4강. Basic Methods for Classification

◈ 담당교수 : 김 동 하

■ 학습개요

이번 강의에서는 범주형 반응변수를 예측하는 분류문제에 대해서 다룬다. 먼저, 분류문제의 정의에 대해서 배우고, 이를 해결하는 모형 중 가장 간단한 형태인 로지스틱 모형에 대해 학습한다. 또한, 로지스틱 모형을 Python 프로그램에서 구현하는 방법을 알아본다.

■ 학습목표

1	분류 문제의 개념에 대해 학습한다.
2	랜덤성분, 체계적성분, 연결함수에 대해 학습한다.
3	로지스틱 모형에 대해 학습한다.

■ 주요용어

용어	해설
분류 문제	독립변수를 이용하여 범주형 종속변수를 예측하는 문제를 뜻한다.
랜덤성분	일반화선형모형을 구성하는 요소 중 하나로, 독립변수가 주어졌을 때 종속변수의 확률분포를 규정한다.
체계적성분	일반화선형모형을 구성하는 요소 중 하나로, 모형에서 종속변수의 기댓값을 설명하기 위해 독립변수를 어떻게 사용할 것인지를 규정한다.
연결함수	일반화선형모형을 구성하는 요소 중 하나로, 체계적 성분과 종속변수의 기댓값과의 관계를 규정한다.
로지스틱 모형	독립 젼수로 이진 종속변수를 예측하는 가장 대표적인 분류 모형으로 랜덤성분은 베르누이 분포, 체계적성분은 선형함수, 연결함수는 로짓함수를 사용한 모형이다.

■ 학습하기

01. 분류 문제

회귀 문제와 분류 문제

- 독립 변수를 이용하여 종속 변수를 예측하는 것은 동일
- 종속 변수의 형태는 크게 두 가지가 존재
- 수치형 변수 : 회귀 모형
- 범주형 변수 : 분류 모형

분류 문제의 예

- 종속 변수는 범주형 변수
  -> 범주형 종속 변수값을 class 혹은 label이라고도 함.
- 분류 문제의 예시
  -> 제품이 불량인지 양품인지를 분류
  -> 고객이 이탈고객인지 잔류고객인지를 분류
  -> 환자의 특정 병에 대해 양성인지 음성인지를 분류
- 본 강의에서는 이진 분류만을 다룰 예정

다양한 분류 모형

- 분류 문제를 해결하기 위해서 매우 다양한 모형들이 존재
- 설명 변수들의 선형식으로 종속 변수를 예측
  -> 로지스틱 모형
- 설명 변수들의 비선형식으로 종속 변수를 예측
  -> 의사결정나무
  -> Support vector machine
  -> 앙상블 기법 (부스팅, random forest 등)

02. 로지스틱 모형

선형 회귀 모형 복습

- 다중선형회귀모형

$$Y=\beta_0+\beta_1X_1+\cdots+\beta_pX_p+\epsilon,\ \epsilon\sim N(0,\sigma^2)$$

- 선형회귀모형의 경우 종속 변수가 연속형 변수일 때만 사용 가능.
- 종속 변수가 범주형이라 가정하자. (0 또는 1)
- 선형회귀모형을 가정할 경우 문제점이 발생함.
- $E(Y\mid X=x)$의 범위가 [0,1]을 벗어날 수 있음.
- 오차항 ϵ의 분포가 정규분포가 될 수 없음.
- 종속 변수 Y 가 수치형이 아닐 때에도 Y와 $\beta_0+\beta_1X_1+\cdots+\beta_pX_p$의 관계를 모형화할 수는 없을지?

선형 회귀 모형의 세 가지 요소

- 랜덤 성분 (Random component)
- 체계적 성분 (Systematic component)
- 연결함수 (Link function)

랜덤 성분

- 독립변수가 주어졌을 때 종속변수의 확률분포를 규정.
- 선형모형에서의 랜덤 성분은 정규분포

$$Y \mid X = x \sim N(\beta_0 + \beta_1 x_1 + \cdots \beta_p x_p, \sigma^2)$$

체계적 성분

- 모형에서 종속변수의 기댓값을 설명하기 위해 독립변수를 어떻게 사용할 것인지를 규정.
- 선형모형에서의 체계적 성분은 선형 함수.

$$E(Y \mid X = x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

연결 함수

- 체계적 성분과 종속변수의 기댓값과의 관계를 규정.
- 선형모형에서의 연결 함수는 항등 함수.

$$g(E(Y \mid X = x)) = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p$$

$g(a) = a$: 항등 함수 (identity function)

로지스틱 모형

- 이진분류만을 고려하기 때문에 종속변수 Y는 0과 1을 갖는다고 가정.
- 앞서 언급한 세 가지 성분 중에서 랜덤 성분과 연결 함수를 변형.

로지스틱 모형의 랜덤 성분

- 베르누이 분포를 사용.

$$Y \mid X = x \sim Ber(\pi)$$
$$P(Y = 1 \mid X = x) = \pi$$
$$P(Y = 0 \mid X = x) = 1 - \pi$$

로지스틱 모형의 체계적 성분

- 선형 회귀 모형과 동일하게 선형 함수를 사용.

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

로지스틱 모형의 연결 함수

- 로짓 함수 (logit function)을 사용.

$$g(E(Y \mid X=x)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$
$$g(a) = \log\left(\frac{a}{1-a}\right)$$

- 즉,

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p,$$
$$\pi = P(Y=1 \mid X=x) = E(Y \mid X=x)$$

- 로지스틱 모형을 다음과 같은 형태로도 수식화할 수 있다.

$$P(Y=1 \mid X=x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}$$
$$P(Y=0 \mid X=x) = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}$$

단순 로지스틱 모형과 다중 로지스틱 모형

- 단순 로지스틱 모형

$$P(Y=1 \mid X=x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

모수: β_0, β_1

- 다중 로지스틱 모형

$$P(Y=1 \mid X=x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}$$

모수: $\beta_0, \beta_1, \cdots, \beta_p$

로지스틱 모형의 모수 추정

- 학습 데이터: $(x_1, y_1), \cdots, (x_n, y_n)$
- 학습 데이터를 이용한 교차 엔트로피를 사용

$$-\sum_{i=1}^{n} \log(P(Y=y_i \mid X=x_i))$$

- 교차 엔트로피를 최소로 하는 모수를 추정한다.

로지스틱 모형을 이용하여 예측하기

- 추정한 모수: $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$
- 새로운 입력 변수 x가 주어졌을 때 출력 변수를 다음과 같이 예측한다.

$$\hat{p} = \hat{P}(Y=1 \mid X=x) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p)}$$

- $\hat{y} = 1$ if $\hat{p} \geq c$
- $\hat{y} = 0$ if $\hat{p} < c$

- $c \in (0,1)$는 절단값이라 부르며, 대게 0.5를 사용.

로지스틱 모형을 이용한 예측 예제

- 예: 모의고사 점수(X)를 이용해 A 대학의 합격 여부(Y) 예측하기 (단순 로지스틱모형)

$$\hat{P}(Y=1 \mid X=x) = \frac{\exp(\hat{\beta}_0+\hat{\beta}_1 x_1)}{1+\exp(\hat{\beta}_0+\hat{\beta}_1 x_1)}$$
$$= \frac{\exp(-50+0.5x)}{1+\exp(-50+0.5x)}$$

- 절단값을 0.5로 사용하자.
- 모의고사 점수가 105점일 때

$$\frac{\exp(-50+0.5*105)}{1+\exp(-50+0.5*105)} = 0.924 > 0.5$$

- 합격으로 예측.

03. Python을 이용한 실습

타이타닉 데이터셋

- 타이타닉 사건 때 배에 있었던 승객들의 명단
- 891명의 12가지 정보를 포함하고 있음.
- Survived: 사망 여부 (0: 사망, 1: 생존)
- Pclass: 1=1등석, 2=2등석, 3=3등석
- Sex: male=남성, female=여성
- 등등
- 성별, 나이, 좌석 등급으로 승객의 사망 여부를 예측하는 로지스틱 모형을 적합하자.

패키지 불러오기

```
import os
import numpy as np
import pandas as pd
import requests
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix
```

데이터셋 불러오기

- read_csv 명령어를 사용하여 titanic.csv 불러오기.

```
data_file = "./data/titanic.txt"
titanic = pd.read_csv(data_file)
print(titanic.shape)
titanic.head()
```

```
(891, 12)
```

	PassengerId	Survived	Pclass	Name	Sex	Age
0	1	0	3	Braund, Mr. Owen Harris	male	22.0
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0
2	3	1	3	Heikkinen, Miss. Laina	female	26.0
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0
4	5	0	3	Allen, Mr. William Henry	male	35.0

데이터 전처리하기

- 종속 변수와 독립 변수로 구분
- 학습 데이터와 평가 데이터로 분리

```
x_titanic = titanic[['Sex', 'Age', 'FirstClass', 'SecondClass']]
y_titanic = titanic['Survived']
```

```
train_x_titanic, test_x_titanic, train_y_titanic, test_y_titanic = \
train_test_split(x_titanic, y_titanic, test_size=0.3, random_state=123)
print(train_x_titanic.head())
print(test_x_titanic.head())
```

로지스틱 모형 적합하기.

```
logistic = LogisticRegression(penalty='none')
logistic.fit(train_x_titanic, train_y_titanic)
```

```
LogisticRegression(penalty='none')
```

```
print(logistic.intercept_)
print(logistic.coef_)
```

```
[-1.08777172]
[[ 2.55838407 -0.04004487  2.38223532  1.05157353]]
```

예측 정확도 살펴보기

```
print(logistic.score(train_x_titanic, train_y_titanic))
print(logistic.score(test_x_titanic, test_y_titanic))
```

```
0.7865168539325843
0.7835820895522388
```

■ 연습문제

(객관식)1. 다음 보기 중 분류 문제에 해당하지 않는 것을 고르시오.

① 이메일이 스팸인지 아닌지 예측.

② 환자의 악성 종양 진행 단계를 예측.

③ 특정 특허에 대해서 앞으로의 소송 여부를 예측.

④ 자동차의 연비 (1리터당 운행거리)를 예측.

정답 : ④
해설 : 연비는 수치형 값이므로 회귀 문제에 해당한다.

(객관식)2. 다음 보기 중 로지스틱모형의 설명으로 잘못된 것을 고르시오.
① 독립변수의 선형식으로 이진분류를 할 수 있는 모형이다.
② 랜덤 성분으로 포아송 분포를 사용한다.
③ 선형회귀모형과 마찬가지로 선형 함수를 체계적 성분으로 사용한다.
④ 로지스틱 모형을 이용하여 예측할 때 보통 절단값을 0.5로 사용한다.

정답) ②
해설) 랜덤 성분으로 베르누이 분포를 사용한다.

(단답형)3. 로지스틱 모형은 종속변수의 기댓값과 체계적 성분 사이의 관계를 나타내기 위해서 어떤 함수를 사용하는가?

정답 : Logit function
해설 : 로지스틱 모형은 연결 함수로 logit function을 사용한다.

■ 정리하기

1. 분류문제는 독립 변수로 범주형인 종속 변수를 예측하기 위한 문제를 뜻한다.

2. 선형회귀모형은 다음의 세 개의 성분으로 구성되어 있다: 1) 랜덤 성분, 2) 체계적 성분, 3) 연결 함수

3. 로지스틱모형은 선형회귀모형에서 랜덤 성분을 베르누이 분포, 연결 함수를 로짓 함수로 변형한 것이다.

■ 참고자료 (참고도서, 참고논문, 참고사이트 등)

없음.