머신러닝 - 회귀

> 여러 개의 독립변수와 한 개의 종속변수 간의 상관관계를 모델링하는 기법
>
> - 최적의 회귀 모델을 만드는 것은 실제 값과 회귀 모델의 차이에 따른 오류 값(잔차)의 합이 최소가 되는 모델을 학습시키는 것
> - 회귀가 선형인가 비선형인가는 독립변수가 아닌 가중치(Weight) 변수가 선형인지 아닌지에 따라 다름

Y = *W1*X1 + *W2*X2 + *W3*X3 + ... + *Wn*Xn

- Y : 종속변수
- X : 독립변수
- W : 회귀 계수 (Regression coefficients)

전체 데이터 오류 합 계산 방식

- Mean Absolute Error : 단순히 오류들에 절댓값을 취하여 합하는 방식
- RSS (Residual Sum of Square) : 각 오류 값의 제곱을 더하는 방식

회귀 평가 지표

1. MAE (Mean Absoulte Error) - 실제 값과 예측값의 차이를 절댓값으로 변환해 평균
2. MSE (Mean Squared Error) - 실제 값과 예측값의 차이를 제곱해 평균
3. RMSE (Root Mean Squared Error) - MSE 값이 실제 오류 평균보다 더 커지는 문제를 막기위해 MSE에 루트를 씌우는 것
4. R^2 - 분산 기반으로 예측 성능 평가, 1에 가까울수록 예측 정확도가 높아짐, 예측값 Variance / 실제값 Variance

---

경사 하강법 (Gradient Descent)

Linear Regression

> RSS 를 최소로 하는 회귀 계수를 학습을 통해서 찾는 것

다항 회귀 (Polynomial Regression)

> 회귀가 독립변수의 단항식이 아닌 2차, 3차 방정식과 같은 다항식으로 표현되는 것

Y = *W1*X1 + *W2*X2 + W3 *X1* X2 + ... + *W5*X2^2

다항 회귀를 이용한 과소적합 및 과적합 이해

> 다항 회귀 (Polynomial Regression) 는 피처의 직선적 관계가 아닌 복잡한 다항 관계를 모델링 - 다항식의 차수가 높아질수록 복잡한 피처 간의 관계도 모델링이 가능
>
> 하지만 다항 회귀의 차수(degree)가 높아질수록 학습 데이터에만 너무 맞춘 학습이 이뤄져서 테스트 데이터 환경에서는 오히려 예측 정확도가 떨어지는 경우가 발생 (과적합의 문제)

- 편향-분산 트레이드오프 (Bias-Variance Trade off)
  - 편향 : 예측 결과가 실제 결과에 근접하는 지를 나타내는 단어 (편향이 클수록 오류가 큼)
  - 분산 : 예측 결과의 변동성을 나타내는 단어
  - 일반적으로, 편향과 분산은 한 쪽이 높으면 한 쪽인 낮아지는 경향이 있음
    - 편향이 높으면 분산은 낮아지고 (**과소적합**), 분산이 높으면 편향이 낮아짐 (**과적합**)

Regularized Linear Models

1. Ridge
   - 선형 회귀에 L2규제를 추가한 회귀 모델, L2 규제는 상대적으로 큰 회귀 계수 값의 예측 영향도를 감소시키기 위해 회귀 계수값을 더 작게 만드는 규제 모델 (RSS + alpha * ||W||22)
2. Lasso
   - 선형 회귀에 L1규제를 추가한 회귀 모델, L1 규제는 예측 영향력이 작은 피처의 회귀 계수를 0으로 만들어 회귀 예측 시 피처가 선택되지 않게 하는 것으로 피처 선택 기능으로도 불림 (RSS + alpha * ||W||1)
3. 엘라스틱넷(Elastic Net) 회귀
   - L2, L1 규제를 함께 결합한 모델, 주로 피처가 많은 데이터 세트에 적용되며, L1 규제로 피처의 개수를 줄임과 동시에 L2 규제로 계수 값의 크기를 조정

선형 회귀 모델을 위한 데이터 변환 처리

> 선형 모델은 피처와 타겟값 간에 선형의 관계가 있다고 가정
>
> 선형 회귀 모델은 피처와 타겟값의 분포가 정규 분포를 따르는 것을 선호
>
> 특히, 타겟값의 경우 정규 분포 형태가 아니라 특정값으로 분포가 치우친 왜곡(Skew)된 형태의 분포도인 경우 예측 성능에 부적적인 영향을 미칠 가능성이 큼
>
> => **이러한 이유로, 선형 회귀 모델을 적용하기 전에 먼저 데이터에 대한 스케일링/정규화 작업을 수행하는 것이 일반적**
>
> => 특히, 중요 피처들이나 타깃값의 분포도가 심하게 왜곡됐을 경우에 변환 작업을 수행 (그렇지 않은 경우에 변환 작업을 한다고 해서 무조건 예측 성능이 향상되는 것은 아님)

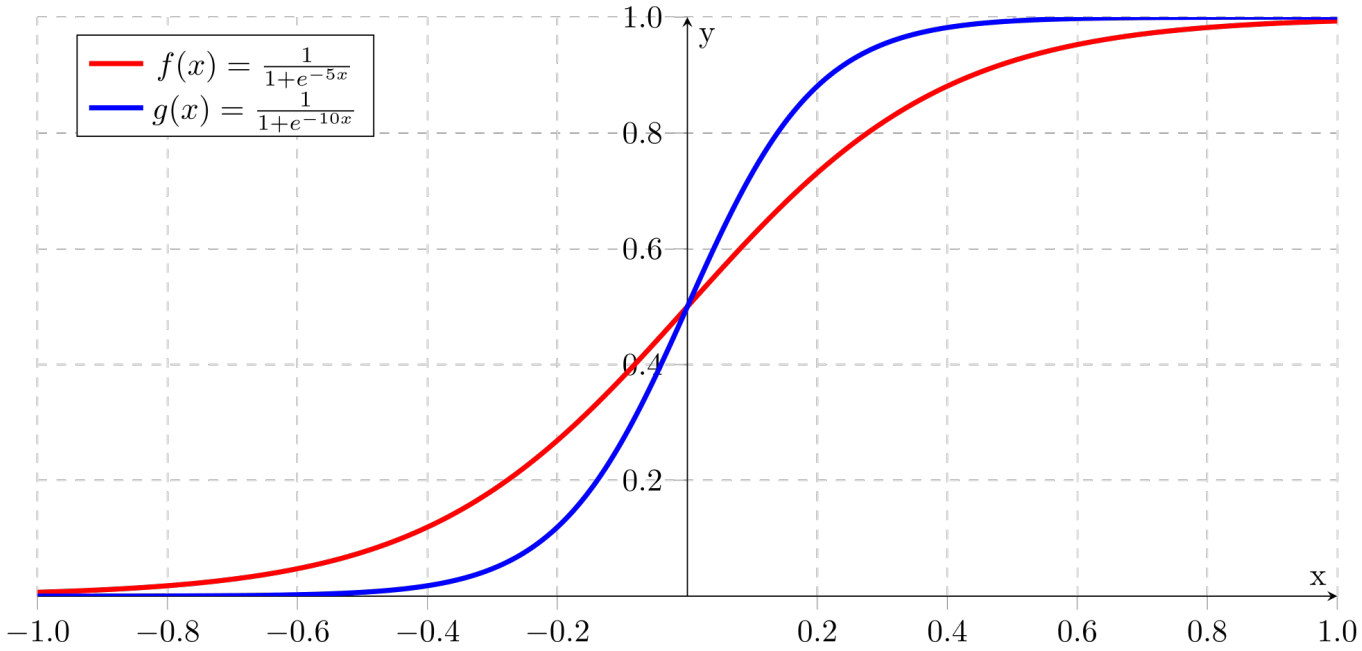
방법

1. 표준 정규 분포 변환 (Standard)
2. 최댓값/최솟값 정규화 (MinMax)
3. 로그 변환 (Log)
   - 일반적으로, 선형 회귀 적용 데이터 세트가 심하게 왜곡되어 있을 경우에 로그 변환을 적용하는 것이 결과가 좋은 편임

로지스틱 회귀 (Logistic Regression)

> 선형 회귀 방식을 분류에 적용한 알고리즘으로 분류에 사용

- 시그모이드 함수 (Sigmoid)



- 선형 회귀 방식을 기반으로 시그모이드 함수를 이용해 분류를 수행하는 회귀 -> 시그모이드 함수의 최적선을 찾고, 이 함수의 반환 값을 확률로 간주해 확률에 따라 분류를 결정하는 것
- 로지스틱 회귀는 가볍고 빠르지만, 이진 분류 예측 성능도 뛰어나기 때문에 이진 분류의 기본 모델로 사용하는 경우가 많음, 또한 희소한 데이터 세트 분류에도 뛰어난 성능을 보여서 텍스트 분류에서도 자주 사용

회귀 트리

> 회귀를 위한 트리를 생성하고 이를 기반으로 회귀 예측을 하는 것

- 트리 생성이 CART(Classification And Regression Trees) 알고리즘에 기반하고 있는 건 분류뿐만 아니라 회귀도 가능하게 해주는 트리 생성 알고리즘 (ex, 결정 트리, 랜덤 포레스트, GBM, XGBoost, LightGBM)