Z 값을 활용한 결측치 대체에 관한 연구

박승현
대전탄방중학교
realtonypark@gmail.com

A Study on Replacement of Missing Data using Z

Seung-Hyeon Park
Daejeon Tanbang Middle School

요 약

데이터에 결측치가 존재할 때 어떤 데이터로 결측치를 대체시켜야 원래의 데이터에 가장 근접한 데이터를 만들어낼 수 있는지에 관한 연구. Z 값을 사용하면 평균으로 결측치를 대체시키는 것보다 더 정확한 결과를 도출해낼 수 있다.

1. 서론

KNN 알고리즘을 사용해서 붓꽃(아이리스) 품종을 구별하는 실험을 하던 중 결측치가 있으면 KNN 알고리즘을 구현시켰을 때 정확도가 현저히 낮게 나오는 것을 알 수 있었다. KNN 알고리즘으로 구현했을 때 n 값을 5 로 설정했을 때 정확도가 약 90% 로 나오면서 다른 실험들보다 현저히 낮은 정확도를 띄었다. 데이터에 결측치가 있을 때 머신러닝을 구현하더라도 결측치를 원래의 데이터에 가장 근접한 값으로 대체해서 결측치를 보완하는 방법을 연구했다. 표준점수(Z 값)를 사용하면 된다. 표준점수로 결측치를 대체한 것과 평균으로 결측치를 대체한 것을 비교, 분석해서 Z 값의 효과를 보일 수 있다.

2. 본론

실험에 앞서 붓꽃(아이리스) 데이터에 결측치를 생성시켜야 했다. 결측치는 4 개의 열에 있는 데이터들 중 한 행에 최대 한 개씩의 데이터를 결측치로 전환시켰다.

결측치를 표준점수, 즉 Z 값으로 대체시키는 방법으로 연구를 진행했다. 표준점수는 원수치인 x 가 평균에서 얼마나 떨어져 있는지를 나타낸다. 음수이면 평균 이하, 양수이면 평균이상이다. 표준점수 계산은 다음 그림과 같다.

› 표준점수 산출 공식

표준점수 = (Z점수 × 해당 영역의 표준편차) + 평균

$$Z점수 = \frac{X-m}{\sigma}$$

X : 수험생의 원점수
m : 해당 과목의 수험생 평균
σ : 해당 과목의 수험생 표준편차

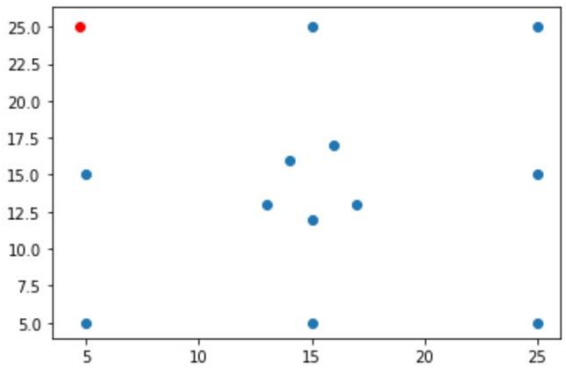
› 각 영역별 문항수 및 원점수·표준점수 범위

영역	문항수	원점수 만점	표준점수		
			평균	표준편차	범위
언어	50	100	100	20	0~200
수리	30	100	100	20	0~200
외국어	50	100	100	20	0~200
사회/과학/직업탐구	20	50	50	10	0~100
제2외국어/한문	30	50	50	10	0~100

(그림 1) 표준점수 계산

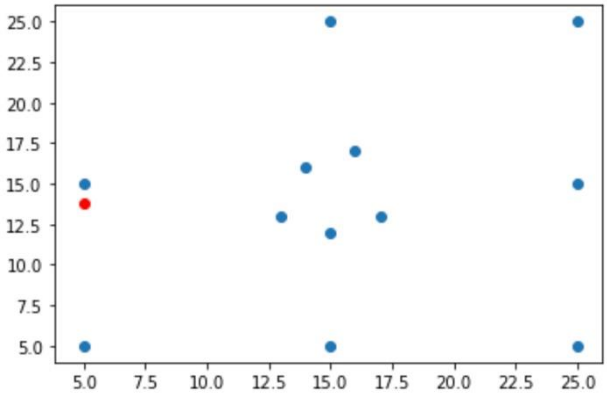
따라서 표준점수를 계산해서 결측치에 적용하기 위해서는 평균, 분산, 표준편차를 계산해야 한다. 위 값들을 계산해서 결측 데이터들을 적용하면 Z 값으로 결측치를 대체시키는 것이 끝난다.

평균으로 결측 데이터를 대체시키는 것과 Z 값으로 결측치를 대체시키는 것을 시각화해서 비교해 보았다. 다음 그림 자료는 직접 만든 13 개의 데이터들 중 하나의 데이터의 y 축 좌표를 결측치로 만들어 빨간색으로 표시해놓은 것이다.



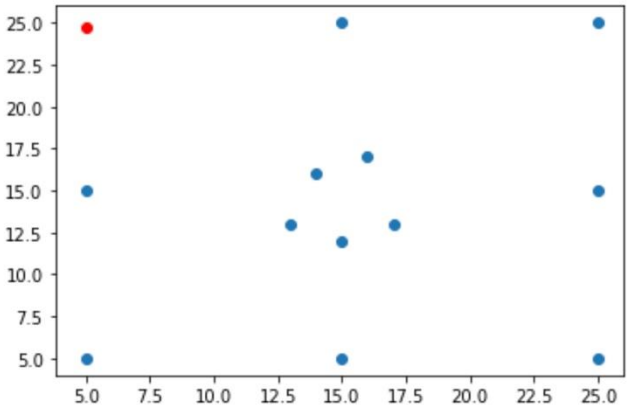
(그림 2) 결측치

다음 그림자료는 결측치가 아닌 12 개의 다른 데이터들의 y 축 좌표값의 평균을 내어 빨간색으로 표시된 데이터의 y 축 좌표값에 대체시킨 것이다. 그림에서 알아낼 수 있듯이 결측 데이터의 원래 y 축 좌표값과는 10 정도나 차이가 나며 부정확한 결과가 도출되었다.



(그림 3) 결측치를 평균으로 대체

다음 그림자료는 결측치의 y 좌표값을 전에 설명했던 계산방법으로 Z 값을 생성해서 대체시킨 결과이다. 원래 데이터와 거의 차이 나지 않는 값이 만들어졌음을 알 수 있다.



(그림 4) 결측치를 Z 값으로 대체

이렇게 Z 값을 결측 데이터들 대신에 사용해서 KNN 알고리즘을 사용해서 붓꽃(아이리스) 데이터를 구분하는 실험을 해보았다. 결측치를 위에서 만든 표준점수로 대체시켜 KNN 알고리즘을 구현하고 결측치를 평균으로 대체한 데이터와 표준점수로 대체한 데이터의 정확도를 비교했을 때 더 높은 정확도를 띄었다.

```
#after
from sklearn.neighbors import KNeighborsClassifier
classifier = KNeighborsClassifier(n_neighbors = 3)

X_train = np_data[:, 0:-1]
Y_train = np_data[:, 4: ]
Y_train = Y_train.reshape(Y_train.size,)
classifier.fit(X_train, Y_train)

print(classifier.score(X_train, Y_train))
```

```
0.9733333333333334
```

(그림 5) 결측치를 표준점수로 대체

```
#after
from sklearn.neighbors import KNeighborsClassifier
classifier = KNeighborsClassifier(n_neighbors = 3)

X_train = np_data[:, 0:-1]
Y_train = np_data[:, 4: ]
Y_train = Y_train.reshape(Y_train.size,)
classifier.fit(X_train, Y_train)

print(classifier.score(X_train, Y_train))
```

```
0.9533333333333334
```

(그림 6) 결측치를 평균으로 대체

3. 결론

붓꽃(아이리스) 데이터를 구별하기 위해서 KNN 알고리즘을 구현하는 실험을 하던 중 결측치가 있는 데이터는 어떻게 해야 결측치를 원래 데이터와 가장 유사한 값으로 복원시킬 수 있을까 고민하던 중 대학수학능력시험에서 성적을 낼때 사용하는 표준점수가 떠올랐고, 표준점수를 결측치 대체값으로 사용하니 평균으로 대체한 것보다 더 좋은 결과가 도출되었다.

참고문헌

[1] 고길곤 외 1, 설문자료의 결측치 처리방법에 관한 연구: 다중대체법과 재조사법을 중심으로, “행정논총”, 제 54 권, 제 4 호, 291~319, 2016 년 12 월

[2] https://ko.wikipedia.org/wiki/%ED%91%9C%EC%A4%80_%EC%A0%90%EC%88%98