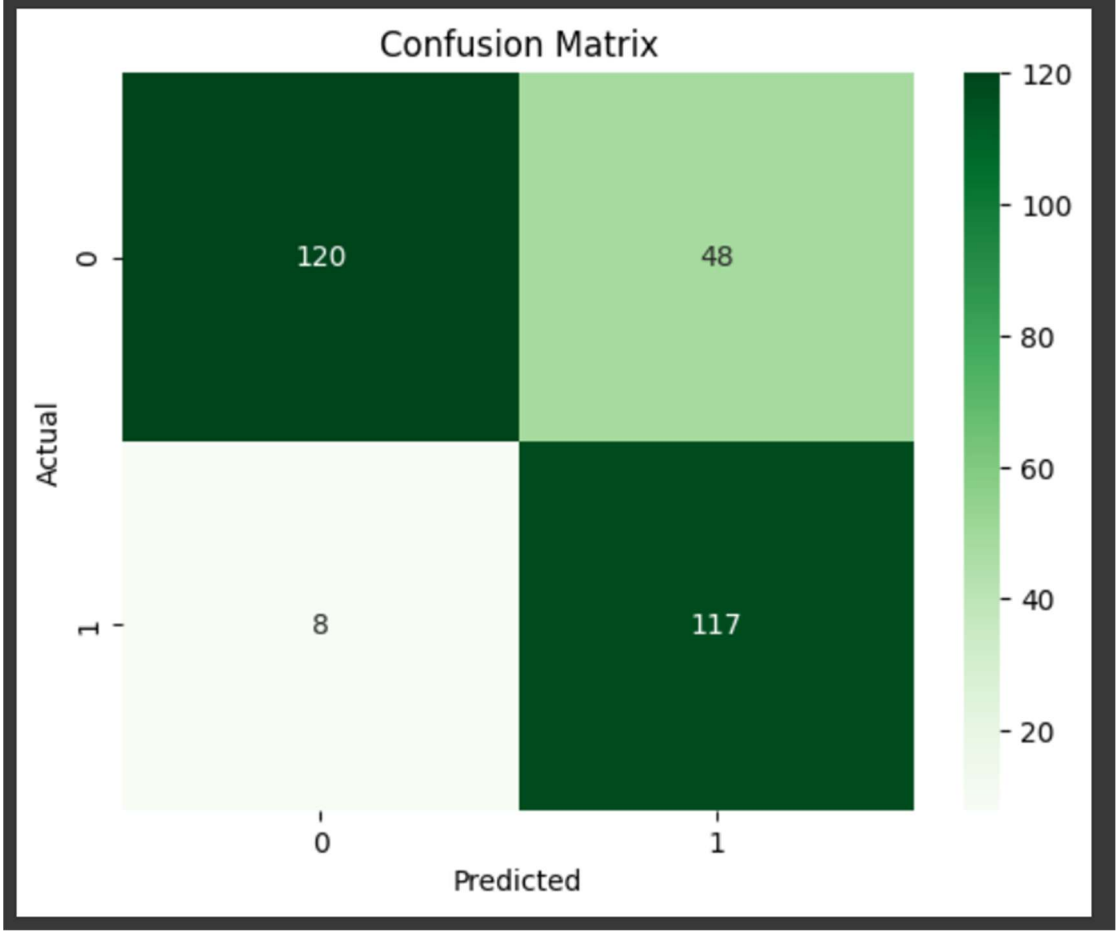
$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

```
Logistic Regression
Accuracy: 0.80887

Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.71      0.81       168
           1       0.71      0.94      0.81       125

    accuracy                           0.81       293
   macro avg       0.82      0.83      0.81       293
weighted avg       0.84      0.81      0.81       293
```



작성한 알고리즘에는 성능을 확인하기 위해서 Accuracy를 구하여 확인하는 코드를 넣었다. 0.80887로 약 80%의 정확도를 보인다. 또한, 얻은 데이터를 시각화하여 아래 그래프를 통해 각 클래스에 얼마나 알맞게 분류하는지 나타냈다. 진한 녹색 부분이 정확히 분류한 경우이고, 흰색과 연한 녹색은 잘못 분류한 경우다. 0(sun)인 경우는 precision이 0.94로 정확하게 분류했지만, 1(sun이 아닌 경우)는 0.71로 비교적 잘못 분류한 경우가 많았다.

3. 결론 및 개선점

Linear Regression과 Logistic Regression의 공통점으로는 선형 모델이다. 하지만 사용 방법과 결과 확인에서 차이를 보인다. 먼저 사용 목적으로는 Linear Regression은 연속적인 수치 예측, Logistic Regression은 2개 이상의 클래스 분류다. 각각 직선 함수와 시그모이드 함수를 사용하며, 산점도 + 예측선과 결정 경계 + 분류 영역을 통하여 시각화한다. 마지막으로 사용 예는 Linear Regresion의 경우, 위의 예제와 같이 온도 예측이나 가격 예측에 사용되며, Logistic Regression의 경우, 날씨 분류나 꽃의 품종 분류에 사용된다. 머신러닝의 첫 과제인 만큼 다양한 곳에서 정보를 찾아보면서 과제를 진행했다. Linear Regression의 경우는 데이터가 목표선 근처에 잘 형성되었지만, Logistic Regression의 경우 정확도가 90%가 넘지 않았다. 또, sun인 경우에 sun이 아니라고 판단한 경우가 48개나 있어서 알고리즘이 개선이 필요하다고 느꼈다. 앞으로의 수업을 통해 조금씩 개선하겠다.