로지스틱 회귀분석과 SVM 의 서포트벡터를 이용한 종양 크기에 따른 유방암 진단 영향 분석

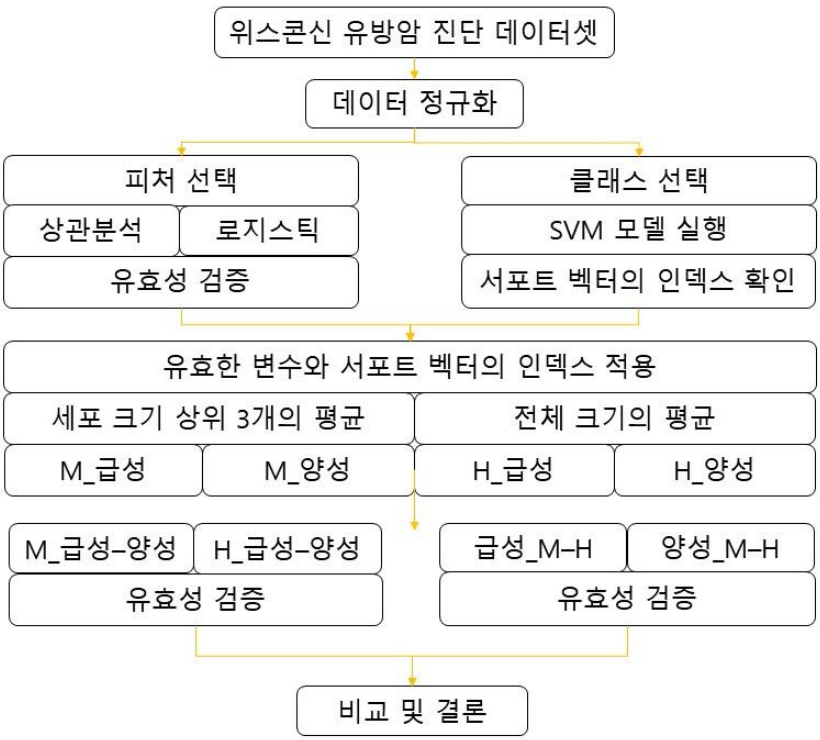
1. 장민승(19930120)
2. alstmd603@naver.com
3. 장민승(19930120)

요약

본 분석은 Wisconsin Breast Cancer Diagnostic Data Set(위스콘신 유방암 진단 데이터 셋)으로 종양의 크기가 유방암 진단에 영향을 주는지 알아본다. 상관분석과 로지스틱 회귀분석으로 유효한 피처를 선택하고, SVM 모델을 실행해 클래스 사이를 가장 멀리 분리하는 최대 마진 초평면에 가장 가까운 점들인 서포트 벡터 클래스로 유방암 진단에 기준이 되는 클래스를 선택했다. 전체 크기의 평균은 'mean' 컬럼의 데이터를 사용했고, 상위 크기 3 개의 평균은 'worst' 컬럼의 데이터를 사용했다.

분석 결과는 급성과 양성 모두 Radius(반지름), Compactness(조밀성), Perimeter(둘레)와 Area(면적) 변수와 다르게 Concavity(오목함)와 Concave points(오목점)는 상위 크기 3 개의 평균이 전체 크기의 평균보다 크다는 것과, 종양이 급성일 경우 양성보다 그 차이가 더 크다는 것을 확인했다. 그리고 모든 변수에서 상이한 차이를 확인할 수 없었기 때문에 종양의 크기는 유방암 진단에 영향을 준다고 할 수 없다.

분석 방법



데이터 탐색

유방암은 영상의학검사를 통해 양성여부를 판단한다. 그런데 양성 결과가 나와도 모두 암이라고 하지 않는다. 양성 결과의 경우 양성 유방 질환과 악성 유방 질환으로 나누는데, 이 악성 유방 질환이 유방암이다. 악성질환은 조직검사 결과로 판별하는데, UCI 에서 제공하는 'Wisconsin Breast Cancer Diagnostic Data Set'이 유방암 진단을 위한 데이터다.

데이터 셋에는 569 개의 암 조직검사 예시가 있으며, 각 예시는 32 개의 특징을 갖는다. 마지막 33 번째 컬럼인 X 는 결측 값이다. ID(아이디)는 환자 식별 번호를 나타내고, Diagnosis(진단)는 Malignant(악성)을 나타내는 "M" 과 Begin(양성)을 나타내는 "B"로 코드화 되어있다. 다른 30 개 컬럼은 '_mean'(전체 크기의 평균), '_se'(표준 오차), '_worst'(상위 크기 3 개의 평균)가 있다. 분석에는 전체 크기의 평균과 상위 크기 3 개의 평균을 사용했다.

Radius (반지름)	중심에서 외벽까지 거리들의 평균값
Texture (질감)	Gray-Scale(광도) 값들의 표준편차
Perimeter (둘레)	중심 종양의 크기
Area (면적)	
Smoothness (매끄러움)	반경 길이의 지역적 변화
Compactness (조밀성)	둘레^2/면적-1
Concavity (오목함)	윤곽의 오목한 부분의 정도
Concave points (오목점)	윤곽의 오목한 부분
Symmetry (대칭성)	
Fractal Dimension (프랙탈 차원)	해안선 근사 -1

분석 단계별 결과

#요약 통계량 확인

변수끼리 정확한 비교를 위해 데이터 분포를 확인했다.

	Radius (반지름)	Texture (질감)	Perimeter (둘레)	Area (면적)
최솟값	6.981	9.71	43.79	143.5
평균	14.127	19.29	91.97	654.9
최댓값	28.110	39.28	188.50	2501.0

편의상 순서대로 4 개 변수까지 확인하더라도 척도 범위가 전부 다르다는 것을 알 수 있다. 변수가 거리 공식에 상대적으로 동일하게 기여하도록 재조정을 해야 한다.

#최소-최대 정규화

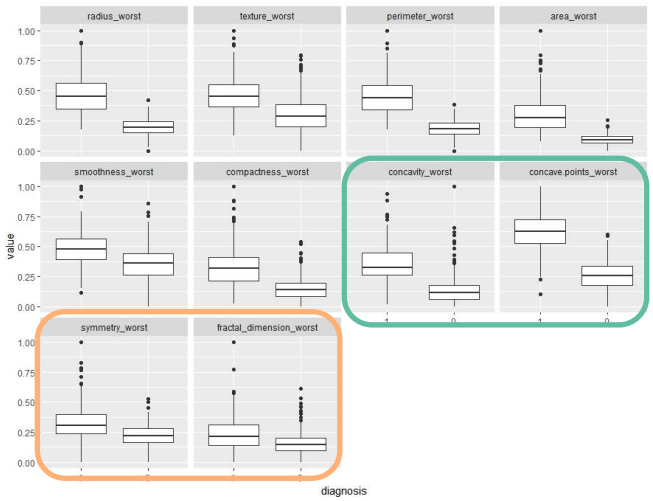
정규화를 위해 최소-최대 정규화를 했다. 이 과정은 변수 X 의 개별 값에서 최솟값을 빼고 X 의 범위로 나눠 모든 값이 0 에서 1 사이의 범위에 있게 한다. 이를 통해 원래 값이 얼마나 멀리 위치하는지 0%에서 100%까지 나타내 해석할 수 있다.

	Radius (반지름)	Texture (질감)	Perimeter (둘레)	Area (면적)
최솟값	0	0	0	0
최댓값	1	1	1	1

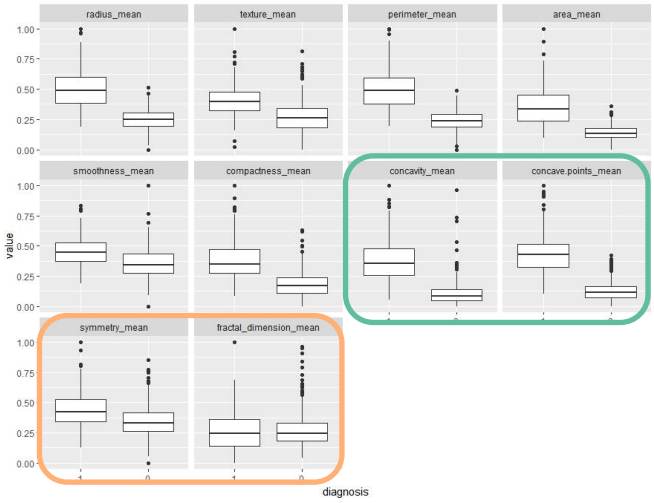
#분포 확인

변수의 개수가 10 개로 많기 때문에 변수 간의 상관성을 확인할 필요가 있다. 세포 크기와 Diagnosis(진단) 컬럼에서 M(악성)과 B(양성)을 기준으로 데이터셋을 나눠 박스 플롯으로 변수 별 분포를 확인했다.

상위 크기의 3 개 평균



전체 종양 크기의 평균

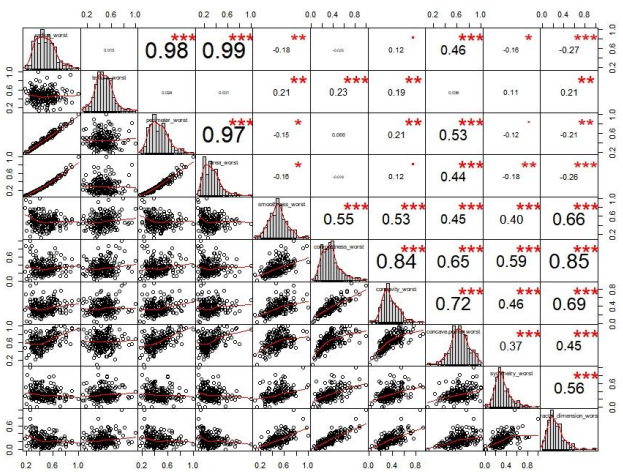


두 클래스(급성과 양성)의 중위 값이 떨어져 있는 정도와 분포를 통해 Concavity(오목함)와 Concave points(오목점)가 중요한 피처일 것이라고 생각할 수 있고, Symmetry(대칭성)와 Fractal Dimension(프랙탈 차원) 피처는 중위 값이 거의 분리되지 않아 적당한 피처는 아니라고 생각 할 수 있다.

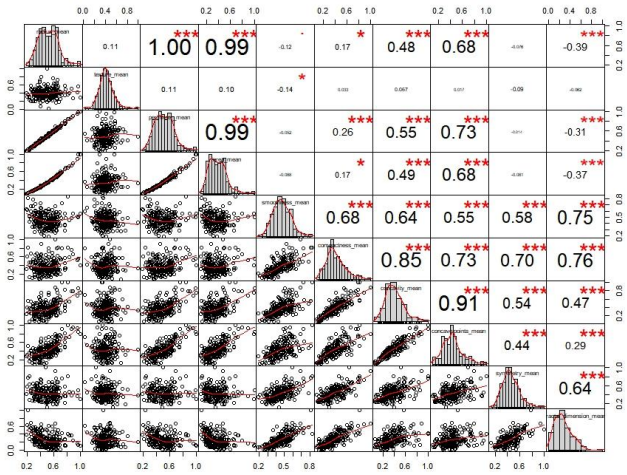
#상관분석

종양 크기와 Diagnosis(진단) 컬럼에서 M(악성)과 B(양성)을 기준으로 변수 간의 상관관계를 확인했다.

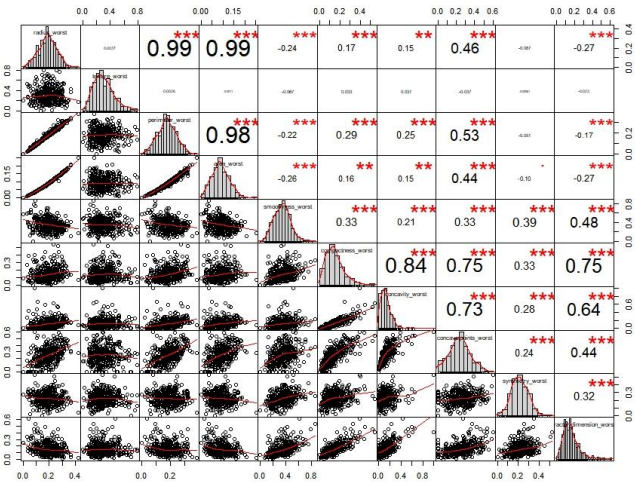
상위 크기 3 개의 평균_급성



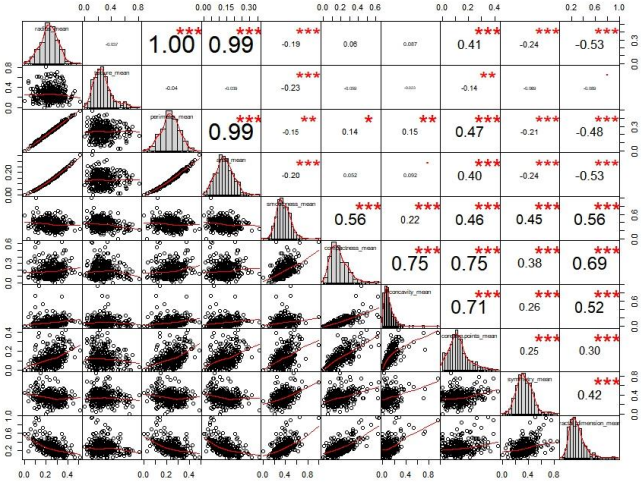
전체 크기의 평균_급성



상위 크기 3 개의 평균_양성



전체 크기의 평균_양성



상관분석 계수의 해석(Rea & Parker, 2005)[1]을 따라 분포를 통해 중요한 피처라고 생각한 Concavity(오목함)와 Concave points(오목점)의 상관계수가 ±0.6 ~ ±0.8 미만인 매우 강한 상관관계를 나타내는 변수를 선택해 변수의 유의성을 확인했다. 분포 확인 단계에서 Symmetry(대칭성)와 Fractal Dimension(프랙탈 차원)은 적당한 피처가 아니라고 생각했기 때문에 고려하지 않았다.

상관계수					
변수 1	변수 2	상위_급성	전체_급성	상위_양성	전체_양성
조밀성(Compactness)	오목함(Concavity)	+0.84	+0.85	+0.84	+0.84
오목함(Concavity)	오목점(Concave points)	+0.72	+0.91	+0.73	+0.73
조밀성(Compactness)	오목점(Concave points)	+0.65	+0.73	+0.75	+0.75
반지름(Radius)	오목점(Concave points)	+0.46	+0.68	+0.46	+0.41
둘레(Perimeter)	오목점(Concave points)	+0.53	+0.73	+0.53	+0.47
면적(Area)	오목점(Concave points)	+0.44	+0.68	+0.44	+0.40

#상관관계의 유효성 검증

Pearson 상관계수로 p-value 값을 확인해 상관관계의 유효성을 검증했다.

유효성 검증					
변수 1	변수 2	상위_급성	전체_급성	상위_양성	전체_양성
조밀성(Compactness)	오목함(Concavity)	< 2.2e-16	< 2.2e-16		< 2.2e-16
오목함(Concavity)	오목점(Concave points)				
조밀성(Compactness)	오목점(Concave points)				
반지름(Radius)	오목점(Concave points)	1.639e-12			4.765e-16
둘레(Perimeter)	오목점(Concave points)	< 2.2e-16			< 2.2e-16
면적(Area)	오목점(Concave points)	1.314e-11			2.106e-15

[1] Rea,LM., & Parker,R.A (2005). Designing & Conducting Survey Research A Comprehensive Guide (3rd Edition). San Francisco, CA: Jossey-Bass.

p-value 를 확인한 결과, 95% 신뢰구간에서 모든 상관관계가 유의하다는 것을 확인했다.

#로지스틱 회귀분석

종양 크기 기준으로 훈련 데이터 80% (455 개)와 테스트 데이터 20% (114 개)로 로지스틱 회귀 분석을 했다. 모든 입력 변수로 10-fold cross validation 을 한 후에 로지스틱 모형을 만든 후 최량 부분 집합과 비교해 성능이 더 나은 모델의 변수를 확인했다.

모든 입력 변수_상위 크기 3 개의 평균						
유의한 변수	p-value	TP	TN(1 종오류)	FN(2 종오류)	FP	정확도
Texture (질감)	4.81e-05	26	0	6	82	94.74%
Smoothness (매끄러움)	0.0208					
Concave points (오목점)	0.0398					

모든 입력 변수_전체 크기의 평균						
유의한 변수	p-value	TP	TN(1 종오류)	FN(2 종오류)	FP	정확도
Texture (질감)	9.63e-09	24	2	1	77	88.6%
Area (면적)	0.42634					
Smoothness (매끄러움)	0.006958					
Concave points (오목점)	0.084046					

모든 입력 변수로 적용한 로지스틱 회귀 모델의 분류 정확도는 각각 94.74%와 88.6%로 확인했다. 입력 변수에 따른 모델 성능 향상을 위해 최량 부분 집합으로 모형을 축소해 확인한다.

최량 부분 집합_상위 크기 3 개의 평균						
유의한 변수	p-value	TP	TN(1 종오류)	FN(2 종오류)	FP	정확도
Texture (질감)	8.731061e-07	26	0	6	82	94.74%
Area (면적)	6.501739e-10					
Smoothness (매끄러움)	8.411179e-03					
Concave points (오목점)	1.088279e-03					

최량 부분 집합_전체 크기의 평균						
유의한 변수	p-value	TP	TN(1 종오류)	FN(2 종오류)	FP	정확도
Texture (질감)	4.994524e-09	24	2	9	79	90.4%
Area (면적)	8.367314e-08					
Concave points (오목점)	9.909373e-15					

최량 부분 집합의 변수로 적용한 로지스틱 회귀 모델의 분류 정확도는 각각 94.74%와 90.4%로 확인했다.

#피처 선택

전체 크기의 평균에서 Texture(질감), Area(면적) 그리고 Concave points(오목점) 변수로 로지스틱 회귀 모델을 만들었을 때 분류 정확도가 최량 부분 집합보다 높았다. 상관계수와 함께 비교하면 Texture(질감)은 다른 변수들과 유의미한 상관성을 보이지 않았고, Smoothness(매끄러움)는 상위 크기 3 개의 평균에서 역시 다른 변수들과 유의미한 상관성을 보이지 않기 때문에 유효한 피처로 고려하지 않았다. 그래서 Area(면적)와 Concave points(오목점)와 매우 강한 상관관계인 Concavity(오목함), Compactness(조밀성), Radius (반지름) 그리고 Perimeter(둘레)를 분석대상 피처로 선택했다.

#SVM 모델 실행

서포트 벡터 클래스로 유방암 진단에 기준이 되는 클래스를 찾기 위해 SVM 모델을 실행했다. 모든 변수와 최량 부분 집합으로 선택한 변수 각각 세포크기를 기준으로 훈련 데이터 80%와 테스트 데이터 20%로 나누고, 10-fold cross validation 으로 최적의 파라미터로 모델을 실행한 뒤 confusion matrix 로 4 가지 커널 함수 별로 성능을 비교했다.

SVM 의 특징은 커널 트릭을 사용해 고차원 공간으로 매핑을 한다는 것이다. 커널 타입으로 Linear(선형)는 데이터를 전혀 변환하지 않고 내적으로 간단히 표현하고, Polynomial(다항)은 간단하게 비선형 변환을 더하고, Radial Basis(방사 기저)는 많은 유형의 데이터에서 잘 작동되는데 유클리드 공간을 일반화한 개념인 힐베르트 공간이 되고, Sigmoid(시그모이드)는 시그모이드 활성 함수를 사용한 신경망과 유사한 SVM 모델을 만든다.

모든 입력 변수_상위 크기 3 개의 평균					
커널 함수	TP	TN(1 종오류)	FN(2 종오류)	FP	정확도
Linear(선형)	26	2	0	86	98.25%
Polynomial(다항)	19	0	7	88	93.86%
Radial Basis(방사기저)	26	11	0	77	90.35%
Sigmoid(시그모이드)	25	5	1	83	94.74%

모든 입력 변수_전체 크기의 평균					
커널 함수	TP	TN(1 종오류)	FN(2 종오류)	FP	정확도
Linear(선형)	24	2	10	78	89.47%
Polynomial(다항)	21	5	0	88	95.61%
Radial Basis(방사기저)	25	1	12	76	88.6%
Sigmoid(시그모이드)	24	2	4	84	94.74%

최량 부분 집합_상위 크기 3 개의 평균					
커널 함수	TP	TN(1 종오류)	FN(2 종오류)	FP	정확도
Linear(선형)	25	1	7	81	92.98%
Polynomial(다항)	21	5	1	87	94.74%
Radial Basis(방사기저)	26	0	5	83	95.61%
Sigmoid(시그모이드)	25	1	10	78	90.35%

최량 부분 집합_전체 크기의 평균					
커널 함수	TP	TN(1 종오류)	FN(2 종오류)	FP	정확도
Linear(선형)	24	2	3	85	95.61%
Polynomial(다항)	21	5	1	87	94.74%
Radial Basis(방사기저)	25	1	8	80	92.11%
Sigmoid(시그모이드)	25	1	6	82	93.86%

모든 입력 변수로 실행한 SVM 모델이 최량 부분 집합보다 분류 정확도가 더 높다는 것을 확인했다. 그래서 모든 입력 변수를 반영한 SVM 모델로 상위 크기 3 개의 평균은 Linear(선형) 함수를 사용하고, 전체 크기의 평균은 Polynomial(다항) 함수를 사용한다.

#클래스 선택

본 분석에서는 모델의 Overfitting 의 한계가 있지만, 최대한 정확하고 많은 서포트 벡터를 가져오는게 중요하다고 생각하기 때문에 전체 데이터를 테스트 데이터로 서포트 벡터를 확인한다.

서포트 벡터			
상위 평균_ Linear(선형)	전체 평균_ Polynomial(다항)	중복 서포트 벡터	유일 서포트 벡터
54	233	51	236

236 개의 서포트 벡터로 클래스를 확인했다.

#상관관계인 변수와 서포트벡터 클래스를 적용한 전체 크기의 평균 급성과 양성 차이 확인

전체 크기의 평균 내 급성과 양성 차이					
Radius	Perimeter	Area	Compactness	Concavity	Concave points
0.1469164	0.1397205	0.1112294	0.09413234	0.1154362	0.1942037
p-value					
<2.2e-16	<2.2e-16	<2.2e-16	5.109e-05	4.06e-12	<2.2e-16

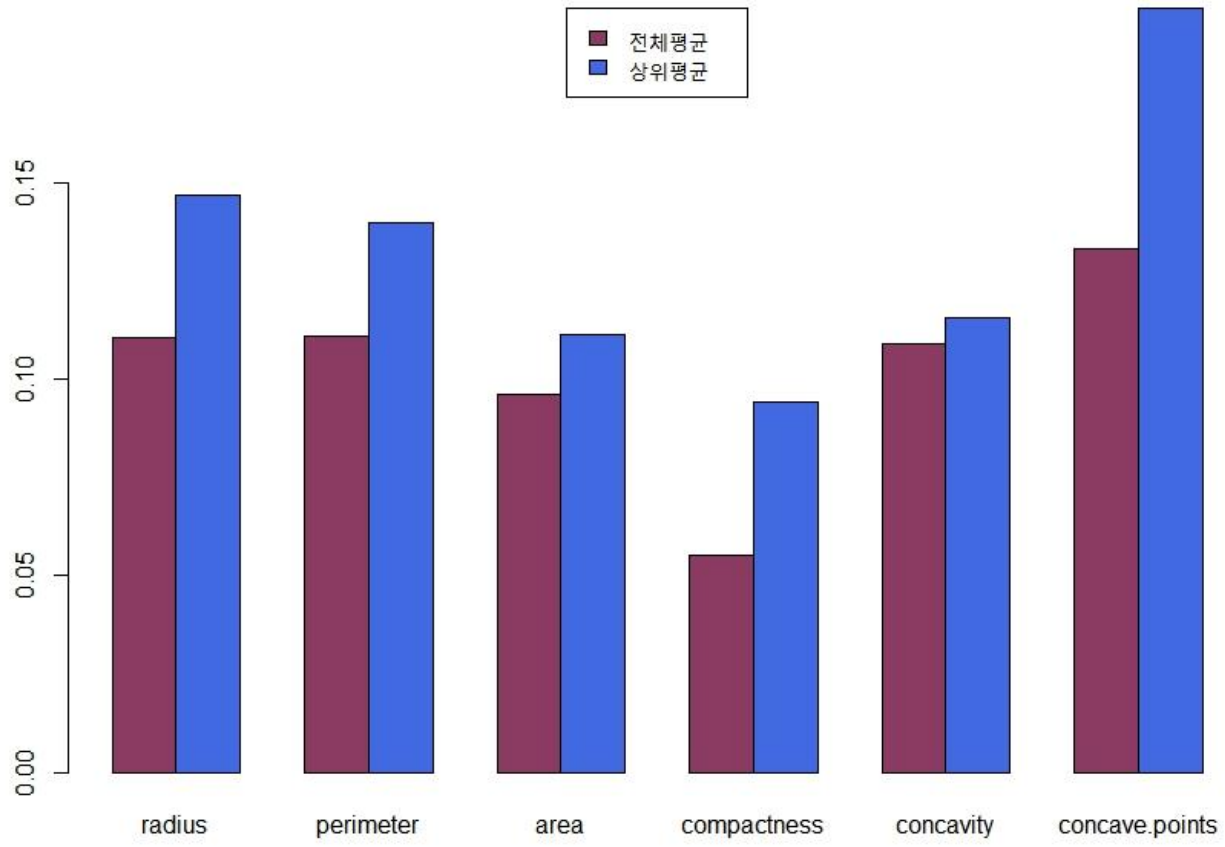
p-value 를 확인한 결과, 95% 신뢰구간에서 모든 차이가 유의하다는 것을 확인했다.

#상관관계인 변수와 서포트벡터 클래스를 적용한 상위 크기 3 개의 평균 급성과 양성 차이 확인

상위 크기 3 개의 평균 내 급성과 양성 차이					
Radius	Perimeter	Area	Compactness	Concavity	Concave points
0.1469164	0.1397205	0.1112294	0.09413234	0.1154362	0.1942037
p-value					
<2.2e-16	<2.2e-16	<2.2e-16	2.709e-09	7.957e-13	<2.2e-16

p-value 를 확인한 결과, 95% 신뢰구간에서 모든 차이가 유의하다는 것을 확인했다.

#그래프 확인 및 해석



상위 크기 3 개의 평균과 전체 크기의 평균 각각 급성과 양성 차이 부등호					
Radius	Perimeter	Area	Compactness	Concavity	Concave points
전체<상위	전체<상위	전체<상위	전체<상위	전체<상위	전체<상위
0.03633203	0.02862638	0.0151359	0.0388511	0.006661515	0.06093432

비교 결과, 급성과 양성 모두 상위 크기 3 개의 평균이 전체 크기의 평균보다 크고, 그 중 Concave points(오목점)가 0.06093432 로 다른 피처의 차이 보다 크다는 것을 확인했다.

#상관관계인 변수와 서포트 벡터 인덱스를 적용한 상위 크기 3 개의 평균과 전체 크기의 평균 급성 차이

급성을 기준으로 상위 크기 3 개의 평균과 전체 크기의 평균 차이					
Radius	Perimeter	Area	Compactness	Concavity	Concave points
-0.02272828	-0.3525359	-0.03979614	-0.014509	0.03240577	0.2173412
p-value					
0.07874	0.003563	0.0003883	0.3905	0.0202	<2.2e-16

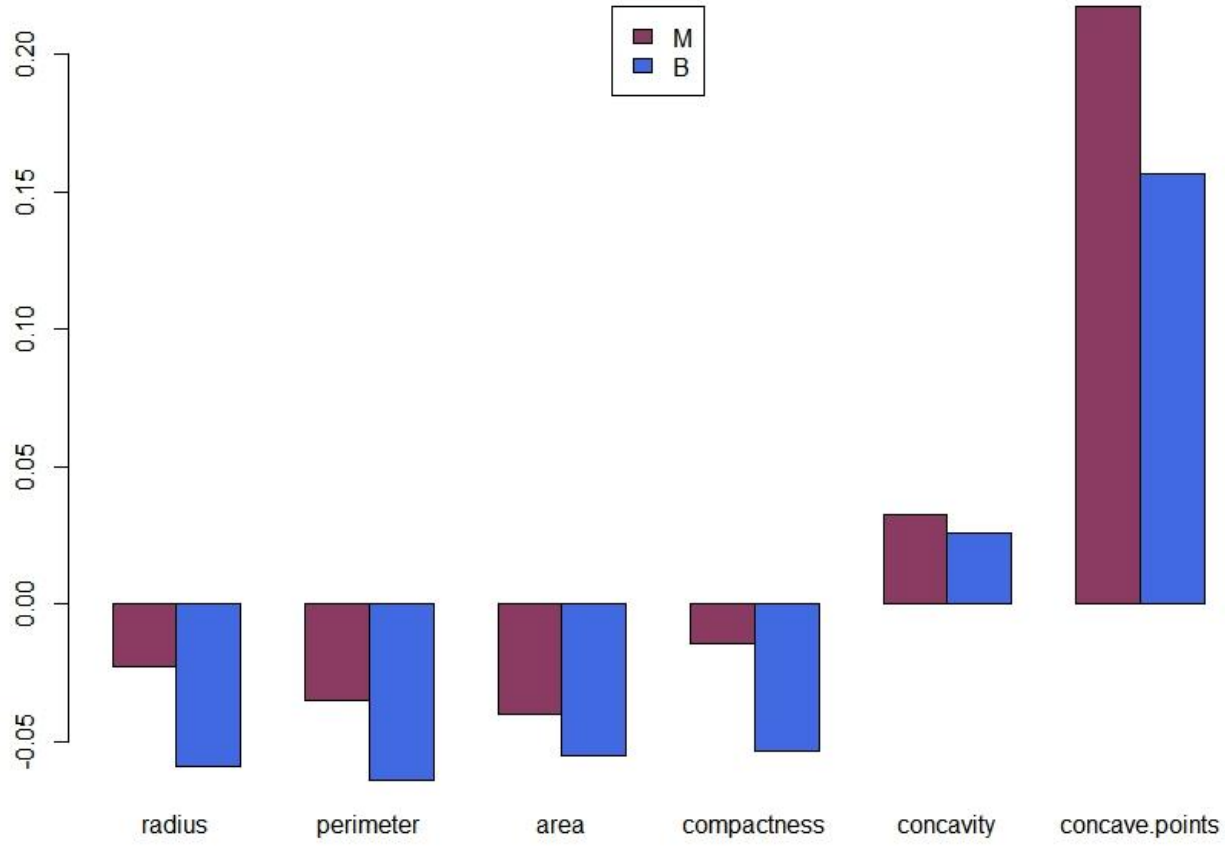
p-value 를 확인한 결과, 95% 신뢰구간에서 Radius(반지름)와 Compactness(조밀성)는 급성을 기준으로 상위 크기 3 개의 평균과 전체 크기의 평균 차이에서 p-value 가 0.05 보다 크므로 유효하지 않다는 것을 확인했다. 그리고 Concavity(오목함)와 Concave points(오목점)는 다른 변수와 다르게 상위 크기 3 개의 평균이 전체 크기의 평균보다 크다는 것을 확인했다.

#상관관계인 변수와 서포트 벡터 인덱스를 적용한 상위 크기 3 개의 평균 평균과 전체 크기의 평균 양성 차이

양성을 기준으로 상위 크기 3 개의 평균과 전체 크기의 평균 차이					
Radius	Perimeter	Area	Compactness	Concavity	Concave points
-0.05906031	-0.06386228	-0.05493204	-0.0533601	0.02574425	0.1564069
p-value					
1.442e-12	3.669e-15	<2.2e-16	8.447e-06	<2.2e-16	<2.2e-16

p-value 를 확인한 결과, 95% 신뢰구간에서 모든 차이가 유의하다는 것을 확인했다.

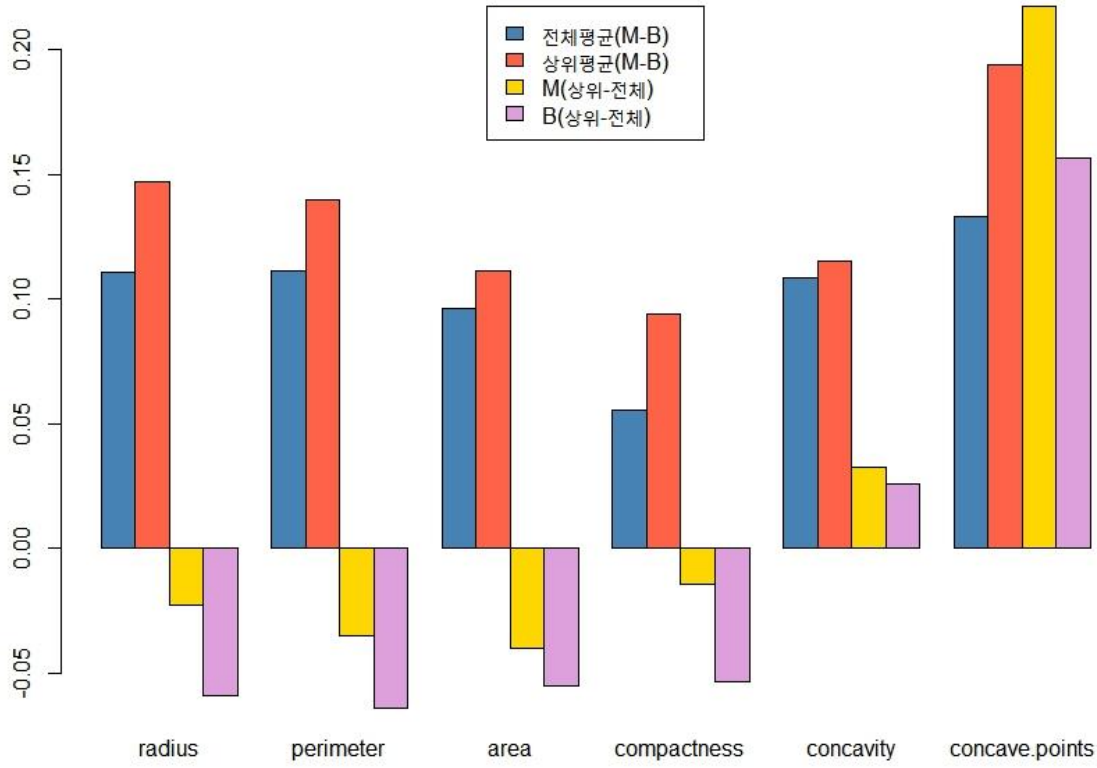
#그래프 확인 및 해석



암 진단 기준 급성과 양성 차이의 부등호					
Radius	Perimeter	Area	Compactness	Concavity	Concave points
M <B	M < B	M <B	M < B	M > B	M > B
0.03633203	0.02862638	0.0151359	0.0388511	0.006661515	0.06093432

p-value 를 확인한 결과, 95% 신뢰구간에서 모든 차이가 유의하다는 것을 확인했다. 그리고 급성과 양성 모두 Radius(반지름), Compactness(조밀성), Perimeter(둘레) 그리고 Area(면적) 변수와 다르게 Concavity(오목함)와 Concave points(오목점)는 상위 크기 3 개의 평균이 전체 크기의 평균보다 크다는 것과 급성일 경우 양성보다 그 차이가 더 크게 나타날 수 있다는 것을 확인했다.

결론



Radius	Perimeter	Area	Compactness	Concavity	Concave points
전체<상위	전체<상위	전체<상위	전체<상위	전체<상위	전체<상위
0.03633203	0.02862638	0.0151359	0.0388511	0.006661515	0.06093432
M <B	M < B	M <B	M < B	M > B	M > B
0.03633203	0.02862638	0.0151359	0.0388511	0.006661515	0.06093432

분석결과는 모든 변수에서 상이한 차이를 확인할 수 없었기 때문에 종양의 크기는 유방암 진단에 영향을 준다고 할 수 없다. 또한, 급성과 양성 모두 Radius(반지름), Compactness(조밀성), Perimeter(둘레)와 Area(면적) 변수와 다르게 Concavity(오목함)와 Concave points(오목점)는 상위 크기 3 개의 평균이 전체 크기의 평균보다 크다는 것과, 종양이 급성일 경우 양성보다 그 차이가 더 크다는 것을 확인했다. 이는 의료인이 급성 종양임을 판단할 경우에 크기에 영향을 받지 않아야 하고, 환자에게 종양 크기로 인한 공포감과 관리 소홀에 대한 주의를 주어야 할 필요성이 있다는 것을 알 수 있다.

코드

https://github.com/ChangMinSeung/dsc2018_the_preliminaries_wbcd