1. 배유진(19950926)

2. seow0124@naver.com

3. 이광호(19940510)/배유진(19950926)/송지원(19951223)

유방암 진행 정도 구별하기 위한 데이터 분석

데이터 스케치에 사용할 데이터를 탐색하는 과정에서, 유방암 진단과 관련된 데이터를 찾을 수 있었고, 이 데이터를 사용하게 되면 굉장히 유의미한 결과를 얻을 수 있을 것이라고 판단하여 스케치를 시작하게 되었습니다.

관련 배경 지식

유방암은 유방 조직에 암이 발생하는 것으로, 전 세계 여성들이 가장 무서워하는 암이며 암의 발병률은 2위, 사망률은 1위를 차지하고 있는 암입니다. 최근에는 남성에게도 유방암이 발병할 수 있다는 것이 알려져 그 관심이 커졌습니다. 또 그 진행 정도에 따라 5년 내의 생존율이 큰 차이를 보이는데 1-2기(초기)의 경우 90%이상 3-4기(말기)의 경우 약 20%입니다.

세침흡인세포검사법(FNA)은 본 데이터셋에서 데이터를 추출하는데 사용한 검사방식으로, 유방암의 검사법 중 하나이며 간단하면서도 정확도가 높고, 환자의 불편이 적다는 장점이 있습니다. 환자의 세포를 직접 채취하여, 이를 세포학적으로 검사하는 방식으로 세포의 특질을 검사하는 방식입니다.

자료 설명

데이터의 세부적인 출처는 UCI에서 제공하는 Breast-Cancer-Wisconsin 데이터 입니다.

- 데이터 출처 : UCI Machine learning repository
- 위스콘신 대학교에서 1984부터 유방암 환자들의 FNA검사에서 수집한 9가지의 세포의 특질 데이터
- 699명의 데이터
- 16개의 결측치 존재 (본 스케치에서는 데이터의 결측치를 제외하고 사용)

변수 이름 및 설명

(Sample Code Number와 Class를 제외한 모든 변수는 1 ~ 10 사이의 값을 가지며 변수의 정도를 나타냅니다)

Sample Code Number : 환자들의 식별번호

Clump Thickness : 세포의 층의 정도

Uniformity of Cell size : 세포 크기의 일관성

Uniformity of Cell shape : 세포의 모양의 균일성

Marginal Adhesion : 상피 바깥쪽 세포의 밀집 정도

Single epithelial cell size(단일 상피세포 크기) : 상피세포의 확장 정도(세포가 커진 정도)

Bare Nuclei : 세포질로 둘러싸이지 않은 세포의 비율

Bland Chromatin : 핵의 균일한 질감

Normal Nucleoli : Nucleoli의 가시성

(Nucleoli – 세포 핵에서 RNA를 보관하고 있는 작은 부분)

Mitoses : 세포의 분열 활성도 수준

Class : 암의 진행 정도(2-초기 , 4-말기)

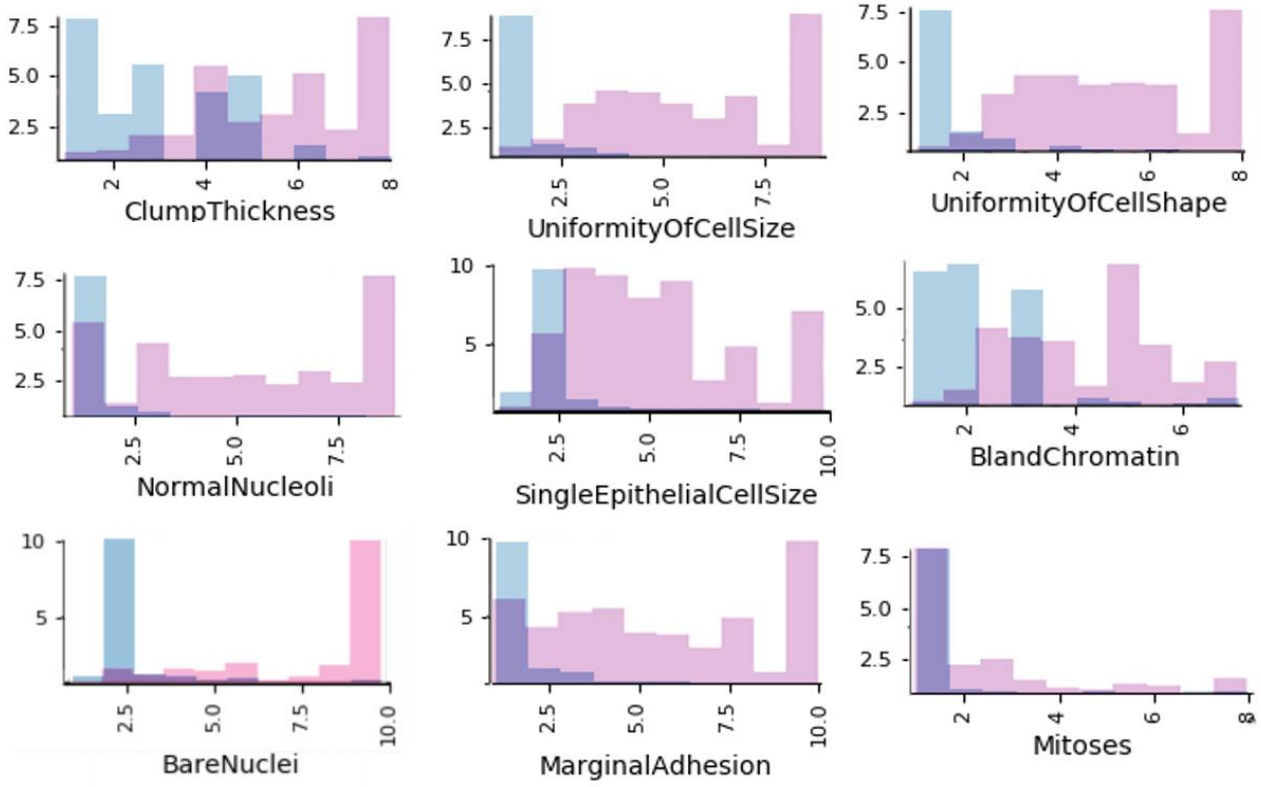
Class2 (초기) 와 Class4 (말기) 의 차이

Class2 (초기)와 Class4 (말기)의 차이를 알아보기 위해서 일부 표본의 데이터를 비교해 보았습니다. Class2 의 5개의 데이터와 Class4 의 5개의 데이터는 다음과 같습니다.

Class	Clump Thickness	Uniformity OfCellSize	Uniformity OfCellShape	Marginal Adhesion	SingleEpithelial CellSize	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses
2	5	1	1	1	2	1	3	1	1
2	2	1	1	1	2	1	2	1	1
2	3	1	1	1	2	2	3	1	1
2	6	8	8	1	3	4	3	7	1
2	4	1	1	3	2	1	3	1	1
4	8	10	10	8	7	10	9	7	1
4	5	3	3	3	2	3	4	4	1
4	8	7	5	10	7	9	5	5	4
4	7	4	6	4	6	1	4	3	1
4	10	7	7	6	4	10	4	1	2

여러 변수 중 Bare Nuclei의 클래스 간 차이가 분명합니다. Class4의 Bare Nuclei 값은 Class2의 Bare Nuclei 값에 비해 값이 월등히 큰 것을 알 수 있습니다. 또한 Class2의 변수가 Class4의 변수에 비해 모든 수치가 전체적으로 낮은 것을 알 수 있습니다.

아래는 각 변수 별 빈도수를 비교한 그래프입니다. 파란 막대는 Class2, 분홍 막대는 Class4의 빈도를 나타냅니다.



Mitoses 변수의 분포를 보면 상당히 겹쳐 있는 것을 볼 수 있고, Bare Nuclei 변수는 두 그룹의 분포가 확연히 구분 되어있는 것을 알 수 있습니다. 이를 통해 Bare Nuclei 변수는 Mitoses 변수에 비하여 클래스를 구분 하는 데 더 적합하다고 예상할 수 있습니다. 또한 Uniformity of cell size와 Uniformity of cell shape 변수의 경우, Class2는 주로 0~2 사이에 집중 되어있고, Class4는 주로 3~10 사이에 집중 되어있는 것을 확인하였습니다. 이러한 내용들을 통해 영향력 있는 변수를 정할 수 있었습니다.

추가적으로 아래는 Class2와 Class4의 각 변수의 평균과 각 평균의 차이입니다.

	Class 2 평균	Class 4 평균	평균의 차이
ClumpThickness	2.956331878	7.195020747	4.238688869
UniformityOfCellSize	1.325327511	6.572614108	5.247286597
UniformityOfCellShape	1.443231441	6.560165975	5.116934534
MarginalAdhesion	1.364628821	5.547717842	4.183089021
SingleEpithelialCellSize	2.120087336	5.298755187	3.17866785
BareNuclei	1.346846847	7.627615063	6.280768216
BlandChromatin	2.100436681	5.979253112	3.878816431
NormalNucleoli	1.290393013	5.863070539	4.572677526
Mitoses	1.063318777	2.589211618	1.525892841
		'평균의 차이'의 평균	4.24698021

평균의 차이가 가장 큰 변수는 Bare Nuclei 이고 가장 작은 변수는 Mitoses 입니다. 각 변수 값들은 1~10사이로 범위를 정했기 때문에 평균의 차이가 클래스를 구분 하는 데 의미 있는 요소가 될 수 있습니다. 따라서 '평균의 차이'의 평균인 약 4.24보다 큰 값이 클래스를 구분하는데 유용한 값이라고 판단을 할 수 있습니다.

유방암 판별 모형

지금까지 결과를 바탕으로 보면 유방암에 가장 영향을 미치는 요소는 Bare Nuclei으로 판단할 수 있습니다. 하지만 그래프를 보면 Bare Nuclei 뿐만 아니라 Uniformity Of Cell Size 와 Uniformity Of Cell Shape 도 의미가 있는 속성이라는 것을 예측할 수 있습니다. 위에서 뽑은 3가지 속성 Uniformity Of Cell Size, Uniformity Of Cell Shape, Bare Nuclei을 가지고 로지스틱 회귀 모델을 구성해 보았습니다.

Logistic Regression Model

로지스틱 회귀 모델을 작성해보니 아래와 같은 결과가 나왔습니다.

```
Call:
glm(formula = Nclass ~ Uniformity_of_Cell_Size + Uniformity_of_Cell_Shape +
    Bare_Nuclei, family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.8280  -0.1434  -0.1434   0.0326   2.2466

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              -6.2813     0.5289 -11.875  < 2e-16 ***
Uniformity_of_Cell_Size   0.5571     0.1549   3.597 0.000322 ***
Uniformity_of_Cell_Shape  0.6183     0.1715   3.605 0.000313 ***
Bare_Nuclei               0.5332     0.0770   6.925 4.36e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Class = -6.2813 + (0.5571) * Uniformity_of_Cell_Size + (0.6183) * Uniformity_of_Cell_Shape + (0.5332) * Bare_Nuclei

그리고 위 수식의 Accuracy 를 계산하기 위하여 Data Set 의 80%를 모델링에 사용하고, 20%를 모델을 검증하는데 사용하였습니다. 그 결과 아래와 같이 98.54%의 Accuracy 가 나왔습니다.

In [7]:
```
from sklearn.preprocessing import LabelEncoder
X= df.iloc[:,[2,3,6]]
y= df.iloc[:,10]

encoder = LabelEncoder()
y1 = encoder.fit_transform(y)
Y = pd.get_dummies(y1).values
```

In [8]:
```
from sklearn.model_selection import train_test_split
X_train,X_test, y_train, y_test= train_test_split(X,Y,test_size=0.2,random_state=1)
```

In [11]:
```
from keras.models import Sequential
from keras.layers import Dense
from keras.optimizers import Adam

model = Sequential()

model.add(Dense(64,input_shape=(3,),activation = "tanh"))
model.add(Dense(64,activation = "tanh"))
model.add(Dense(2,activation = "sigmoid"))

model.compile(loss='binary_crossentropy', optimizer = 'Adam', metrics=['accuracy'])
model.summary()
```

...

In [15]:
```
hist = model.fit(X_train,y_train,epochs =100, batch_size =32)
```

...

In [16]:
```
a=model.evaluate(X_test,y_test)
```

```
137/137 [==============================] - 0s 131us/step
```

In [17]:
```
print(a)
```

```
[0.03767997616507711, 0.9854014598540146]
```

저희가 추론한 수식이 정확한 지와 좀 더 높은 Accuracy를 가진 수식이 있는지 확인하기 위해 Stepwise를 해보았습니다 그 결과 다음과 같은 수식이 나왔습니다.

Class = -9.9773 + (0.5341) * Clump_Thickness + (0.3447) * Uniformity_of_Cell_Shape + (0.3422) * Marginal_Adhesion + (0.3880) * Bare_Nuclei + (0.4620) * Bland_Chromatin + (0.2259) * Normal_Nucleoli + (0.5306) * Mitoses

추가적으로 모든 속성을 적용하여 모델을 설계하였더니 다음과 같은 수식이 나왔습니다.

Class = -10.098215 + (0.535079) * Clump_Thickness + (-0.006199) * Uniformity_of_Cell_Size + (0.322113) * Uniformity_of_Cell_Shape + (0.330368) * Marginal_Adhesion + (0.096511) * Single_Epitheial_Cell_Size + (0.382786) * Bare_Nuclei + (0.447316) * Bland_Chromatin + (0.212834) * Normal_Nucleoli + (0.534258) * Mitoses

다음의 표는 처음 예측한 모델, Stepwise의 결과 모델 그리고 모든 속성을 다 적용한 모델의 Accuracy 입니다.

모델	Accuracy
그래프를 통해 예측한 모델	0.985
Stepwise를 실행한 모델	0.9452
모든 속성을 적용한 모델	0.9416

각 모델 간 Accuracy의 비교를 해 보니 모델 간에 큰 차이가 존재하지 않다는 것을 알게 되었습니다.

본 스케치에서는 R을 사용하여 독립변수와 종속변수간의 상관관계 및 구체적인 수식 확인을 진행하였으며, 예측 결과에 대한 Accuracy 확인에는 python의 Deep learning library인 Keras와Tensorflow를 사용하였습니다.

본 스케치를 진행한 결과, 계획 단계에서 생각했던 의미 있는 결과를 도달하지 못하였으며 그 원인으로 샘플 데이터 수의 부족이라고 생각하였습니다. 또한 속성들이 모두 의미 있는 값들이기 때문에 서로 다른 모델간의 Accuracy의 차이가 거의 없다고 생각하게 되었습니다.