4_Logistic_Regression

문제 정의

- 로지스틱 회귀는 **주어진 입력 변수들과 이산적인 목표 변수 간의 관계를 모델링하는 통계적 방법**이다. 이 방법은 주로 분류 문제에 사용되며, 선형 회귀와는 달리 **이산형 결과를 예측**한다. 로지스틱 회귀의 핵심 요소는 **로지스틱 또는 시그모이드 함수**로, 이 함수는 0과 1 사이의 값을 출력하는 S자 모양의 곡선이다. 입력 변수들의 선형 결합을 시그모이드 함수에 입력함으로써 결과의 확률을 계산한다. **로지스틱 회귀는 선형 회귀와 달리 볼록 함수(convex function) 형태**를 가지기 때문에, **최적의 모델 파라미터를 찾기 위해 경사 하강법과 같은 반복적인 최적화 알고리즘이 필요**하다. 이러한 특성 덕분에 로지스틱 회귀는 다양한 분류 문제에서 유용하게 활용된다. 특히 이진 분류에서 좋은 성능을 보인다.

해당 문제에 대한 일반적인 접근

- 로지스틱 회귀는 단일 로지스틱 회귀와 다중 로지스틱 회귀로 나뉜다. **단일 로지스틱 회귀**는 **하나의 독립 변수를 사용하여 종속 변수의 이진 결과를 예측**한다. 활성화 함수로는 **시그모이드 함수**를 사용하여 입력 값을 0과 1 사이의 값으로 변환하며, 이 확률은 종속 변수가 특정 클래스에 속할 가능성을 나타낸다. **비용 함수(Cost function)**로는 **로그 손실 함수를 사용하여 모델의 예측이 실제 값과 얼마나 잘 맞는지 측정하고, 이를 기반으로 모델의 가중치를 경사 하강법을 통해 최적화**한다.

  다중 로지스틱 회귀는 **두 개 이상의 독립 변수를 사용하며, 종속 변수가 여러 클래스 중 하나에 속할 확률을 예측**한다. 다중 로지스틱 회귀에서는 **소프트맥스 함수**를 활성화 함수로 사용하여 각 클래스에 대한 확률을 출력하고, 여기에도 로그 손실 함수를 적용하여 모델을 최적화한다. 이 과정에서도 경사 하강법이 주로 사용되며, 각 파라미터의 업데이트는 이전 단계의 정보를 고려하여 이루어진다.

일반적인 접근법의 제한 사항

- 로지스틱 회귀는 주로 **선형 결정 경계를 사용하여 데이터를 분류**한다. 이는 입력 변수와 종속 변수 사이의 관계가 선형적일 때 잘 작동하지만, **데이터 간의 관계가 복잡하고 비선형적인 경우 성능에 한계**를 보일 수 있다. 예를 들어, 데이터가 복잡한 패턴을 보이거나 변수 간 상호작용이 중요한 역할을 할 때, 로지스틱 회귀만으로는 이를 효과적으로 모델링하기 어렵다. 이런 경우, 모델은 **실제 데이터 구조를 제대로 반영하지 못하여 과적합이 발생**할 수 있다.

제한 사항에 대한 해결 방안

- 로지스틱 회귀 모델에 **L1 또는 L2 규제를 추가하는 것은 과적합을 방지하고 모델의 일반화 성능을 향상**시키는 데 큰 도움이 된다. L1 규제는 가중치의 절대값에 대한 페널티를 부과하는 방식으로 작동한다. 이 방법은 특정 가중치를 완전히 0으로 만들 수 있기 때문에, 불필요한 특성을 모델에서 제거하는 효과가 있다.
  반면, L2 규제는 가중치의 제곱에 대한 페널티를 부과한다.(수식 포함) 이 규제는 모든 가중치를 작게 유지하여 모델의 복잡성을 줄이는데 기여하며, 이로 인해 모델이 훈련 데이터에 덜 민감하게 되어 과적합의 위험을 감소시킨다. L2 규제는 가중치 값들이 완전히 0이 되는 것을 피하면서도 전체적으로 작은 값들로 분포하게 한다.
  로지스틱 회귀에서 이러한 규제 방법을 적용하면, 각기 다른 문제 상황에 맞게 모델을 더욱 효과적으로 조정할 수 있다.