와인 특징 분석

HYOJUN ACADEMY WINE CLASS

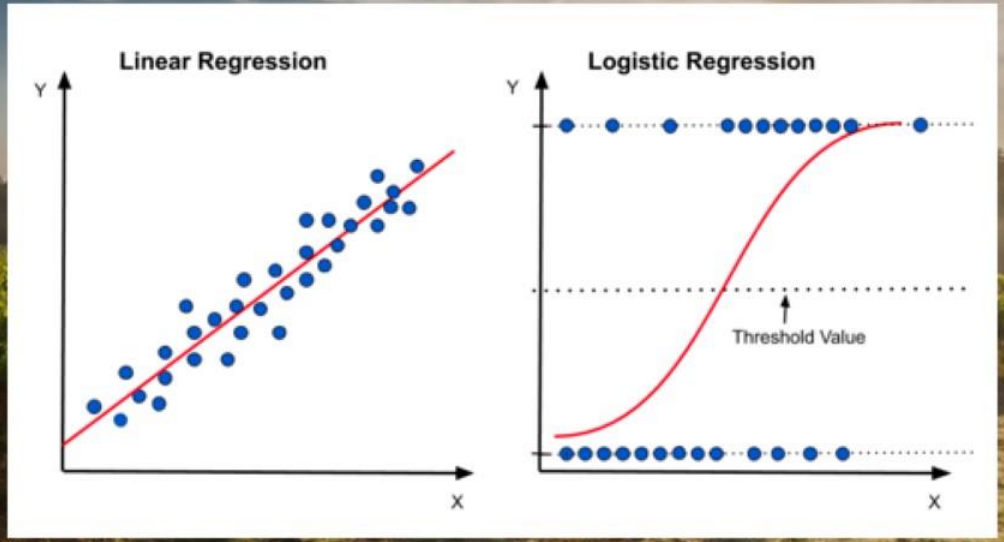
효준 소믈리에와
함께하는 와인의 기초

일 시 2024.08.23 장 소 경북대학교 복현회관

와인 색깔 예측 모형 구축

04 part

안효준



로지스틱 회귀분석이란

- 독립 변수의 선형 결합을 이용하여 사건의 발생 가능성을 확률로 예측하는 데 사용되는 통계 기법
- 선형 회귀 : (종속 변수 : 수치형)
- 로지스틱 회귀 : (종속 변수 : 범주형)

```python
X = wine.iloc[:, :-1] # 독립변수
y = wine['color'] # 종속변수

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

✓ 0.0s Python

훈련(train), 테스트(test) 세트 분할 (80 : 20)

```python
# 데이터 스케일링 (수치형 변수)
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X_train)
```

✓ 0.0s Python

설명변수의 단위를 맞추기 위해 스케일링(Scaling)

로지스틱 회귀모델 훈련 결과

Confusion Matrix:

```
[[3899   13]
 [  19 1266]]
```

Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	3912
1	0.99	0.99	0.99	1285
accuracy			0.99	5197
macro avg	0.99	0.99	0.99	5197
weighted avg	0.99	0.99	0.99	5197

Model Coefficients: <- 회귀계수(베타) 값

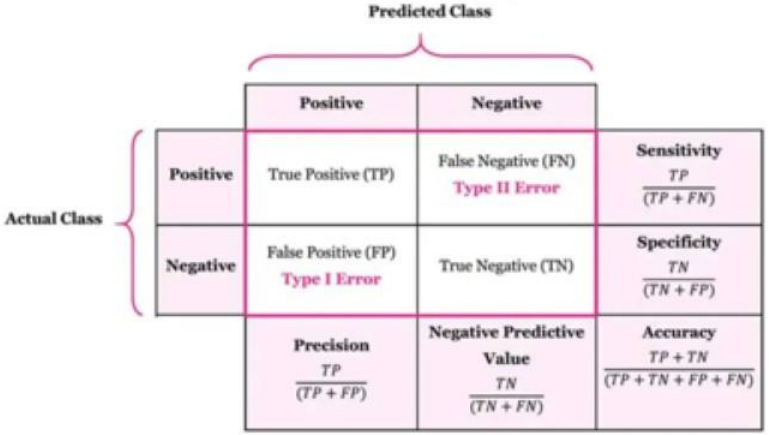
```
[[ 0.36907798  1.29401329 -0.36287684 -3.57862293  0.79205655  0.93706383
  -2.95310731  3.5087579   0.33391421  0.61707462  1.22368381  0.20372537]]
```

Intercept:

```
[-4.19044926]
```

<- 상수항(알파) 값

일반식: $\text{logit}(P(Y=1)) = \alpha + \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{12} x_{12}$



혼동행렬 (Confusion Matrix)

테스트 데이터에 적합한 결과

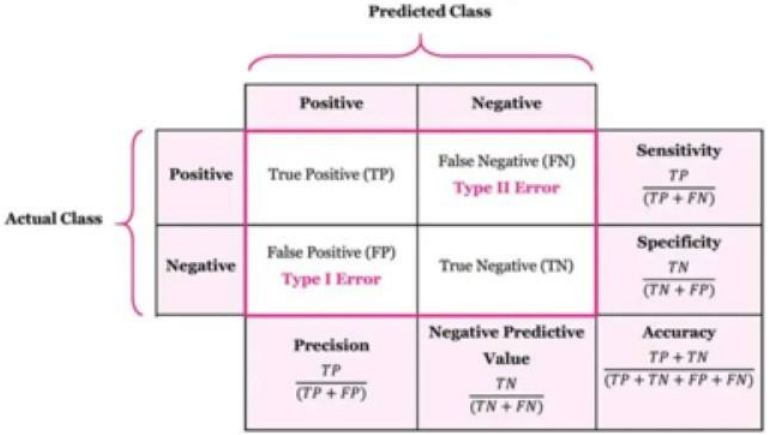
Confusion Matrix:

```
[[981   5]
 [  8 306]]
```

일반식: $\text{logit}(P(Y=1)) = \alpha + \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{12} x_{12}$

Classification Report:

	precision	recall	f1-score	support
0	0.99	0.99	0.99	986
1	0.98	0.97	0.98	314
accuracy			0.99	1300
macro avg	0.99	0.98	0.99	1300
weighted avg	0.99	0.99	0.99	1300



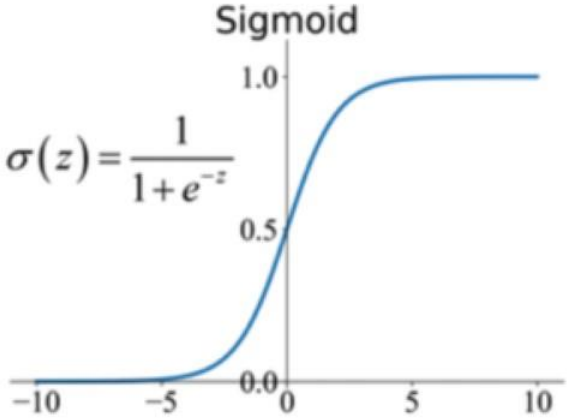
혼동행렬 (Confusion Matrix)

로지스틱 회귀모형 : $\text{logit}(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$

설명변수(x)에 데이터 값을 대입 : $\text{logit}(p) = \beta$

로짓 함수 정의 : $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$ 여기서 $\frac{p}{1-p}$는 확률 p의 오즈

$$p = \frac{e^{\beta}}{1+e^{\beta}}$$



<- 시그모이드 함수

```
print(f"표준화 한 설명변수 값: (고정산도, 휘발성 산도, ..., 알코올, 품질)"
      f"\n{X_test_scaled[0]}\n와인의 종류: {y_test.to_list()[0]} (0: 화이트, 1: 레드)")
```

✓ 0.0s Python

```
표준화 한 설명변수 값: (고정산도, 휘발성 산도, ..., 알코올, 품질)
[-0.1721767  -0.54303355  0.90013042 -0.65328671 -0.32082382  0.53259574
  0.04149221 -1.34998708 -0.3632329  -0.35397899  1.17959322  1.34345237]
와인의 종류: 0 (0: 화이트, 1: 레드)
```

<- 화이트 와인 데이터

```
logit = model.decision_function(X_test_scaled[0].reshape(1, -1))
logit
```

✓ 0.0s Python

```
array([-6.182377])
```

로짓값

$$\text{logit}(p) = -6.182377$$

$$\text{logit}(p) = \beta$$

```
probability = 1 / (1 + np.exp(-logit[0]))
probability
```

✓ 0.0s Python

```
0.0020612547458882333
```

$$p = \frac{e^{-6.182377}}{1+e^{-6.182377}} \approx 0.002$$

따라서 레드 와인(Y=1)일 확률은 약 0.2%

해당 데이터는 화이트 와인으로 추측

```
print(f"표준화 한 설명변수 값: (고정산도, 휘발성 산도, ..., 알코올, 품질)"
      f"\n{X_test_scaled[0]}\n와인의 종류: {y_test.to_list()[0]} (0: 화이트, 1: 레드)")
```

✓ 0.0s Python

```
표준화 한 설명변수 값: (고정산도, 휘발성 산도, ..., 알코올, 품질)
[-0.1721767  -0.54303355  0.90013042 -0.65328671 -0.32082382  0.53259574
  0.04149221 -1.34998708 -0.3632329  -0.35397899  1.17959322  1.34345237]
와인의 종류: 0 (0: 화이트, 1: 레드)
```

<- 화이트 와인 데이터

실제로도 해당 데이터는 화이트 와인 데이터

로지스틱 회귀모형이 잘 작동하는 것으로 보임

아쉬운 점

- **교차 검정**(Cross Validation)을 수행하지 않아 모델의 **과적합** 우려가 존재
- 로지스틱 회귀모형 이외의 **다른 분류 모형**을 비교해보지 못함
  (Random Forest, Gradient Boosting, Light GBM 등)

종합 결론

- 와인 간의 **특징 차이**를 실제로 **통계적**으로 **검정**해 볼 수 있었다.
- 간단한 **분류 모형**을 만들어 봄으로써 머신러닝의 기초를 맛볼 수 있었다.