투빅스 10기 정규과정

ToBig's 9기 김명진

Logistic Regression & Classification

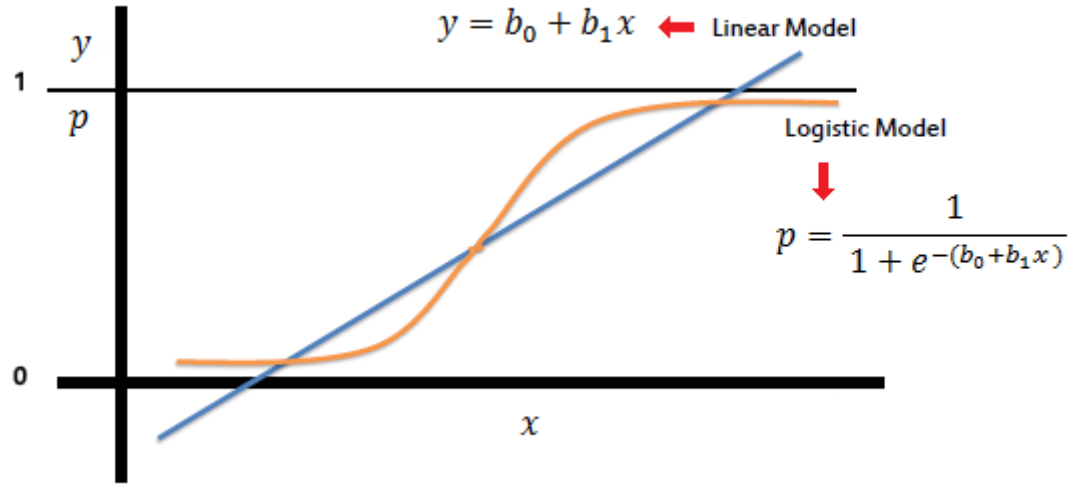
1. 회귀분석

- Output Variable이 연속형
- X가 주어졌을 때 Y의 조건부 평균을 모델링

2. 로지스틱 회귀분석

- Output Variable이 범주형
- X가 주어졌을 때 Y의 조건부 확률을 모델링
- 입력 데이터가 주어졌을 때 결과가 특정 분류로 나뉘기 때문에 분류기법으로 볼 수 있다.

- 왜 조건부 확률로 모델링하지?
- Output이 Binary class(0 or 1)인 경우를 생각해보자!
- Linear Regression을 하는 경우 0보다 작고 1보다 큰 값이 나온다.



잔차 폭발!

- 조건부 확률을 모델링하기 때문에
- 종속 변수의 결과가 [0,1]에 속한다.
- Output이 binary이기 때문에 조건부 확률의 분포가 정규분포 대신 이항 분포를 따른다.

- Odds : y=1확률과 y=0확률의 비

$$Odds = \frac{p(Y=1|X)}{1-p(Y=1|X)}$$

- Logit : Odds에 로그를 취한 것

-Logit을 반응변수로 선형 모형을 만드는 것이 로지스틱 회귀

$$\text{logit}(p) = \log \frac{p}{1-p}$$

- Logistic function

-Sigmoid라고도 불림

-Logit function의 inverse

$$\text{logistic function} = \frac{e^{\beta . X_i}}{1+e^{\beta . X_i}}$$

- 왜 Logit을 반응변수로 두고 선형모델을 만들지?

- 조건부 확률을 반응변수로 둔다면?

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

좌변 : [0,1], 우변 : (-inf,inf)
안 맞는다!

- Odds를 반응변수로 둔다면?

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

좌변 : [0,inf), 우변 : (-inf,inf)
안 맞는다!

- Logit을 반응변수로 둔다면?

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

좌변 : (-inf,inf), 우변 : (-inf,inf)
맞는다!!

로지스틱 함수 유도 [편집]

로지스틱 회귀가 다른 회귀 분석과 구분되는 가장 큰 특징은 결과 값이 0 또는 1이라는 것이다. 따라서 결과 값의 범위가 $[-\infty, +\infty]$인 선형 회귀의 식을 결과 값의 범위가 [0,1]이 되도록 로짓 변환을 수행한다. 로지스틱 함수를 구하는 과정은 아래와 같다.

일단, 오즈비를 종속 변수 값에 상관 없이 결과 값이 항상 [0,1] 사이에 있도록 하기 위해 로짓 변환을 수행한다.

$$\operatorname{logit}(\mathbb{E}[Y_i|x_{1,i},\cdots,x_{m,i}]) = \operatorname{logit}(p_i) = \ln\frac{p_i}{1-p_i}$$

그리고 로지스틱 회귀에서 로짓 변환의 결과는 x에 대한 선형 함수와 동일하므로,

$\operatorname{logit}(p_i) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_m x_{m,i} = \beta \cdot X_i$ 가 되고,

위 두식을 결합하면,

$\ln\frac{p_i}{1-p_i} = \beta \cdot X_i$ 이 된다.

따라서, 우리가 구하고자 하는 특정 독립 변수 x가 주어졌을 때, 종속 변수가 1의 카테고리에 속할 확률은

$p_i = \operatorname{logit}^{-1}(\boldsymbol{\beta}\cdot\mathbf{X}_i) = \frac{1}{1+e^{-\boldsymbol{\beta}\cdot\mathbf{X}_i}}$ 이다.

이를 확률 질량 함수로 표현하면 다음과 같다.

$$\Pr(Y_i = y_i \mid \mathbf{X}_i) = {p_i}^{y_i}(1-p_i)^{1-y_i} = \left(\frac{1}{1+e^{-\boldsymbol{\beta}\cdot\mathbf{X}_i}}\right)^{y_i}\left(1-\frac{1}{1+e^{-\boldsymbol{\beta}\cdot\mathbf{X}_i}}\right)^{1-y_i}$$

- 사전 정보가 없으면 0.5이상이면 1 /미만이면 0으로 분류

- 절단값 c를 직접 정하는 경우

1. 사전 정보 활용

-Y =1 인 자료가 많다면 c를 작게 택한다

2. 손실함수를 고려해서

-Y=1 클래스를 잘못 분류하는 손실이
Y=0을 잘못 분류하는 손실보다 크다면
절단값 c를 작게 택한다

- 결과 해석 방법

-Odds Ratio : x+1에서의 Odds와 x에서의 Odds의 비

즉, x가 한 단위 증가할 때 Odds의 증가율

$$\frac{\mathbb{P}(Y=1|x+1)/\mathbb{P}(Y=1|x)}{\mathbb{P}(Y=0|x+1)/\mathbb{P}(Y=0|x)} = \frac{\mathbb{P}(Y=1|x+1)\mathbb{P}(Y=0|x)}{\mathbb{P}(Y=0|x+1)\mathbb{P}(Y=1|x)} = \exp(\beta_1)$$

2. (20점) 다음은 지방간 자료에 대하여 로지스틱 회귀에서 변수 fatliver (0은 정상, 1은 지방간)를 target으로 하여 단계별회귀로 변수선택을 했을 때 최종모형에 대한 결과이다. 여기서 유의한 변수로 AGE(나이), BMI(비만도), Gender(1: 남성, 2: 여성), GPT(간기능검사 수치)가 선택되었다. 적합된 모형의 식을 쓰고 각 계수에 대하여 해석하라.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	-15.0280	1.9084	62.01	<.0001		0.000
AGE		1	0.0447	0.0150	8.88	0.0029	0.2594	1.046
BMI		1	0.1045	0.0159	43.41	<.0001	0.7929	1.110
GENDER	1	1	1.1650	0.3524	10.93	0.0009		3.206
GENDER	2	0	0	.	.	.	.	.
GPT		1	0.0703	0.0216	10.58	0.0011	0.5992	1.073

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	-15.0280	1.9084	62.01	<.0001		0.000
AGE		1	0.0447	0.0150	8.88	0.0029	0.2594	1.046
BMI		1	0.1045	0.0159	43.41	<.0001	0.7929	1.110
GENDER	1	1	1.1650	0.3524	10.93	0.0009		3.206
GENDER	2	0	0	.	.	.	.	.
GPT		1	0.0703	0.0216	10.58	0.0011	0.5992	1.073

logit(Pr(지방간))=-14.4455+0.0447AGE+0.1045BMI+1.1650I(Gender=1)+0.0703GPT

나이가 1단위 증가하면 지방간이 있을 확률의 오즈비가 1.046배 증가

비만도가 1단위 증가하면 지방간이 있을 확률의 오즈비가 1.110배 증가

성별이 남성인 경우 여성인 경우에 비해 지방간이 있을 확률의 오즈비가 3.206배 증가

간기능검사 수치 GPT가 1단위 증가하면 지방간이 있을 확률의 오즈비가 1.073배 증가

- 클래스가 3개 이상이면? K개라면?? (Multinomial Classification)
- 1/2/3
- 1 or not

$$\log \frac{P(Y=1|X=\overrightarrow{x})}{P(Y=3|X=\overrightarrow{x})} = \beta_1^T \overrightarrow{x}$$

- 2 or not

$$\log \frac{P(Y=2|X=\overrightarrow{x})}{P(Y=3|X=\overrightarrow{x})} = \beta_2^T \overrightarrow{x}$$

-K-1개의 binary logistic classifier로 분류가능!

- 이런 상황을 일반화한 것이 Softmax function!

$$\text{softmax}: \ P(Y = k|x) = \frac{e^{\beta_k X}}{\sum_k e^{\beta_k X}}$$

-Sigmoid처럼 0과 1 사이의 값으로 변환한다
-결과들의 합이 1이 되도록 만들어 준다.

Softmax Classifier (Multinomial Logistic Regression)

$$L_i = -\log\left(\frac{e^{sy_i}}{\sum_j e^{s_j}}\right)$$

unnormalized probabilities

cat	**3.2**	**24.5**	**0.13**
car	5.1	164.0	0.87
frog	-1.7	0.18	0.00

exp → normalize

unnormalized log probabilities → probabilities

First axiom [edit]

The probability of an event is a non-negative real number:

$$P(E) \in \mathbb{R}, P(E) \geq 0 \qquad \forall E \in F$$

where F is the event space. In particular, $P(E)$ is always finite, in contrast with more general measure theory. Theories which assign negative probability relax the first axiom.

Second axiom [edit]

See also: Unitarity (physics)

This is the assumption of unit measure: that the probability that at least one of the elementary events in the entire sample space will occur is 1.

$$P(\Omega) = 1.$$

Third axiom [edit]

This is the assumption of σ-additivity:

Any countable sequence of disjoint sets (synonymous with *mutually exclusive* events) $E_1, E_2, \ldots$ satisfies

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i).$$

- 그럼 Softmax를 왜 쓰는거지??

- Loss Function : 학습의 기준!
- Cross Entropy : 분류에서 주로 사용하는 Loss Function
  -예측 확률분포와 정답 확률분포의 차이를 나타냄!

$$CE = -\sum_{x} p(x) \log q(x)$$

p(x) : one hot encoded true label / q(x) : predicted probability by model

Softmax Classifier (Multinomial Logistic Regression)

$$L_i = -\log\left(\frac{e^{sy_i}}{\sum_j e^{s_j}}\right)$$

unnormalized probabilities

	unnormalized log probabilities	exp		normalize	probabilities	
cat	**3.2**	→	**24.5**	→	**0.13**	→ L_i = -log(0.13) = **0.89**
car	5.1	→	164.0	→	0.87	
frog	-1.7	→	0.18	→	0.00	

Regression	Binomial	Multinomial (n>2)
$\hat{y}$	$\hat{y} = \frac{1}{1+e^{-\theta X}}$	$\hat{y} = \frac{e^{\theta_y X}}{\sum e^{\theta X}}$
cross entropy	$J(\theta) = -\frac{1}{n}\sum_{i}^{n}[y_i log(\hat{y}_i) + (1-y_i) log(1-\hat{y}_i)]$	$J(\theta) = -\sum_{i} y_i ln(\hat{y}_i)$
forward pass	$[NX1] = [NXD][DX1]$	$[NXC] = [NXD][DXC]$
response	$y = 0$ or 1	$y =$ one-hot-encoded

- Confusion Matrix

		Actual class	
		Cat	**Non-cat**
Predicted class	**Cat**	5 True Positives	2 False Positives
	Non-cat	3 False Negatives	17 True Negatives

accuracy (ACC)

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{P + N} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

sensitivity, recall, hit rate, or true positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

specificity or true negative rate (TNR)

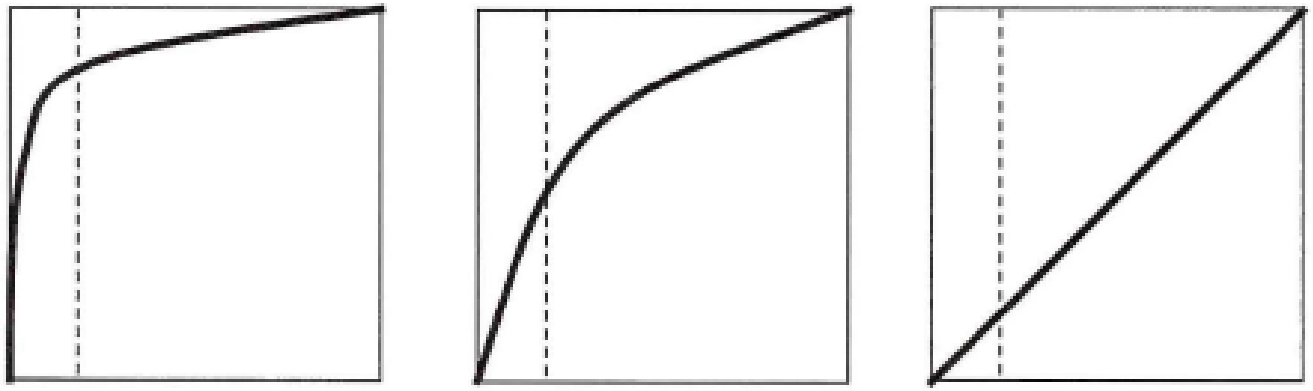
$$\mathrm{TNR} = \frac{\mathrm{TN}}{N} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

-민감도 : 실제 참인 것 중 참으로 판정한 비율

-특이도 : 실제 거짓인 것 중 거짓으로 판정한 비율

- ROC : 여러 절단값에서의 민감도와 특이도의 관계를 보여주는 그래프

- 절단값이 변하면서 분류기의 성능이 어떻게 변해가는지 눈으로 확인할 수 있다.
- x축 : 1-특이도 , y축 : 민감도로 하여 모든 가능한 절단값 c에 대해 연결!



- 곡선 아래 면적을 AUC라 하고 Classifier의 성능 평가 지표로 활용

- 과제

- 1. mlbench 패키지의 BreastCancer data를 7:3으로 train set/test set으로 나눈 후 로지스틱 회귀를 적합하여 예측하세요. 정확한 결과를 위해 랜덤분할을 50회 실시하여 Accuracy의 평균을 구해주세요.

- 2. psub.Rdata 데이터를 로드해 7:3으로 train set/test set으로 나누세요. AGEP, SEX, COW, PINCP, SCHL 변수만을 이용하여 학사학위 이상 소지자와 그렇지 않은 사람의 두 범주로 bachdeg변수를 생성해 SCHL을 제외한 나머지 변수들에 대해 로지스틱 회귀 모형을 적합하고 예측하여 Confusion Matrix를 만들어 주세요.

  (COW: class of worker, SCHL: level of education, PINCP: personal income, AGEP : age)

들어주셔서 감사합니다.