로지스틱 회귀분석

로지스틱 회귀분석 이란

로지스틱 회귀(영어: logistic regression)는 D.R.Cox가 1958년에 제안한 확률 모델로서 독립 변수의 선형 결합을 이용하여 사건의 발생 가능성을 예측하는데 사용되는 통계 기법이다.

로지스틱 회귀의 목적은 일반적인 회귀 분석의 목표와 동일하게 종속 변수와 독립 변수간의 관계를 구체적인 함수로 나타내어 향후 예측 모델에 사용하는 것이다. 이는 독립 변수의 선형 결합으로 종속 변수를 설명한다는 관점에서는 선형 회귀 분석과 유사하다.

하지만 로지스틱 회귀는 선형 회귀 분석과는 다르게 종속 변수가 범주형 데이터를 대상으로 하며 입력 데이터가 주어졌을 때 해당 데이터의 결과가 특정 분류로 나뉘기 때문에 일종의 분류 (classification) 기법으로도 볼 수 있다.

흔히 로지스틱 회귀는 종속변수가 이항형 문제(즉, 유효한 범주의 개수가 두개인 경우)를 지칭할 때 사용된다. 이외에, 두 개 이상의 범주를 가지는 문제가 대상인 경우엔 다항 로지스틱 회귀 (multinomial logistic regression) 또는 분화 로지스틱 회귀 (polytomous logistic regression)라고 하고 복수의 범주이면서 순서가 존재하면 서수 로지스틱 회귀 (ordinal logistic regression) 라고 한다.

로지스틱 회귀 분석은 의료, 통신, 데이터마이닝과 같은 다양한 분야에서 분류 및 예측을 위한 모델로서 폭넓게 사용되고 있다

다중 선형 회귀분석

다중선형회귀(Multiple Linear Regression)는 수치형 설명변수 X와 연속형 숫자로 이뤄진 종속변수 Y 간의 관계를 선형으로 가정하고 이를 가장 잘 표현할 수 있는 회귀계수를 데이터로부터 추정하는 모델입니다.

이 회귀계수들은 모델의 예측값과 실제값의 차이, 즉 오차제곱합(error sum of squares)을 최소로 하는 값들입니다.

이를 만족하는 최적의 계수들은 회귀계수에 대해 미분한 식을 0으로 놓고 풀면 명시적인 해를 구할 수 있습니다

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

로지스틱 파라미터 추정

최대우도추정법

예를들어 데이터가 임의의 파라메터 θθ에 의존하는 확률분포를 따르고 표본데이터 v1,v2,…,vnv1,v2,…,vn이 주어졌다고 해보겠습니다.

p(v1,v2,.....vn|θ)를 알고 싶은데 θ에 대해서 모르는 상황

이럴때는 θ의 우도(likelihood)를 이용해 추정할수 있다.

다시 말해 조건절과 결과절을 뒤집어 원하는 값을 구하는 것입니다.

L(θ|v1,v2,...,vn) 이 경우 가장 적절한 θ는 우도를 최대화해주는 값이 될 것입니다. 다시 말해 관측된 데이터가 발생할 경우를 가장 높게 만들어주는 값이라는 의미입니다.

베르누이 분포

베르누이 시행이란 어떤 실험이 두 가지 결과만을 가지는 실험을 가리킵니다.

베르누이 시행의 결과에 따라 0(실패) 또는 1(성공)의 값을 대응시키는 확률변수(random variable)를 베르누이 확률변수라 합니다. 이 확률변수의 확률분포를 베르누이 분포라고 합니다.

베르누이 확률변수 Y의 분포는 아래 표와 같습니다.

Y	0	1
$P(Y=y_i)$	$1-p$	p

베르누이 분포

수식으로 정리한 베르누이 분포:

$$P(Y = y_i) = p^{y_i}(1-p)^{1-y_i} \quad (y_i = 0, 1)$$

베르누이 확률변수 **Y**에 관한 우도함수**(likelihood function)**는 아래와 같습니다.

$$L = \prod_i p^{y_i}(1-p)^{1-y_i}$$

로지스틱 우도(likelihood)함수

학습데이터에 관측치가 i개가 있고, 정답 범주가 2개(1 또는 0)뿐인 이항로지스틱 모델의 파라메터 β가 주어졌다고 가정해 보겠습니다.

그러면 i번째 관측치의 종속변수 yi는 σ(βTxi)의 확률로 1, 1−σ(βTxi)의 확률로 0이 됩니다. 여기에서 xi는 i번째 관측치의 독립변수, σ는 로지스틱 함수를 가리킵니다. 따라서 로지스틱회귀의 우도함수는 다음과 같이 쓸 수 있습니다.

$$L = \prod_i \sigma(\beta^T \overrightarrow{x_i})^{y_i} \left\{1 - \sigma(\beta^T \overrightarrow{x_i})\right\}^{1-y_i}$$

여기에서 xi는 i번째 관측치의 독립변수, σ는 로지스틱 함수를 가리킵니다.

로지스틱 우도(likelihood)함수

로지스틱 회귀의 파라메터 β는 앞서 언급한 MLE(최대우도측정법)로 구합니다. 로그는 단조

증가함수이므로 로그 우도함수(log-likelihood function)를 최대로 하는 회귀계수 β는 동시에 우도를 최대화하는 β이며 그 역도 성립합니다. 로그 우도함수는 아래와 같이 정리할 수 있습니다.

$$\ln L = \sum_i y_i \ln\left\{\sigma(\beta^T \overrightarrow{x_i})\right\} + \sum_i (1 - y_i) \ln\left\{1 - \sigma(\beta^T \overrightarrow{x_i})\right\}$$

다만 위 로그 우도함수는 추정 대상 파라메터인 회귀계수 β에 대해 비선형이기 때문에 선형회귀와 같이 명시적인 해가 존재하지 않습니다. 따라서 Stochastic Gradient Descent(SGD) 같은 반복적이고 점진적인 방식으로 해를 구하게 됩니다.

그라디언트

Gradient는 공간에 대한 기울기를 말한다.

Gradient 공식:

$$\nabla f = \frac{\partial f}{\partial x}\underline{e}_x + \frac{\partial f}{\partial y}\underline{e}_y$$

여기서 **e(x)**는 **x**방향으로의 단위 벡터이고 **e(y)** **y**방향으로의 단위 벡터 이다.

그라디언트의 방향은 함수값이 커지는 방향이다.

그라디언트 방향의 정반대는 함수의 최솟값으로 가는 뱡향이다.

비용 함수

모델의 정확도를 측정할때 사용되며 예측값과 실제값 차이의 평균을 의미한다.

$$\frac{1}{N}\sum_{i=1}^{n}(\overline{Y_i}-Y_i)^2=\frac{1}{N}\{(\overline{Y_1}-Y_1)^2+(\overline{Y_2}-Y_2)^2+\cdots+(\overline{Y_n}-Y_n)^2\}$$

GD(Gradient Descent)

경사 하강법(GD):
경사/기울기 하강법(Gradient Descent)은 초기값(initial weight)부터 경사를 따라 천천히 내려가서 손실의 최저점을 찾는 방법.

$$\theta_{step+1} = \theta_{step} - pg$$

여기서 P는 학습률 G는 경사값을 의미한다. 경사값은 방향을 나타내고 학습률은 속도를 나타낸다.

수식을 보면 경사값의 반대 방향으로 P만큼 내려간다는 것을 알수있다.

이러한 과정을 손실 함수의 출력값이 많이 줄어들지 않을 때까지 반복한다.
(손실 함수의 출력값이 많이 줄어들지 않았다는 것 -> 경사를 거의 다 내려왔다는 의미)

GD(Gradient Descent)

경사값은 접선의 기울기를 의미하는 것으로 미분을 통해 구한다.

학습률이 크면 매 스텝마다 세타를 많이 변화시켜서 빠른 시간에 학습이 이루어지지만

최솟값 근처에서 완전히 수렴하지 못하고 오른쪽 왼쪽을 왔다 갔다 할 수 있다.

SGD

확률적 경사하강법(SGD)은 다른 GD 방법이 전체 데이터를 대상으로 계산을 한다면

SGD는 일부 데이터만 이용해서 손실 함수와 1차 미분값을 근사적으로 계산한다.

SGD에는 일반적인 SGD, 배치 SGD가 존재한다. -

일반적인 SGD은 한샘플의 그레디언트를 계산하여 속도가 빠르다.하지만 각데이터의 미분값은 전체 데이터를 활용한 미분값 보다 부정확 하므로 정확하지 않다는 단점이 있다.

배치 SGD는 샘플의 그라디언트의 평균을 구한후 한꺼번에 갱신한다. 따라서 정확도는 높지만 속도는 빠르지 않다.

SGD 알고리즘

하나의 샘플에 대한 그라디언트를 계산하고 바로 갱신한다.

3과 6번 사이의 알고리즘을 반복한다.

일반적인 SGD는 한 샘플의 gradient를 계산한 후
즉시 갱신하여 속도가 빠르기는 하지만,
각 데이터의 1차 미분값은 전체 샘플 데이터를
이용한 1차 미분값보다 부정확하므로
손실 함수의 출력값이 정확하지 않다는 단점이 있다.

알고리즘 2-5 스토케스틱 경사 하강 알고리즘(SGD)

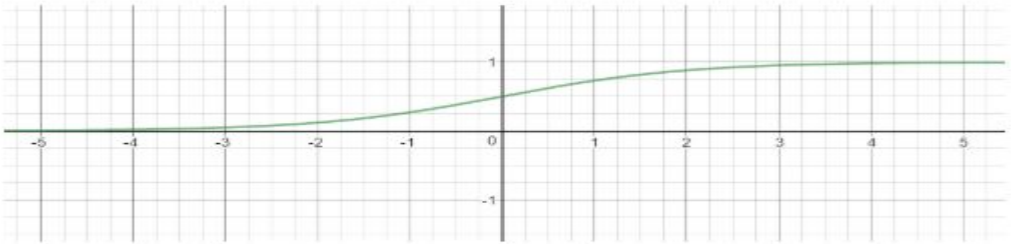
입력 : 훈련집합 $\mathbb{X}$와 $\mathbb{Y}$, 학습률 ρ
출력 : 최적해 $\hat{\Theta}$

1 난수를 생성하여 초기해 Θ를 설정한다.
2 repeat
3 $\mathbb{X}$의 샘플의 순서를 섞는다.
4 for (i=1 to n)
5 i번째 샘플에 대한 그레이디언트 ∇_i를 계산한다.
6 $\Theta = \Theta - \rho\nabla_i$
7 until(멈춤 조건)
8 $\hat{\Theta} = \Theta$

로지스틱 함수

우선 로지스틱 함수(**Logistic Function**)와 승산(**Odds**)에 대해 알아보겠습니다. 로지스틱 회귀의 뼈대가 되는 아이디어이기 때문입니다.
실제 많은 자연, 사회현상에서는 특정 변수에 대한 확률값이 선형이 아닌 **S**-커브 형태를 따르는 경우가 많다고 합니다. 이러한 **S**-커브를 함수로 표현해낸 것이 바로 로지스틱 함수입니다. 분야에 따라 시그모이드 함수로도 불리기도 합니다.
로지스틱 함수는 **x**값으로 어떤 값이든 받을 수가 있지만 출력 결과는 항상 **0**에서 **1**사이 값이 됩니다. 즉 확률밀도함수 **(probability density function)** 요건을 충족시키는 함수라는 이야기입니다. 그 식과 그래프 모양은 아래와 같습니다.

$$y = \frac{1}{1 + e^{-x}}$$



Odds

승산(Odds)이란 임의의 사건 **A**가 발생하지 않을 확률 대비 일어날 확률의 비율을 뜻하는 개념입니다. 아래와 같은 식으로 쓸 수가 있습니다.

$$odds = \frac{P(A)}{P(A^c)} = \frac{P(A)}{1 - P(A)}$$

만약 **P(A)**가 **1**에 가까울 수록 승산은 치솟을 겁니다. 반대로 **P(A)**가 **0**이라면 **0**이 될 겁니다. 바꿔 말하면 승산이 커질수록 사건 **A**가 발생할 확률이 커진다고 이해해도 된다

이항로지스틱 회귀

이제 우리는 범주가 두 개인 분류 문제를 풀어야 합니다.

종속변수 Y가 연속형 숫자가 아닌 범주일 때는 기존 회귀 모델을 적용할 수 없습니다.

회귀식의 장점은 그대로 유지하되 종속변수 Y를 범주가 아니라 (범주1이 될)확률로 두고 식을 세워보면 다음과 같다

$$P(Y=1|X=\vec{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$
$$= \vec{\beta}^T \vec{x}$$

이항로지스틱 회귀

방금의 식을 **odds**를 이용해 바꿔보면 다음과 같다

$$\frac{P(Y=1|X=\vec{x})}{1-P(Y=1|X=\vec{x})} = \vec{\beta}^T\vec{x}$$

범위를 문제 때문에 양변에 로그를 취한다.

$$\log\left(\frac{P(Y=1|X=\vec{x})}{1-P(Y=1|X=\vec{x})}\right) = \vec{\beta}^T\vec{x}$$

이항로지스틱 회귀

앞에 식의 의미는

예를들어 입력벡터 **x**의 첫번째 요소인 **x1**에 대응하는 회귀계수 **β1**이 학습 결과 **2.5**로 정해졌다고 칩시다. 그렇다면 **x1**이 1단위 증가하면 범주 1에 해당하는 로그 승산이 **2.5** 커집니다.

이항로지스틱 회귀

$$\log\left(\frac{P(Y=1|X=\vec{x})}{1-P(Y=1|X=\vec{x})}\right)=\vec{\beta}^{T}\vec{x}$$

위 식을 입력벡터 x가 주어졌을 때 범주1일 확률을 기준으로 정리해주면 다음과 같습니다. (x가 주어졌을 때 범주1일 확률을 p(x), 위 식 우변을 a로 치환해 정리)

$$\frac{p(x)}{1-p(x)}=e^{a}$$

$$\begin{aligned} p(x) &= e^{a}\{1-p(x)\} \\ &= e^{a}-e^{a}p(x) \end{aligned}$$

$$p(x)(1+e^{a})=e^{a}$$

$$p(x)=\frac{e^{a}}{1+e^{a}}=\frac{1}{1+e^{-a}}$$

$$\therefore P(Y=1|X=\vec{x})=\frac{1}{1+e^{-\vec{\beta}^{T}\vec{x}}}$$

이항로지스틱 회귀 결정경계

로지스틱 회귀식에 미지의 정보가 들어왔을때 **범주 1에 속할 확률**을 계산하는것이 필요합니다.

$$P(Y=1|X=\vec{x}) > P(Y=0|X=\vec{x})$$ ← 가정

$$p(x) > 1-p(x)$$

$$\frac{p(x)}{1-p(x)} > 1$$

$$\log\frac{p(x)}{1-p(x)} > 0$$

$$\therefore \vec{\beta}^{T}\vec{x} > 0$$