머신러닝응용 제04강

Basic Methods for Classification

첨단공학부 김동하교수

제04강 Basic Methods for Classification

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

핵심 단어

- 분류 문제
- 랜덤성분, 체계적성분, 연결함수
- 로지스틱 모형

04강. Basic Methods for Classification

01. 분류 문제

1) 회귀문제 vs 분류문제

- 독립변수를 이용하여 종속 변수를 예측하는 것은 동일.

- 종속 변수의 형태는 크게 두 가지가 존재
  - 수치형 변수 -> 회귀 모형
    - 선형 회귀 모형
  - 범주형 변수 -> 분류 모형
    - 로지스틱 모형

2) 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

3) 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)

04강. Basic Methods for Classification

02. 로지스틱 모형

1) 선형 회귀 모형

- **다중선형회귀모형**

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$

- **선형회귀모형의 경우 종속 변수가 연속형 변수일 때만 사용 가능.**

1) 선형 회귀 모형

- 종속 변수가 범주형이라 가정하자. (0 또는 1)

- 선형회귀모형을 가정할 경우 문제점 발생
  - $E(Y|X = x)$의 범위가 [0,1]을 벗어날 수 있음.
  - 오차항 ϵ의 분포가 정규분포가 될 수 없음.

- 종속 변수 Y 가 수치형이 아닐 때에도 Y와 $\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$의 관계를 모형화할 수는 없을까?

1) 선형 회귀 모형

다중선형회귀모형

- $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$

선형회귀모형을 세 가지 요소로 나누어볼 수 있다.

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

1) 선형 회귀 모형

랜덤 성분 (Random component)

- 독립변수가 주어졌을 때 종속변수의 확률분포를 규정.

선형모형에서의 랜덤 성분은 정규분포

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

1) 선형 회귀 모형

체계적 성분 (Systematic component)

- 모형에서 종속변수의 기댓값을 설명하기 위해 독립변수를 어떻게 사용할 것인지를 규정.

선형모형에서의 체계적 성분은 선형 함수.

- $E(Y|X = x) = \boxed{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}$

1) 선형 회귀 모형

연결 함수 (Link function)

- 체계적 성분과 종속변수의 기댓값과의 관계를 규정.

선형모형에서의 연결 함수는 항등 함수.

- $g\big(E(Y|X=x)\big) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$
- $g(a) = a$: 항등 함수 (Identify function)

2) 로지스틱 모형

- 이진분류만을 고려하기 때문에 종속변수 Y는 0과 1을 갖는다고 가정.

- 앞서 언급한 세 가지 성분 중에서 랜덤 성분과 연결 함수를 변형.

2) 로지스틱 모형

랜덤 성분: 베르누이 분포

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

2) 로지스틱 모형

- **체계적 성분: 선형 함수**
  - **선형회귀모형과 동일**
  - $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$

2) 로지스틱 모형

연결 함수: 로짓 함수 (Logit function)

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

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

  - $\pi = P(Y=1|X=x) = E(Y|X=x)$

2) 로지스틱 모형

로지스틱 모형을 다음과 같이도 쓸 수 있음.

- $P(Y=1|X=x)=\frac{\exp(\beta_0+\beta_1x_1+\cdots+\beta_px_p)}{1+\exp(\beta_0+\beta_1x_1+\cdots+\beta_px_p)}$
- $P(Y=0|X=x)=\frac{1}{1+\exp(\beta_0+\beta_1x_1+\cdots+\beta_px_p)}$

2) 로지스틱 모형

단순 로지스틱모형

- $P(Y=1|X=x)=\frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}$
- **모수**: β_0, β_1

다중 로지스틱모형

- $P(Y=1|X=x)=\frac{\exp(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p)}{1+\exp(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p)}$
- **모수**:$\beta_0, \beta_1, \cdots, \beta_p$

3) 로지스틱 모형의 추정

로지스틱 모형의 추정

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

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

- 교차 엔트로피를 최소로 하는 모수를 모수 추정값으로 사용.

4) 예측하기

로지스틱 모형을 이용하여 예측하기

$$\hat{P}(Y=1|X=x)=\frac{\exp(\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_px_p)}{1+\exp(\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_px_p)}$$

- **새로운 입력 변수 x 에 대한 출력 변수를 다음과 같이 예측한다.**
  - $\hat{y}=1,\ if\ \hat{P}(Y=1|X=x)\geq c$
  - $\hat{y}=0,\ if\ \hat{P}(Y=1|X=x)<c$
- $c\in(0,1)$**는 절단값이라 부르며, 대게 0.5를 사용.**

4) 예측하기

로지스틱 모형을 이용하여 예측하기

- 예측 과정을 다음과 같이 표시할 수도 있다.
  - $\hat{y} = 1,\ if\ \ \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p \geq \log\left(\frac{c}{1-c}\right)$
  - $\hat{y} = 0,\ if\ \ \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p < \log\left(\frac{c}{1-c}\right)$
- 절단값을 0을 사용할 경우 $\log\left(\frac{c}{1-c}\right) = 0$
- 즉, $\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$가 0 이상이면 1로, 음수면 0으로 예측.

4) 예측하기

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

$$\hat{P}(Y=1|X=x)=\frac{\exp(\hat{\beta}_0+\hat{\beta}_1 x)}{1+\exp(\hat{\beta}_0+\hat{\beta}_1 x)}$$

$$=\frac{\exp(-50+0.5x)}{1+\exp(-50+0.5x)}$$

  - **절단값을 0.5로 사용하자.**

4) 예측하기

- 로지스틱 모형을 이용하여 예측하기
  - 모의고사 점수가 105점일 때
    - $\frac{\exp(-50+0.5*105)}{1+\exp(-50+0.5*105)} = 0.924 > 0.5$
    - 합격으로 예측
  - 모의고사 점수가 95점일 때
    - $\frac{\exp(-50+0.5*95)}{1+\exp(-50+0.5*95)} = 0.076 < 0.5$
    - 불합격으로 예측

04강. Basic Methods for Classification

03. Python을 이용한 실습

1) 데이터 설명

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.

1) 데이터 설명

- 성별, 나이, 좌석 등급으로 승객의 사망 여부를 예측하는 로지스틱 모형을 적합하자.

2) 환경설정

- 로지스틱 모형을 구현하기 위해 필요한 패키지를 불러오자.

```
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
```

3) 데이터 불러오기

◆read_csv를 이용해 titanic.csv 파일을 불러오자.

```
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

4) 데이터 전처리

- Sex: 남성은 0, 여성은 1로 변환
- Age: 결측치는 평균값으로 대체
- Pclass: 가변수 생성
  - 1,2,3 등급이므로 2개의 가변수면 충분.
  - 3등급을 기준 범주로 하였음.

4) 데이터 전처리

```
titanic['Sex'] = titanic['Sex'].map({'female':1,'male':0})
titanic['Age'].fillna(value=titanic['Age'].mean(), inplace=True)
titanic['FirstClass'] = titanic['Pclass'].apply(lambda x: 1 if x == 1 else 0)
titanic['SecondClass'] = titanic['Pclass'].apply(lambda x: 1 if x == 2 else 0)
```

Sex	FirstClass	SecondClass
0	0	0
1	1	0
1	0	0
1	1	0
0	0	0

4) 데이터 전처리

- 종속변수와 독립변수 구분
- 학습데이터와 평가데이터 나누기

```
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())
```

4) 데이터 전처리

- 종속변수와 독립변수 구분
- 학습데이터와 평가데이터 나누기

	Sex	Age	FirstClass	SecondClass
416	1	34.0	0	1
801	1	31.0	0	1
512	0	36.0	1	0
455	0	29.0	0	0
757	0	18.0	0	1

	Sex	Age	FirstClass	SecondClass
172	1	1.000000	0	0
524	0	29.699118	0	0
452	0	30.000000	1	0
170	0	61.000000	1	0
620	0	27.000000	0	0

5) 모형 적합하기

◆ 로지스틱모형 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]]
```

6) 적합된 모형 평가하기

◆ 예측값 살펴보기.

```
print(test_x_titanic.head())
print(logistic.predict(test_x_titanic)[0:5])
```

```
      Sex        Age  FirstClass  SecondClass
172     1   1.000000           0            0
524     0  29.699118           0            0
452     0  30.000000           1            0
170     0  61.000000           1            0
620     0  27.000000           0            0
[1 0 1 0 0]
```

6) 적합된 모형 평가하기

◆ 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
```

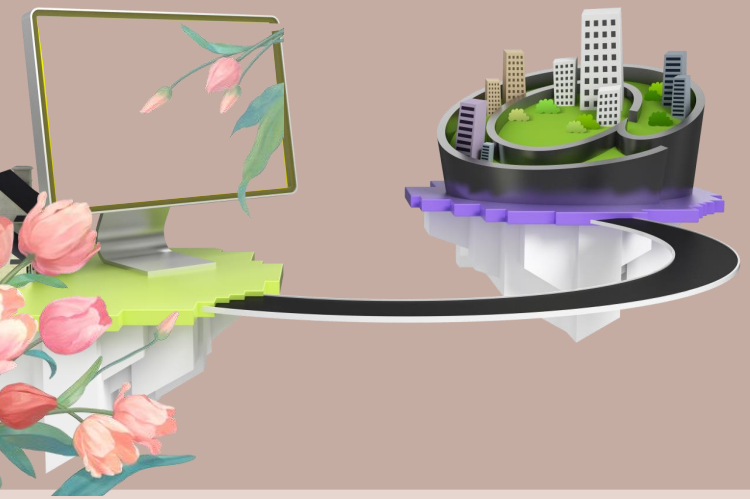
6) 적합된 모형 평가하기

◆ 오차행렬 구하기 (Confusion matrix)

```
test_y_prediction = logistic.predict(test_x_titanic)
print(test_y_titanic.sum())
print(test_y_prediction.sum())
confusion_matrix(test_y_titanic, test_y_prediction)
```

```
98
108

array([[136,  34],
       [ 24,  74]])
```



다음시간안내

제05강

Discriminant Analysis