toMerge

JeonJunho

2023-06-02

##### Data Load and Declare its notation

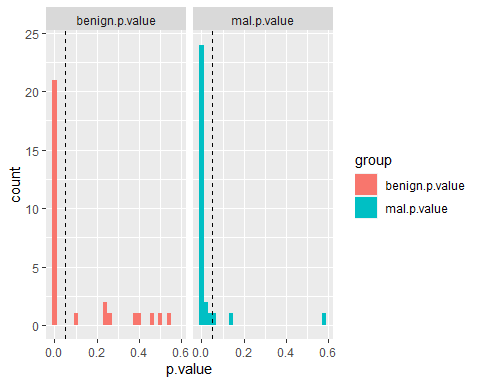
수업 시간에 배운 내용을 토대로 분석하고자 하였음. 수업에서는 MVN인 dataset에 대해 classification을 진행하는 방식을 배웠으므로,

1. 주어진 Dataset이 normal인지 확인
2. 해당 Daset이 normal이 아니면, normalize
3. 수업 내용을 바탕으로 최적 ECM\_rule을 작성
4. Cross Validation을 통해 효율성 확인

### 1. 주어진 Dataset이 normal인지 확인

## # A tibble: 30 × 3  
## variableName mal.p.value benign.p.value  
## <chr> <dbl> <dbl>  
## 1 radius\_mean 0.0233 3.85e- 1  
## 2 texture\_mean 0.00377 2.29e-12  
## 3 perimeter\_mean 0.0168 4.57e- 1  
## 4 area\_mean 0.00000916 2.60e- 1  
## 5 smoothness\_mean 0.138 2.69e- 3  
## 6 compactness\_mean 0.00000890 2.51e-13  
## 7 concavity\_mean 0.00000815 3.7 e-24  
## 8 concave.points\_mean 0.00000747 1.25e-11  
## 9 symmetry\_mean 0.000926 5.89e- 4  
## 10 fractal\_dimension\_mean 0.00000752 5.39e-19  
## # ℹ 20 more rows

정규성 검정의 p.value가 0.05 이상인 variable 확인



## # A tibble: 12 × 3  
## variableName group p.value  
## <chr> <chr> <dbl>  
## 1 radius\_mean benign.p.value 0.385   
## 2 perimeter\_mean benign.p.value 0.457   
## 3 area\_mean benign.p.value 0.260   
## 4 smoothness\_mean mal.p.value 0.138   
## 5 radius\_worst benign.p.value 0.491   
## 6 perimeter\_worst benign.p.value 0.535   
## 7 area\_worst benign.p.value 0.106   
## 8 smoothness\_worst mal.p.value 0.0665  
## 9 smoothness\_worst benign.p.value 0.247   
## 10 concave.points\_worst mal.p.value 0.581   
## 11 concave.points\_worst benign.p.value 0.248   
## 12 symmetry\_worst benign.p.value 0.395

주어진 dataset이 normal을 따르지 않음을 알 수 있다. 따라서, normalize를 한 뒤 classification을 진행하려고 한다.

### 2. 해당 Dataset이 normal이 아니면, normalize(transformation)

자료를 정규화할건데, 대표적 transformation method인 BoxCox-변환을 이용하고자 한다.

forecast packages의 BoxCox() 함수를 사용하였다.

##### After normalize

BoxCox-변환을 통해 정규화 과정을 거친 뒤 실제로 normal을 보이는지 정규성 검정을 진행하였다.

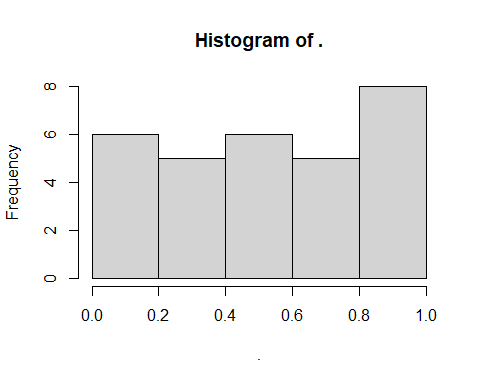
normal과 관련하여 정규성 검정을 할 때, 그 기준은 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3693611/> 이 링크를 참조하여 shapiro.test를 사용하려고 한다.

## texture\_mean area\_mean concavity\_mean concave.points\_mean   
## 0.020 0.041 0.005 0.034   
## radius\_se texture\_se   
## 0.003 0.000

## [1] "Normal이 아닌 변수의 갯수 : 13"

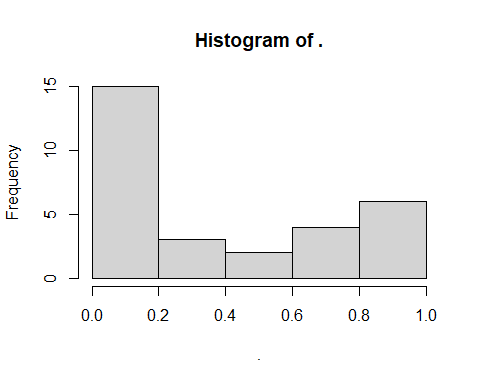
적지 않은 변수가 BoxCox-변환을 거치고 나서도 normal을 따르지 않음을 알 수 있다.

해당 변수들의 분포를 확인한 결과, outlier가 존재하는 것처럼 보인다. 따라서 outlier를 제거한 다음에 다시 정규변환을 거치고자 한다. 또한, BoxCox-변환은 양수에 대해서만 적용할 수 있으므로, 음수가 존재하는 variable에 대해서는 실수 전체로 확장시킨 변환인 Yeo-Johnson-변환을 적용하고자 한다.



## area\_se radius\_se fractal\_dimension\_mean   
## 0.05741079 0.06490157 0.09597181   
## area\_mean concavity\_worst texture\_mean   
## 0.17205173 0.18148719 0.18319468

Malignant group(이하 M)에 대해 정규변환을 실시한 결과, 모든 변수가 잘 정규화됐다.



## fractal\_dimension\_se compactness\_se concavity\_se   
## 0 0 0   
## smoothness\_se concavity\_mean fractal\_dimension\_worst   
## 0 0 0

## [1] "Normal이 아닌 변수의 갯수 : 14"

Benign group(이하 B)에 대해서는 잘 정규화가 되지 않았다. 아무래도 음수인 data에 대해 Boxcox based method를 시도하다보니 잘 적용이 되지 않은 모습이다.

### 3. 수업 내용을 바탕으로 최적의 ECM\_rule을 작성

해당 식을 그대로 구현하였다