5_Linear_Regression

1) 문제 정의

- 선형 회귀 모델에서는 **입력 변수에 대한 응답 변수가 정규분포의 평균으로 가정**된다. 정규 분포의 '평균'이 '입력 변수에 대한 응답 변수', '분산'이 '오차 정도'가 되는 것이다. 선형회귀에서는 **최소제곱회귀(Least Square Regression, LSR)를 사용하여 모델을 최적화**하는데, 이는 **잔차(오차)의 분산이 입력변수에 따라 동일하다는 가정**을 포함한다. 따라서 응답 변수의 확률 분포는 다음과 같다. (N 안에 y_j임 y_i 아님)

$$p(y_i \mid x, W) = \prod_{j=1}^{J} \mathcal{N}(y_i \mid w_j^T x, \sigma_j^2)$$

- 최소제곱회귀(Least Square Regression, LSR)에 데이터를 학습시키기 위해, **목적 함수 NLL(Negative Log Likelihood)을 최소화**한다. **선형 회귀 NLL 식에는 가우시안 분포 식이 포함**된다. 로지스틱 회귀에서와 달리 NLL 식에서 **데이터 포인트 개수 N으로 나누어주지 않은 이유는, 선형 회귀에서는 데이터 포인트들이 어떻게 연관되어있는지를 찾는 것이 목적이기 때문**이다.

$$\mathrm{NLL}(w, \sigma^2) = -\sum_{n=1}^{N} \log\left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_n - w^T x_n)^2}{2\sigma^2}\right)\right]$$

$$= \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \hat{y_n})^2 + \frac{N}{2} log(2\pi\sigma^2)$$

- 위 식에서 왼쪽은 RSS(Residual Sum of Square), 오른쪽은 분산에 해당하며, **분산이 입력값에 따라 달라지지 않는 상수**임을 확인할 수 있다.

2) 해당 문제에 대한 일반적인 접근법

- 선형회귀의 최소제곱회귀(LSR, Least Square Regression)에서는 주로 w에 대한 최적화를 한 다음에 σ에 대한 최적화를 수행한다. 최적의 파라미터를 찾기 위한 방법은 크게 두 가지이다.

1. 첫 번째는, NLL을 미분했을 때 0이 되는 지점을 찾는 것이다. $\nabla_{w,\sigma} NLL(w, \sigma^2) = 0$ 이러한 '수치적' 방법은 대규모 데이터를 다룰 때 적합하다.
2. 두 번째는 잔차 제곱의 합(RSS, Residual Sum of Square)을 최소화하는 것이다. 잔차 벡터는 X의 모든 열과 직교여야한다. $X^T(y - Xw) = 0 \quad \rightarrow \quad w = (X^T X)^{-1} X^T y$. 이 방식을 정규방정식을 통한 최적화라고도 하며, 계산 비용이 적어 적은 데이터셋을 다룰 때 적합하다.

3) 일반적인 접근법의 제한사항(from 논문 or 본인생각)

OLS(Ordinal Least Square)은 LSR(Least Square Regression) 중에서도 응답 변수가 순서형인 경우에 적합한 회귀 분석 기법이다. 다음과 같은 경우는 OLS가 적합하지 않을 수 있다.

1. 첫 번째는 지수 그래프인 경우이다. **OLS는 입력 변수와 응답 변수 간의 선형 관계를 가정하기 때문**이다.
2. 두 번째는 **오차의 분산이 일정하지 않고 입력 변수에 따라 변하는 경우이다. OLS는 오차의 분산이 동일하고 정규 분포를 따른다는 가정을 기반**으로 하기 때문이다.
3. 세 번째는 **응답 변수가 순서형이 아닌, 범주형**인 경우이다. OLS는 연속적인 응답 변수를 다루는데 적합하기 때문이다.

4) 제한사항에 대한 해결방안(from 논문 or 본인생각)

GLM(Generalized Linear Model)을 사용하여 일반적인 선형 모델을 다룰 수 있다. GLM은 보통 3가지 요소로 구성된다.

1. 첫 번째는 **선형 예측자(linear predictor)**이다. **입력 변수와 그에 상응하는 가중치를 선형적으로 결합하여 응답 변수를 계산**하는 부분이며, Least Square Regression과 유사하다.
2. 두 번째는 **log link function**이다. linear predictor은 **응답 변수의 스케일, 입력 변수와의 선형 관계를 보장하기 위해 링크 함수를 통해 변환**된다. 여러가지 link function 중, log link function을 사용하면 General Linear Model이 된다. identify link function을 사용할 경우 linear regression, logic link function을 사용할 경우 logistic regression이 되는 것이다.
3. 세 번째는 **확률 분포를 포아송 분포로 사용하는 것**이다. 포아송 분포는 분산과 평균이 λ로 동일하다. λ가 커질수록 확률분포의 모양이 오른쪽으로 치우치게 된다. (=점점 왼쪽으로 긴 꼬리(long-tailed)를 가지는 형태가 된다.) 그래서 입력 변수 **x가 증가함에 따라 분산이 증가하는 경우, 응답 변수의 확률 분포를 나타내는 데 Poisson Regression을 사용하는 것이 적합**하다. (Normal(=Gaussian) Distribution을 사용하면 Linear Regression, Binomial/Bernoulli Distribution을 사용하면 Logistic Regresion이 되는 것이다.)