1.분석목적

분석목적으로는 위스콘신 유방암 진단 데이터를 가지고 유방암 여부에 대한 분류를 하기 위한 것이다. 종속변수인(diagnosis)유방암 여부를 기준으로 두고 나머지 변수들(독립변수)을 가지고 분석을 하려 한다. 이를 통해 얼만큼 분류가 잘 되었는가, 유방암 여부를 분류할 때 사용한 모델의 정확성은 얼마나 되는가를 통계적으로 알아보려고 한다.

2. 분석절차

위에서 말했듯이 종속변수를 분류하려고 하는데 특히 이런 이진분류에서는 많은 방법들이 존재한다. Knn, svm, 군집분석, 로지스틱 회귀분석 등 정말 다양한 방법이 있는데 나는 k-means 클러스터링 방법을 사용하려고 한다. 클러스터링 분석법을 사용하기전 분석절차에 대해서 설명해 보려고 한다.

(1) 데이터셋에 결측 값이나 중복된 값이 존재하는지를 확인해본다.

(2) 독립변수 간의 다중 공선성을 확인하여 그에 따른 후속처리를 한다.

(3) T.test를 통해서 변수들 선별 처리를 한다.

(4) 로지스틱 회귀분석을 통해서 종속변수에 어떠한 변수들이 얼만큼 영향을 미치는지 확인해본다.

(5) 남은 독립변수들을 가지고 분류모델링(K-means)을 시행한다.

(6) 모델의 성능을 확인해본다.

3. 분석결과

3-1 변수탐색

먼저, 데이터 전처리 과정에서 결측 값이나 중복된 값은 존재하지 않았다. 다행히 깔끔한 데이터로 분석을 시작하게 되었다.

종속변수에 대한 분류분석을 시행하기 전에 적절한 독립변수의 개수가 중요하다. 너무 많지도 않고 그렇다고 너무 적은 개수도 아닌 적절하게 독립변수의 수를 조절하여 종속변수를 분류를 해보았다. 이러한 변수탐색 과정에서 분산팽창지수, t.test의 p-value값, 로지스틱 회귀분석의 후진소거법 이렇게 3가지의 방법이 사용되었다.

설명변수 간의 다중공선성을 확인하는 과정이 첫번째였다. 이것은 로지스틱 회귀분석을 들어가기 전 설명변수간의 독립성을 확인해보기 위한 것이었다. 이 과정에서 상관계수를 구하였을 때 여러 변수들 간의 높은 상관성을 볼 수 있었다. 즉, 다중공선성이 존재하였고 이상태로 회귀분석을 시

시 하였을 때 잘못된 결과가 나올 가능성이 높았다. 그래서 VIF(분산팽창지수)값 기준으로 10이상인 변수들, 다시 말해서 다중공선성이 높게 존재하는 변수들을 제거하기로 하였다. VIF값이 10 이상인 가장 높은 변수를 제거하고 다시 구동하여 또 높은 변수를 제거하고, 이렇게 반복적인 과정을 거쳐 VIF값이 10이상인 변수가 없을 때까지 반복하였다. 이렇게 하여 30개의 독립변수중 17개만 추려내게 되었다. 이렇게 추려낸 변수들을 스케일링 하여 표준화 작업을 해주었다.

그 후로는, 종속변수(B,M)의 그룹별 독립변수 간의 차이가 존재하는지에 대한 여부를 t.test를 통해 알아 보았고 그에 대한 p-value 값이 0.05 보다 큰 변수들을 제거해주었다.

1	points_mean	7.101150e-116
2	area_worst	2.828848e-97
3	perimeter_se	1.651905e-47
4	smoothness_worst	6.575144e-26
5	symmetry_worst	2.951121e-25
6	texture_mean	4.058636e-25
7	points_se	3.072309e-24
8	smoothness_mean	1.051850e-18
9	symmetry_mean	5.733384e-16
10	dimension_worst	2.316432e-15
11	compactness_se	9.975995e-13
12	concavity_se	8.260176e-10
13	dimension_se	6.307355e-02
14	smoothness_se	1.102966e-01
15	dimension_mean	7.599368e-01
16	texture_se	8.433320e-01
17	symmetry_se	8.766418e-01

(t.test)

Dimension_se,smoothness_se, dimension_mean, texture_se, symmetry_se 이렇게 5개의 변수들이 유의수준 0.05보다 커서 제거 대상이 되었다.

남은 12개의 변수들을 독립변수로 가지고 좀더 변수탐색 과정을 하게 되었다. 마지막 과정으로는 로지스틱 회귀분석법이었다. 종속변수 유방암의 여부를 1과 0으로 설정을 한 후에 회귀분석을 시행하였다. 다중공선성의 문제에 대비하여 후진 소거법을 선택하여 종속변수에 유의한 영향을 미치지 않는 변수들을 하나씩 제거해가면서 더 이상 제거할 변수가 없을 때까지 시행되었다.

	Df	Deviance	AIC
<none>		70.529	88.529
- ndat5$concavity_se	1	72.900	88.900
- ndat5$points_mean	1	77.683	93.683
- ndat5$compactness_se	1	79.509	95.509
- ndat5$perimeter_se	1	80.475	96.475
- ndat5$symmetry_worst	1	85.453	101.453
- ndat5$smoothness_worst	1	87.683	103.683
- ndat5$texture_mean	1	108.147	124.147
- ndat5$area_worst	1	138.745	154.745

(로지스틱 회귀분석 후진제거법)

3번의 변수탐색 과정을 거쳐서 30개의 많은 변수에서 최종적으로는 왼쪽의 결과와 같은 8개의 독립(설명)변수가 선택이 되었다.

3-2 분류 모델링

이제는 종속변수인 유방암 여부를 지금까지의 선택된 8개의 독립변수들로 분류모델을 만들려 한다. 분류 모델링으로 사용할 모델은 K-means clustering 이다.

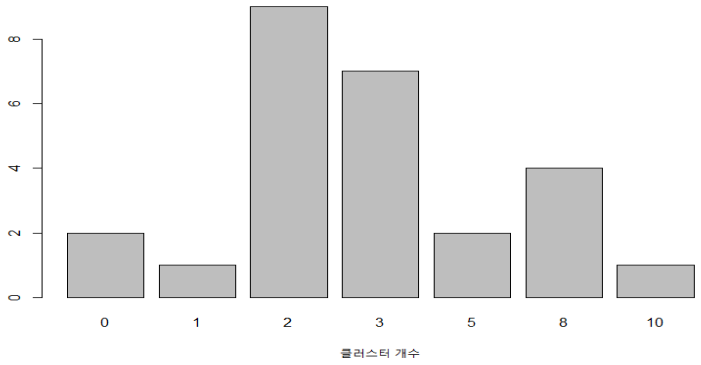
내가 k-means를 사용한 이유는 분류하기에 있어서 여러 데이터에 대한 적용이 간편하다고 생각한 점이 가장 크다. 물론 종속변수에 대한 해석적 관점으로 보았을 때는 로지스틱 회귀분석도 좋은 선택이지만 k의 개수만 적절하게 설정해주면 분류 목적인 차원에서만 보았을 때는 더욱 알맞다고 생각한 경향이 크기 때문이다. 또한 이렇게 많고 복잡한 설명변수들에 있어서 큰 사전정보 없이도 쉽게 할 수 있다는 장점도 사용 이유 중 하나이다.

방금 얘기한 것처럼 적절한 k의 개수가 k-means 클러스터링에서 가장 중요하면서도 까다로운 요소이다. 그래서 임의로 정하는 것이 아닌 여러가지 방법을 사용하여 k의 개수를 설정하였다.

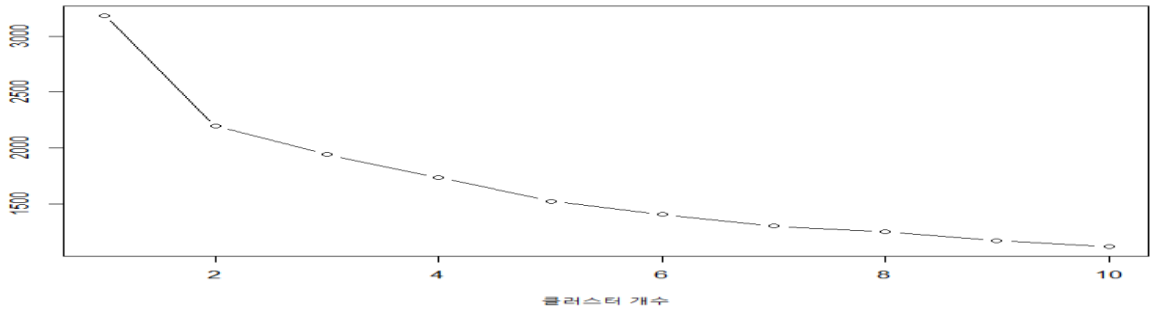
(1) 최적의 군집수 투표 결과

```
* Among all indices:
* 9 proposed 2 as the best number of clusters
* 7 proposed 3 as the best number of clusters
* 2 proposed 5 as the best number of clusters
* 4 proposed 8 as the best number of clusters
* 1 proposed 10 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  2
```

(2) 군집 수 결정 그래프



(3) WSS 그래프(오차제곱합)



2개의 군집수가 가장 적절하다는 투표값에 더해 오차제곱합이 많이 떨어져 완만하게 감소한다는 것을 확인하여 k는 2개로 결정을 하였다.

본격적으로 train데이터와 test데이터로 나누어 k-means 클러스터링을 실시하였다. Train과 test의

비율은 7:3으로 나누었다. 또한 1군집과 2군집에 이름을 맞추기 위하여 그전의 "1=M" , "0=B"으로 전처리 했던 데이터를 "2=M", "1=B"로 바꾼 후 군집화를 하였다.

학습을 시킨 후에 테스팅 데이터를 가지고 얼만큼의 정확도를 가지고 있는지 확인해 보았다.

```
testpred   1   2
       1 101   6
       2   6  57
> mean(testpred == testing$a)
[1] 0.9294118
```

처음 돌렸을 때의 결과이다. 실제 1인 군집 107개중 101개를 맞추었고 실제 2인 군집 63개중 57개를 맞추어 90프로에 가까운 정확도를 확인 할 수 있었다.

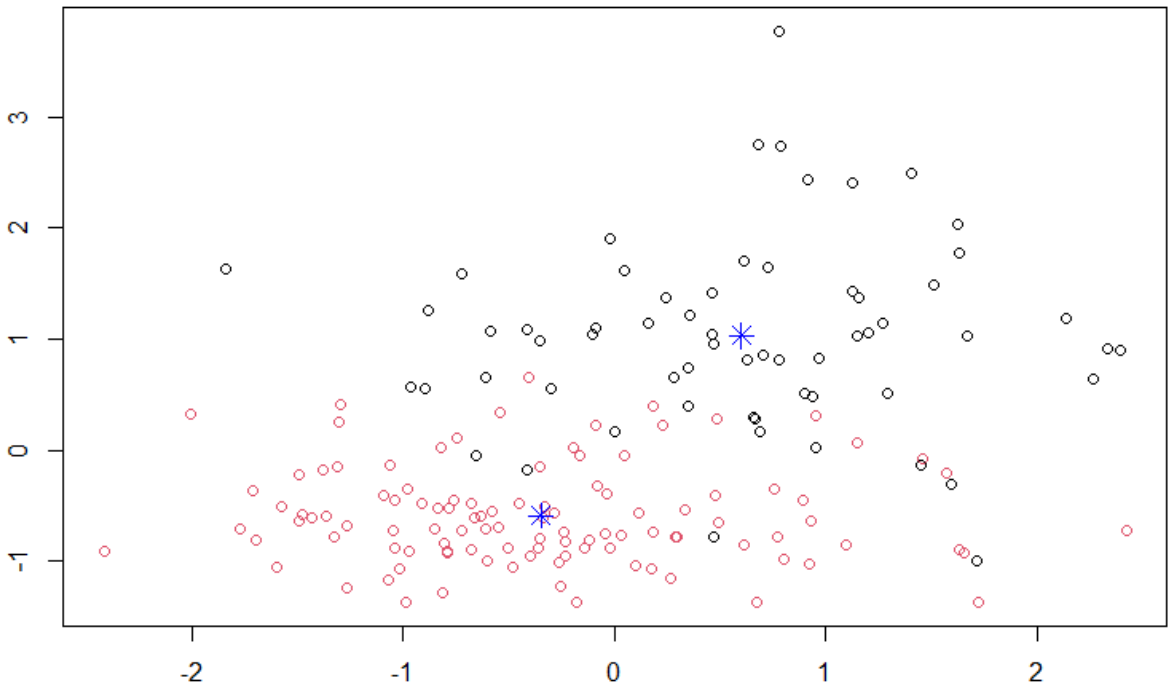
```
(0.9294118) (0.9) (0.9) (0.9) (0.9294118) (0.8294118) (0.9058824) (0.9) (0.9) (0.8705882)
(0.9) (0.9058824) (0.9) (0.8647059) (0.8705882) (0.8941176) (0.9) (0.8764706) (0.9058824) (0.8823529)
```

이렇게 테스팅 데이터를 20번 반복으로 돌렸을 때 나오는 결과 이다.

```
      평균     표준편차
0.8917353 0.02474896
```

정확성(Accuracy)의 20회 평균은 대략 0.89정도가 나오고 표준편차 또한 0.025정도로 낮은 편이라 좋은 결과라고 할 수 있다.

클러스터링 모델의 모델링 성능 평가지표로 확인하였을 때는 좋은 결과를 보여 주고 있었고 테스팅 데이터에 대한 클러스터링이 어떻게 되었는지를 시각적으로 확인하려고 한다.



두개의 군집으로 나누어서 그린 것이고 파란색 별 모양 점은 각 군집 별 중심점이다. 중심점 기준에서 보았을 때 두개의 군집이 고르게 분류 된 것을 볼 수 있었다. 또 한 위의 시각화는 정확도 0.9를 가지고 있는 그래프이다.

마지막으로 randindex함수를 통해서 실제 유방암의 여부와 군집간의 일치도를 나타낸 수정된 순

위지수를 구하였는데 대략 0.74정도가 나오게 되었다. 즉 수정된 순위지수를 사용함으로서 우연에 의해 발생하는 경우를 고려하여 74프로정도의 일치도를 보여주고 있다.

(my comments)

이렇게 여러가지의 과정들을 통해서 유방암 양성여부에 대한 분류 모델링을 하였는데, 정확도만이 중요한 것이 아니라 여러가지 분류 분석과 다양한 탐색방식에 의한 결과는 전부 다르기 때문에 데이터의 형식과 그에 맞는 상황에 따라 분석법을 선택하는 것이 맞다고 생각한다. 어떠한 방법은 맞고 틀리다가 아니라 그 분석방식의 장점을 이용하여 유용한 분석을 하는 것이 가장 중요하다고 생각하면서 분석을 마친다.