PART 3. 데이터 분석 - 5장. 정형 데이터 마이닝

데이터 준전문가

ADSP, Advanced Data Analytics semi-Professional

류영표 강사
ryp1662@gmail.com

강사 소개

류영표

Youngpyo Ryu

동국대학교 수학과/응용수학 석사수료

現 Upstage AI X 네이버 부스트 캠프 AI tech 1~4기 멘토

前 Innovation on Quantum & CT(IQCT) 이사

前 한국파스퇴르연구소 Image Mining 인턴(Deep learning)

前 ㈜셈웨어(수학컨텐츠, 데이터 분석 개발 및 연구인턴)

강의 경력

- 현대자동차 연구원 강의 (인공지능/머신러닝/딥러닝/강화학습)
- ㈜ 모두의연구소 Aiffel 1기 퍼실리테이터(인공지능 교육)
- 인공지능 자연어처리(NLP) 기업데이터 분석 전문가 양성과정 멘토
- 공공데이터 청년 인턴 / SW공개개발자대회 멘토
- 고려대학교 선도대학 소속 30명 딥러닝 집중 강의
- 이젠 종로 아카데미(파이썬, ADSP 강사) / 강남 : ADSP
- 최적화된 도구(R/파이썬)을 활용한 애널리스트 양성과정(국비과정) 강사
- 한화, 하나금융사 교육
- 인공지능 신뢰성 확보를 위한 실무 전문가 자문위원
- 보건 · 바이오 AI활용 S/W개발 및 응용전문가 양성과정 강사
- Upstage AI X KT 융합기술원 기업교육 모델최적화 담당 조교

주요 프로젝트 및 기타사항

- 개인 맞춤형 당뇨병 예방·관리 인공지능 시스템 개발 및 고도화(안정화)
- 폐플라스틱 이미지 객체 검출 경진대회 3위
- 인공지능(AI)기반 데이터 사이언티스트 전문가 양성과정 1기 수료
- 제 1회 산업 수학 스터디 그룹 (질병에 영향을 미치는 유전자 정보 분석)
- 제 4,5회 산업 수학 스터디 그룹 (피부암, 유방암 분류)
- 빅데이터 여름학교 참석 (혼잡도를 최소화하는 새로운 노선 건설 위치의 최적화 문제)

데이터 마이닝

- 대용량 데이터에서 의미있는 패턴을 파악하거나 예측하여 의사결정에 활용하는 방법
- 통계분석과 차이점

  : 통계분석은 가설이나 가정에 따른 분석이나 검증을 하지만, 데이터마이닝은 다양한 수리 알고리즘을 이용해 데이터베이스의 데이터로부터 의미 있는 정보를 찾아내는 방법을 통칭.

정보를 찾는 방법론에 따른 종류	분석대상, 활용목적, 표현방법에 따른 분류
• 인공지능(Artificial Intelligence) • 의사결정 나무(Decision Tree) • k-평균군집화(K-means Clustering) • 연관분석(Association Rule) • 회귀 분석(Regression) • 로짓분석(Logit Analysis) • 최근접이웃(Nearest Neighborhood)	• 시각화 분석(Visualization Analysis) • 분류(Classification) • 군집화(Clustering) • 포캐스팅(Forecasting)

데이터 마이닝의 분석 방법

Supervised Data Prediction(지도학습)	UnSupervised Data Prediction(비지도학습)
• 회귀 분석(Regression Analysis) • 로지스틱 회귀 분석(Logistic Regression Analysis) • 일반화 선형 모델(GLM, Generalized Linear model) • 의사결정나무(Decision Tree) (분류) • 인공신경망(Artificial Neural Network) (분류) • 사례기반 추론(Case-based Reasoning) • 최근접 이웃법(K-Nearest Neighborhood)	• OLAP(On-line Analytical Processing) • 연관성 규칙발견(Association Rule Discovery, Market Basket) • 군집 분석(K-means clustering) • SOM(Self Organizing Map)

분석 목적에 따른 작업 유형과 기법

목적	작업유형	설명	사용기법
예측 Predictive Modeling	분류 규칙 Classification	• 가장 많이 사용하는 작업 • 과거의 데이터로부터 고객특성을 찾아내어 분류모형을 만들어 이를 토대로 새로운 레코드의 결과값을 예측 • 목표마케팅 및 고객 신용평가 모형에 활용	회귀분석 판별분석 인공신경망 의사결정나무
설명 Descriptive Modeling	연관규칙 Association	• 데이터 안에 존재하는 항목간의 종속관계를 찾아내는 작업 • 제품이나 서비스의 교차판매, 매장진열, 첨부우편, 사기적발 등 다양한 분야에 활용	동시발생 매트릭스
	연속규칙 Sequence	• 연관규칙에 시간관련 정보가 포함된 형태 • 고객이 구매이력 속성이 반드시 필요하며 목표 마켓팅이나 일대일 마케팅에 활용	동시발생 매트릭스
	군집화 Clustering	• 고객 레코드들을 유사한 특성을 지닌 몇 개의 소그룹으로 분할작업의 특성이 분류규칙과 유사하나 분석대상 데이터에 결과값이 없으며, 판촉활동이나 이벤트 대상을 선정하는 데 활용	K-means Clustering SOM

- **동시발생 매트릭스**는 거래(사건) 속에 포함된 품목(항목)간의 연관관계를 발견하고자 할 때 사용하는 data mining 기법이다.

데이터 마이닝 추진단계

1단계 : 목적설정

· 무엇을 왜 하는지 정확한 목적 설정, 전문가가 참여해 목적에 따라 사용할 모델과 필요한 데이터를 정의

2단계 : 데이터 준비

· 고객정보, 거래정보, 상품 마스터정보, 웹로그 데이터, 소셜 네트워크 데이터 등 다양한 데이터를 활용한다.

· IT 부서와 사전에 협의하고 일정을 조율하여 데이터 접근 부하에 유의, 필요 시 다른 서버에 저장하여 운영에 지장이 없도록 데이터 준비

· 데이터 정제를 통해 데이터의 품질을 보장하고 필요시 데이터를 보강하여 충분한 양의 데이터 확보.

3단계 : 가공

· 모델링 목적에 따라 목적변수를 정의

· 필요한 데이터를 데이터마이닝 소프트웨어에 적용할 수 있는 형식으로 가공

4단계 : 기법, 적용

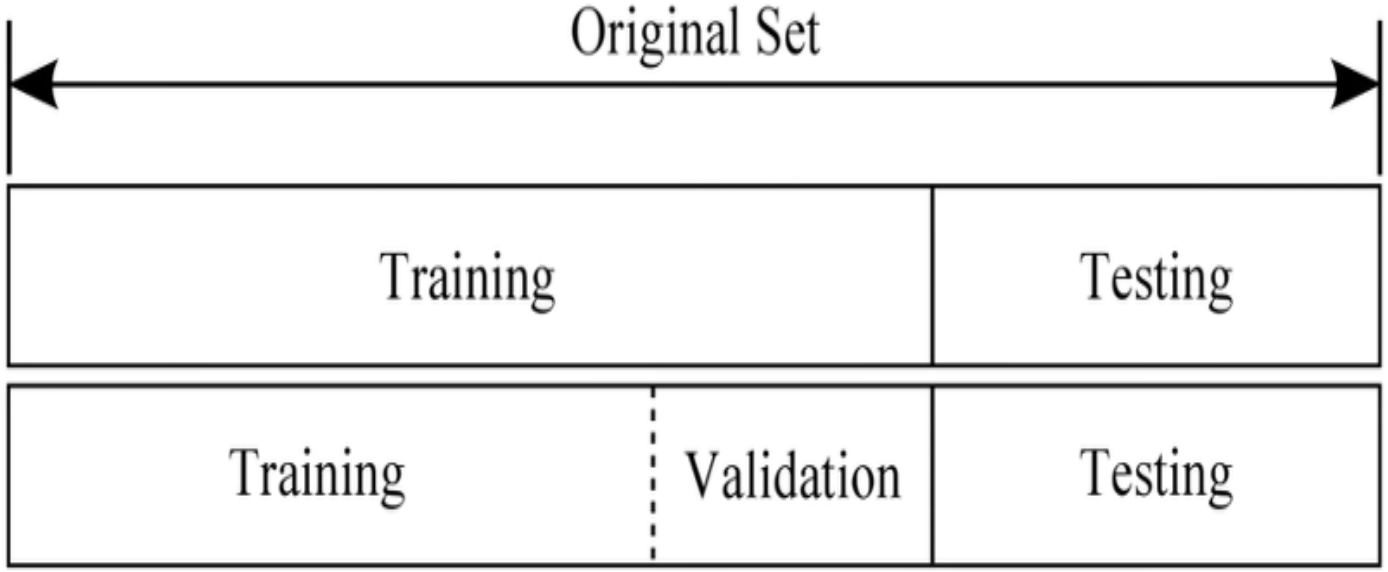
· 1단계에서 명확한 목적에 맞게 데이터마이닝 기법을 적용하여 정보를 추출

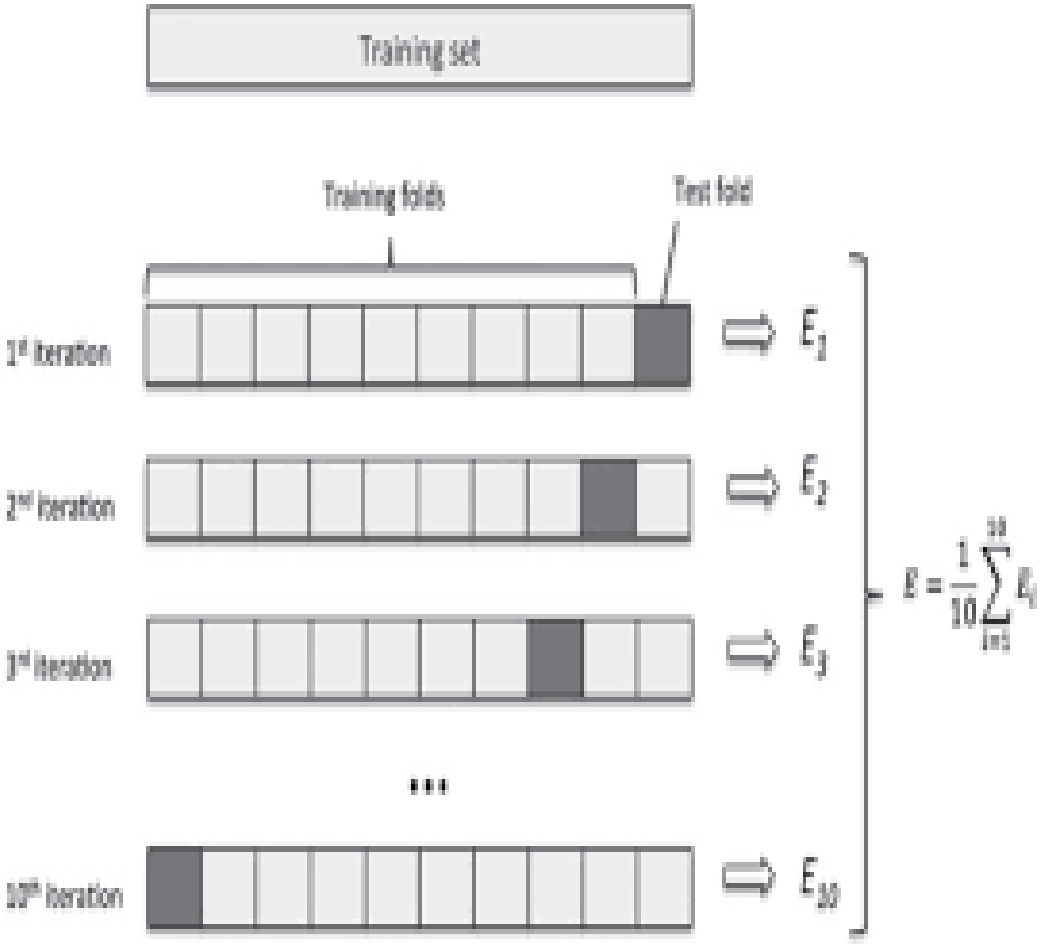
5단계 : 검증

· 추출된 정보를 검증, 테스트 데이터와 과거 데이터를 활용하여 최적의 모델을 선정

· 검증이 완료되면 IT부서와 협의해 상시 데이터마이닝 결과를 업무에 적용하고 보고서를 작성하여 추가 수익 투자대비성과(ROI)등으로 기대효과를 전파.

데이터마이닝을 위한 데이터 분할





오분류표(Confusion Matrix)

		Actual Values	
		Positive	Negative
Predicted Values	Positive	TP	FP
	Negative	FN	TN

- TP(True Positive) : 실제 positive값을 positive로 올바르게 예측한 경우
- TN(True Negative) : 실제 negative값을 negative로 올바르게 예측한 경우
- FP(False Positive or Type I error(1종 오류)) : 실제 negative값을 positive로 잘못 예측한 경우
- FN(False Negative or Type II error(2종 오류)) : 실제 positive값이 negative로 잘못 예측한 경우

정확도(Accuracy)

		Actual Values	
		Positive	Negative
Predicted Values	Positive	TP	FP
	Negative	FN	TN

- Accuracy는 정확도로 전체 예측한 것 중에 올바른 예측을 얼마나 했는지를 지표로 구하는 것.

$$\frac{TP + TN}{TP + TN + FP + FN}$$

- Accuracy값이 높을수록 예측 정확도가 높아질 수 있음.
- Error Rate = 1−Accuracy
- 편향의 함정이 있음.

정밀도(Precision)

		Actual Values	
		Positive	Negative
Predicted Values	Positive	TP	FP
	Negative	FN	TN

- Precision은 정밀도로 Positive로 예측한 것 중 실제로 맞춘 비율.

$$\frac{TP}{TP + FP}$$

- Positive로 예측한 것 중 실제로 Positive가 얼마나 되는지를 보여주는 지표
- Positive predictive value

민감도(Sensitivity, recall)

		Actual Values	
		Positive	Negative
Predicted Values	Positive	TP	FP
	Negative	FN	TN

- True Positive Rate로 실제 Positive를 얼마나 잘 예측했는지를 나타내는 지표.

$$\frac{TP}{TP + FN}$$

- 실제 Positive 중 Positive로 예측한 것이 얼마나 되는지를 보여주는 지표.

특이도(Specificity)

		Actual Values	
		Positive	Negative
Predicted Values	Positive	TP	FP
	Negative	FN	TN

- Specificity는 특이도로 실제 Negative를 얼마나 잘 예측했는지를 나타내는 지표.

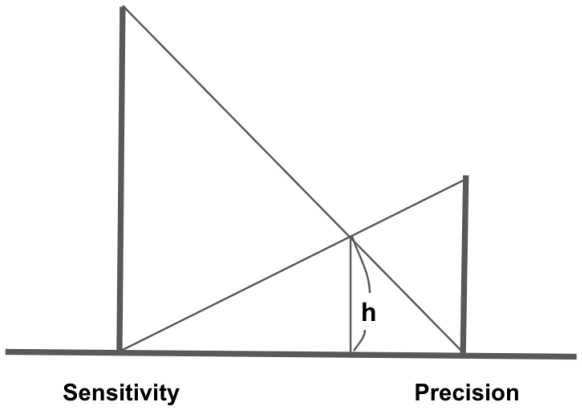
$$\frac{TN}{FP + TN}$$

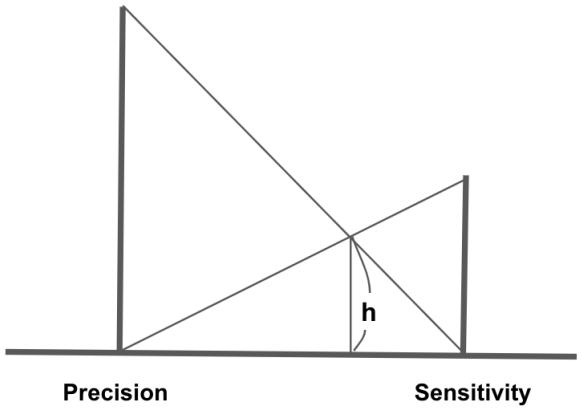
- 특이도를 통해서 False Positive Rate를 구할 수 있음.

  False Positive Rate = 1 − Specificity

F1-Score

- 보통 불균형 분류문제에서 평가척도 주로 사용됨.
- 데이터가 불균형한 상태에서 Accuracy로 성능을 평가하기엔 데이터 편향성이 너무 크게 나타나 올바르게 측정하기 힘들 때 사용.
- Sensitivity와 Precision을 이용한 조화평균

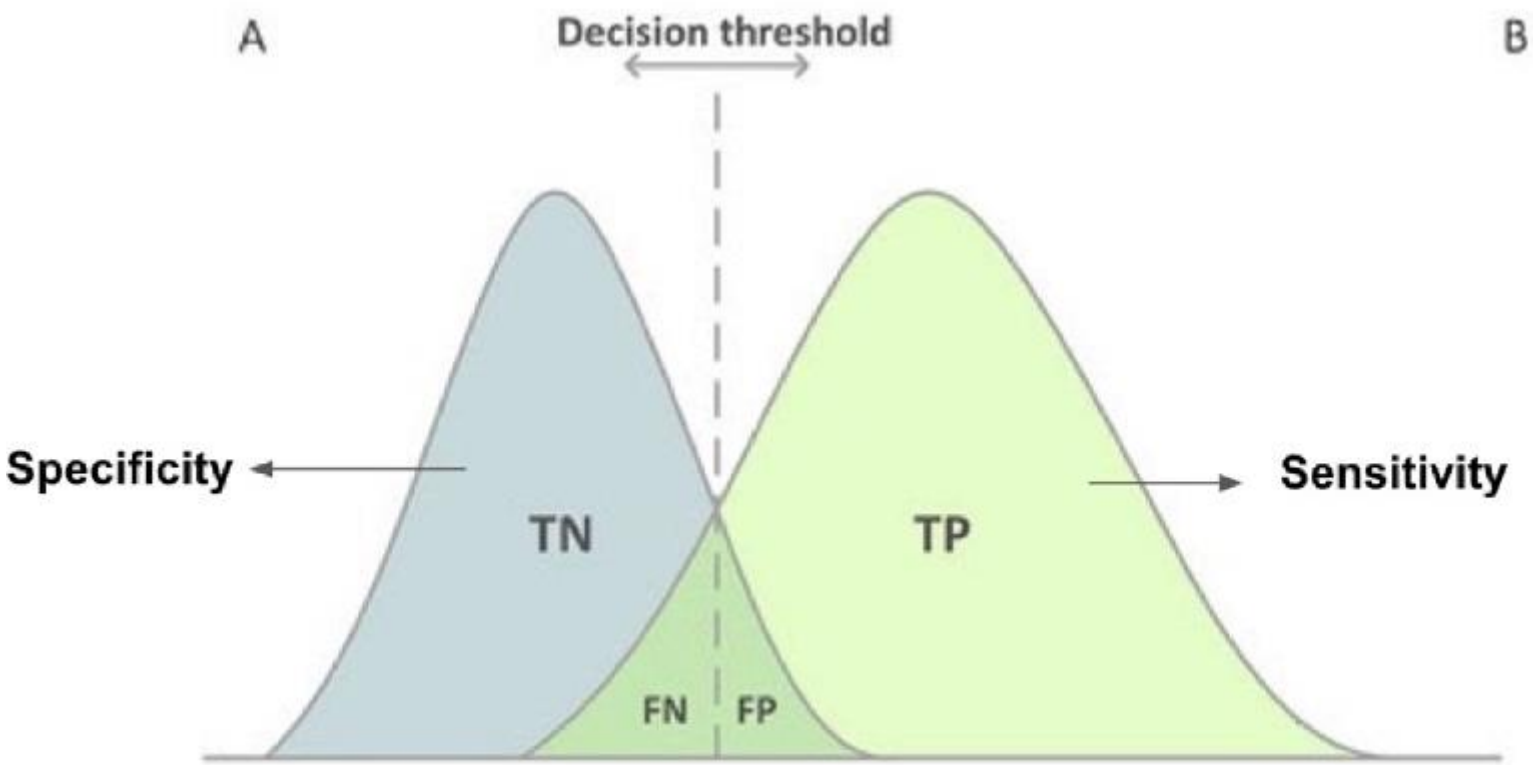




$$2 * \frac{\text{Sensitivity} * \text{Precision}}{\text{Sensitivity} + \text{Precision}}$$

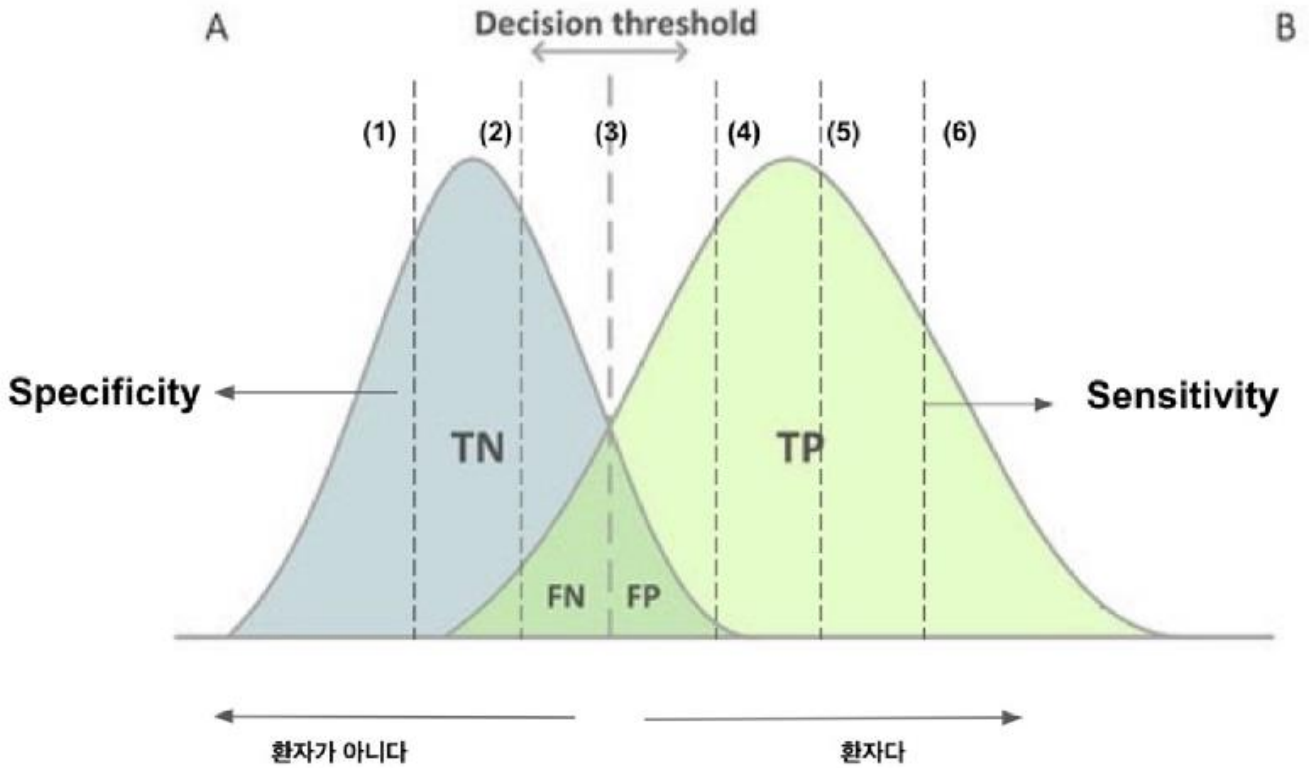
ROC(Receiver Operating Characteristic)

- FPR(False Positive Rate)과 TPR(True Positive Rate)을 각각 x, y축으로 놓은 그래프.
- TPR(True Positive Rate)-〉 실제 Positive를 얼마나 잘 예측했는지를 나타내는 지표(Sensitivity)
- FPR(False Positive Rate) -〉 1 - Specificity(실제 Negative를 얼마나 잘 예측했는지를 나타내는 지표)



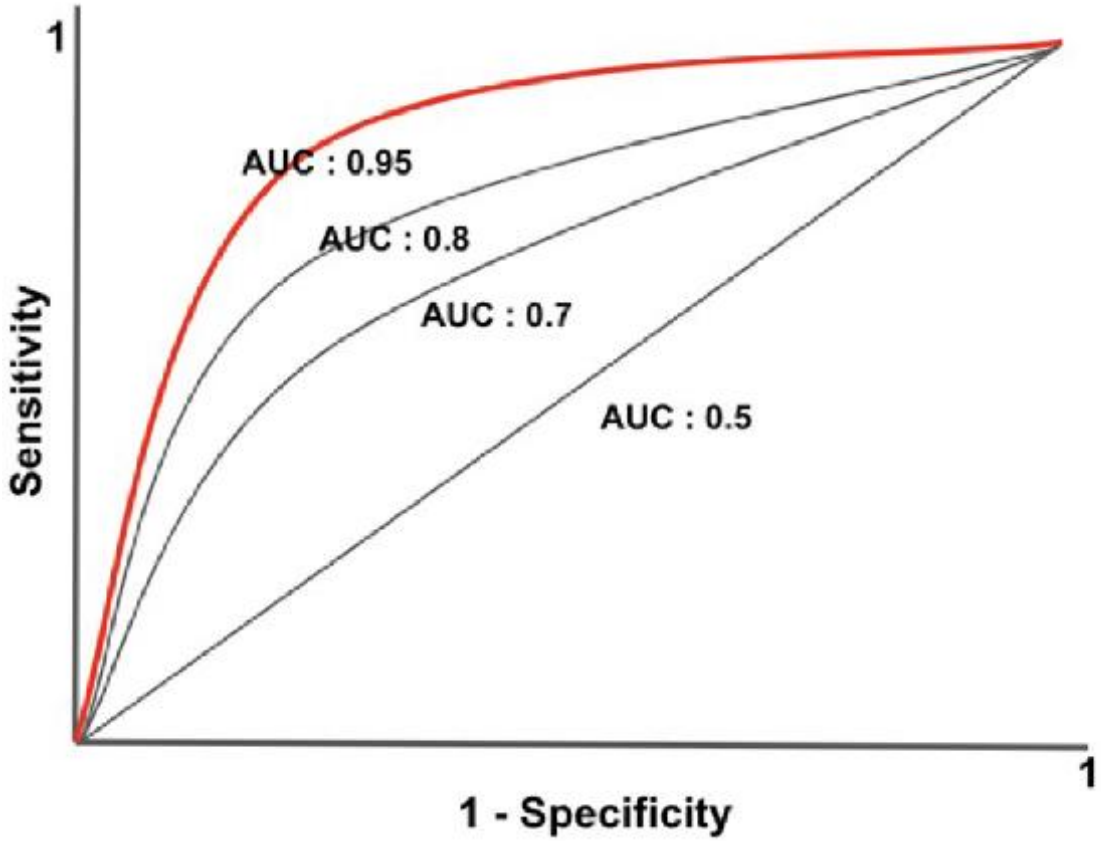
ROC(Receiver Operating Characteristic)

- FPR(False Positive Rate)과 TPR(True Positive Rate)을 각각 x, y축으로 놓은 그래프.
- TPR(True Positive Rate)-〉 실제 Positive를 얼마나 잘 예측했는지를 나타내는 지표(Sensitivity)
- FPR(False Positive Rate) -〉 1 - Specificity(실제 Negative를 얼마나 잘 예측했는지를 나타내는 지표)



AUC(Area Under the Curve)

- FPR(False Positive Rate)과 TPR(True Positive Rate)을 각각 x, y축으로 놓은 그래프.
- TPR(True Positive Rate)-〉 실제 Positive를 얼마나 잘 예측했는지를 나타내는 지표(Sensitivity)
- FPR(False Positive Rate) -〉 1 - Specificity(실제 Negative를 얼마나 잘 예측했는지를 나타내는 지표)



ROC 패키지 R코드 실습

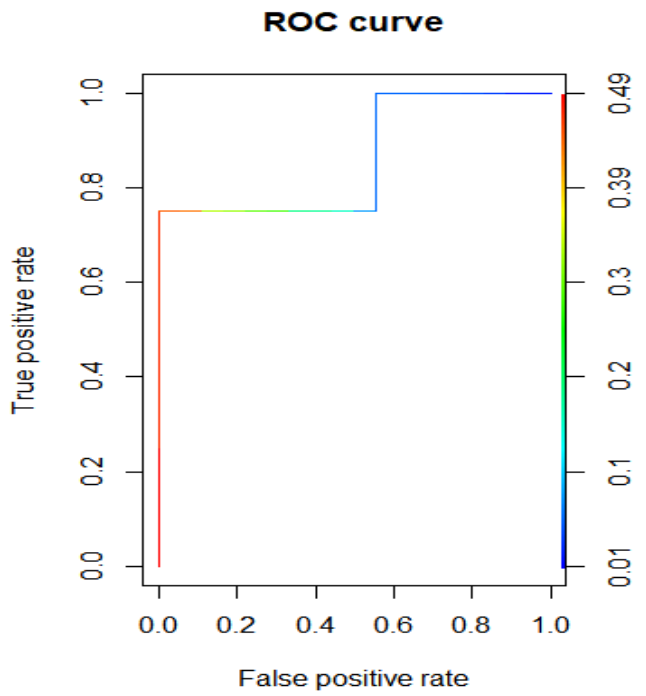
- Binary Classification만 지원 가능

```
library(rpart)
library(party)
library(ROCR)

x <- kyphosis[sample(1:nrow(kyphosis), nrow(kyphosis), replace=F),]
x.train <- kyphosis[1:floor(nrow(x)*0.75),]
x.evaluate <- kyphosis[floor(nrow(x)*0.75):nrow(x),]
x.model <- cforest(Kyphosis~Age+Number+Start, data=x.train)
x.evaluate$prediction <- predict(x.model, newdata=x.evaluate)
x.evaluate$correct <- x.evaluate$prediction == x.evaluate$Kyphosis
print(paste("% of predicted classification correct", mean(x.evaluate$correct)))
x.evaluate$probabilities <- 1- unlist(treeresponse(x.model, newdata=x.evaluate), use.name=F)[seq(1,nrow(x.evaluate)*2,2)]
```

```
[1] "% of predicted classification correct 0.818181818181818"
```

```
pred <- prediction(x.evaluate$probabilities, x.evaluate$Kyphosis)
perf <- performance(pred, "tpr", "fpr")
plot(perf, main="ROC curve", colorize=T)
```



오분류표 예제

		Actual Values	
		Positive	Negative
Predicted Values	Positive	30	40
	Negative	40	60

- 정확도?
- 정밀도?
- 재현율(민감도)
- 특이도?
- FP Rate?
- F1 은?

이익 도표(lift)

이익도표는 분류모형의 성능을 평가하기 위한 척도로, 분류된 관측치에 대해 얼마나 예측이 잘 이루어졌는지를 나타내기 위해 임의로 나눈 각 등급별로 반응검출율, 반응률, 리프트 등의 정보를 산출하여 나타내는 도표.

각 등급은 예측확률에 따라 매겨진 순위이기 때문에. 상위 등급에서는 더 높은 반응률을 보이고 하위등급으로 갈수록 Lift가 빠른 속도로 감소해야 좋은 모형이라고 평가.

Ex) 2000명의 전체 고객 중 381명이 상품이 구매만 경우에 대해 이익도표를 만드는 과정

1) 데이터셋의 각 관측치에 대한 예측 확률을 내림차순으로 정렬

2) 데이터를 10개의 구간으로 나눈 다음 각 구간 반응률(% response)을 산출

3) 기본향상도(baseline lift)에 비해 반응률이 몇 배나 높은지를 계산 = 향상도(Lift)

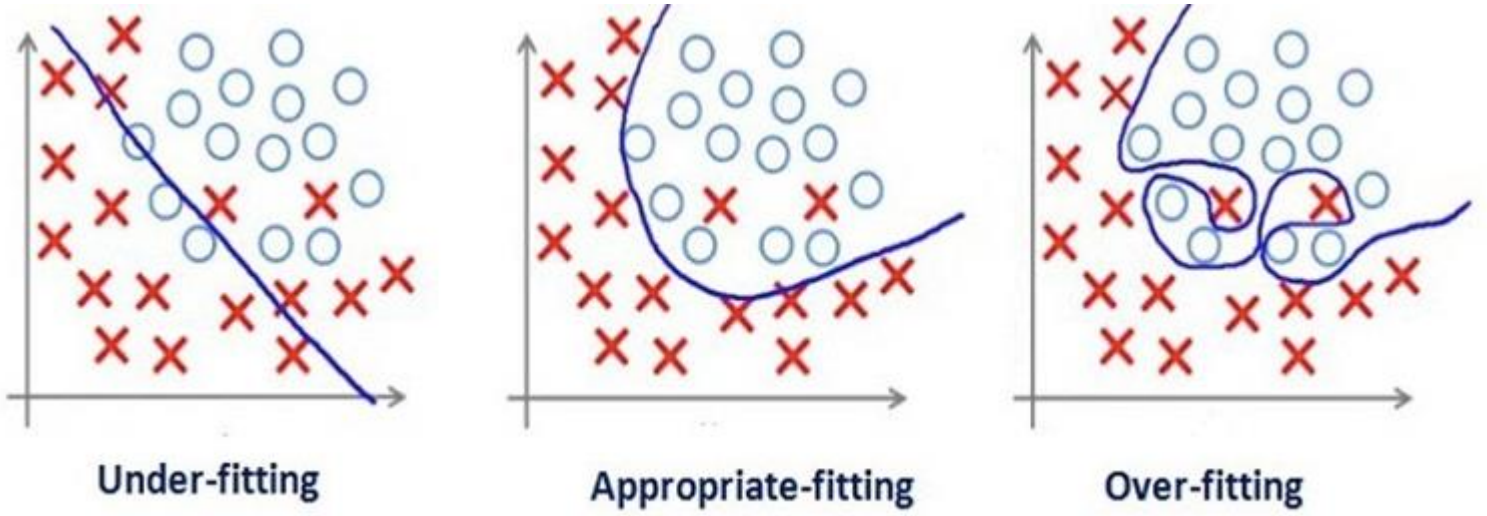
[참고] 과적합, 과대적합, 과소적합

1. 과적합·과대적합(Overfitting)

   : 모형이 학습용 데이터를 과하게 학습하여, 학습 데이터에 대해서는 높은 정확도를 나타내지만 테스트 데이터 혹은 다른 데이터에 적용할 때는 성능이 떨어지는 현상, 훈련데이터에 최적화 되어있기 때문에 테스트 데이터의 작은 변화에 민감하게 반응.

2. 과소적합(Underfitting)

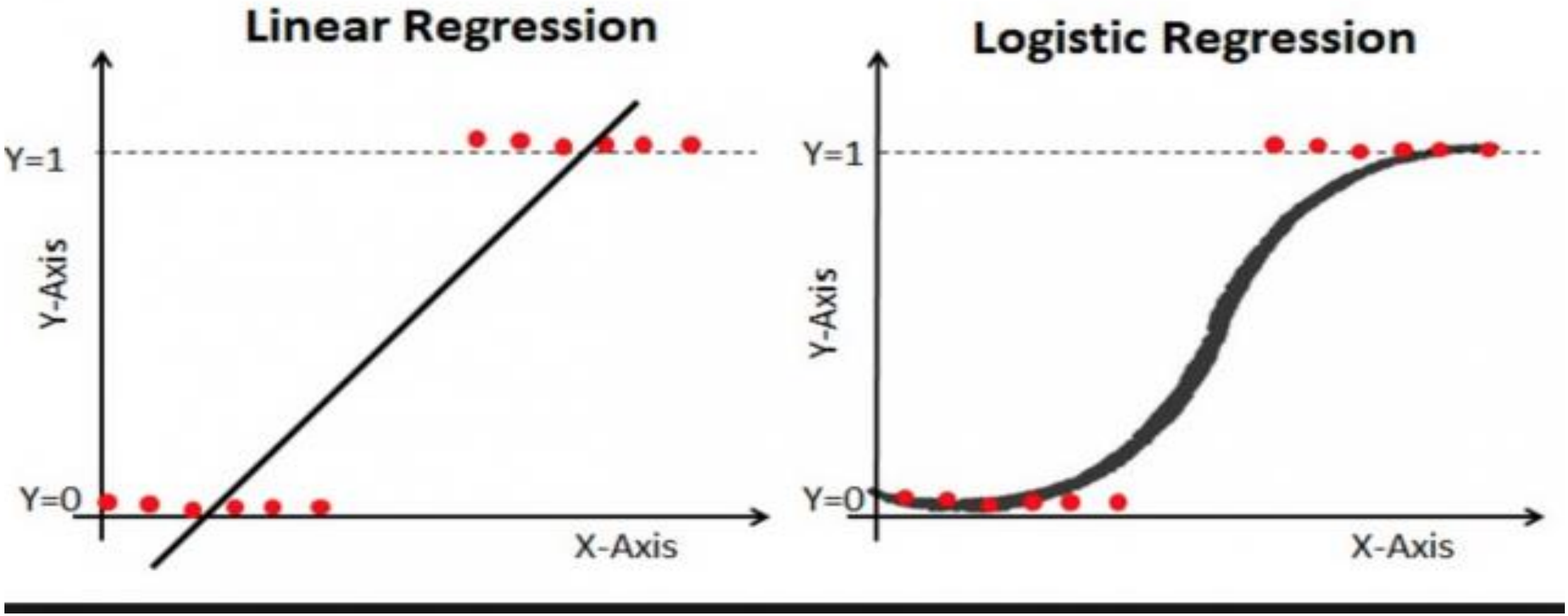
   : 모델이 너무 단순하여 데이터 속에 내제되어 있는 패턴이나 규칙을 제대로 학습하지 못한 경우



Logistic Regression

- 영국의 통계학자 D.R.Cox가 1958년에 제안한 확률 모델로서 독립변수의 선형 결합을 이용하여 사건의 발생 가능성을 예측하는데 사용되는 통계기법.
- 선형 회귀분석과는 다르게 종속변수가 범주형 데이터를 대상으로 하며, 입력 데이터가 주어졌을 때 해당 데이터의 결과가 특정 분류로 나뉘기 때문에 일종의 분류(Classification) 기법으로도 볼 수 있음.
- 종속 변수가 이항변수의 경우(ex) 성공/실패, 업/다운, 생존/죽음, Yes/No 등등)
- OLS로 해석 불가
- 종속변수가 0,1인데 logit을 하면 연속변수처럼 바꾸는 효과가 생겨서 OLS로 해석이 가능하고 확률개념이 생겨서 해석이 용이함.

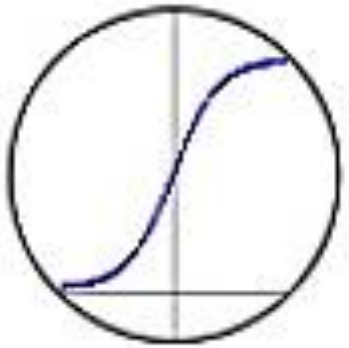
Logistic Regression



출 처 : https://medium.com/@ODSC/logistic-regression-with-python-ede39f8573c7

Logistic Regression

Logistic 함수 및 미분특성



$$f(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x}$$

$$\frac{d}{dx} f(x) = f(x)(1 - f(x))$$

Logistic Regression

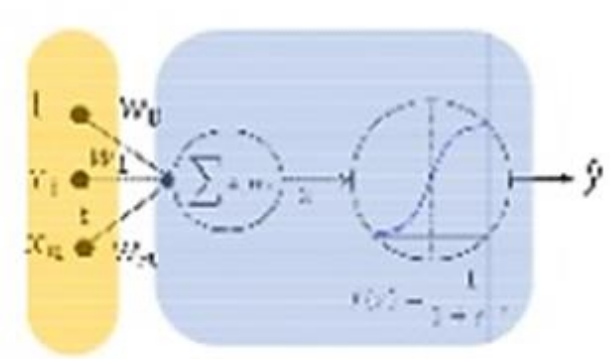
Logistic 함수 및 미분특성

$$\frac{d}{dx}f(x) = \frac{d}{dx}\frac{1}{1+e^{-x}} = \frac{1}{(1+e^{-x})^2}(e^{-x})$$

$$= \frac{1}{(1+e^{-x})^2}\left(1 - \frac{1}{(1+e^{-x})}\right)$$

$$= f(x)f(1-f(x))$$

Logistic Regression



$$\hat{y} = f(z) = \frac{1}{1+e^{-z}}$$

$$z = \sum w_i x_i$$

$$\frac{\partial}{\partial z} f(z) = f(z)\big(1 - f(z)\big)$$

$$\frac{\partial E}{\partial w_i} = \frac{\partial}{\partial w_i} \frac{1}{2} \sum_d (y_d - \hat{y}_d)^2$$

$$= \frac{1}{2} \sum_d \frac{\partial}{\partial w_i} (y_d - \hat{y}_d)^2$$

$$= \frac{1}{2} \sum_d 2(y_d - \hat{y}_d) \frac{\partial}{\partial w_i} (y_d - \hat{y}_d)$$

$$= \sum_d (y_d - \hat{y}_d) \frac{\partial}{\partial w_i} (-\hat{y}_d)$$

$$= -\sum_d (y_d - \hat{y}_d) \frac{\partial \hat{y}_d}{\partial z} \frac{\partial z}{\partial w_i}$$

$$= -\sum_d (y_d - \hat{y}_d) \frac{\partial f(z)}{\partial z} \frac{\partial \left(w_0 + w_1 x_{1,d} + \cdots + w_n x_{n,d}\right)}{\partial w_i}$$

$$= -\sum_d (y_d - \hat{y}_d) \hat{y}_d (1 - \hat{y}_d) x_{i,d}$$

Logistic Regression

· 오즈(Odds)에 대해 알아보자.

· 오즈는 원래 경마장에서 도박 지불률을 정할 때 사용

· 경마장의 예제 : 3마리 말이 출전. 3마리 말의 승리할 확률이 다음과 같음

- A말은 50%의 확률로 승리 : B말은 30%의 확률로 승리 : C말은 20%의 확률로 승리
- 만약, 참가자에게 1000원의 마권을 사게 하고, 승자를 맞춘 사람에게 3000원씩 지불한다면
- A말의 마권을 살 경우 기대값 = 3000원X 50% = 1500원
- B말의 마권을 살 경우 기대값 = 3000원X 30% = 900원
- C말의 마권을 살 경우 기대값 = 3000원X 20% = 600원

· 그러므로, 바보가 아닌 이상 A말의 마권을 사는 것이 비용 1000원을 초과하는 유일한 선택

· 주인의 입장 : 50%의 확률로 A말이 이기면, 판매된 마권의 개수 X 2000원 만큼 손해,

· 그러므로 주인은 망함.

Logistic Regression

· 오즈(Odds)에 대해 알아보자.

· 상금지급 기준 변경 1안

· 경마장의 예제 : 전과 동일

- A말은 50%의 확률로 승리; B말은 30%의 확률로 승리; C말은 20%의 확률로 승리
- 만약, 참가자에게 1000원의 마권을 사게 하고, 예상 승률의 역수배로 상금 지불
- A말의 마권을 살 경우 기댓값 = 1000원 X(1/50%)X50% = 1000원
- B말의 마권을 살 경우 기댓값 = 1000원 X(1/30%)X30% = 1000원
- C말의 마권을 살 경우 기댓값 = 1000원 X(1/20%)X20% = 1000원
- 어느 쪽에 걸더라도 기대값 상금이 같아져 패자의 손실과 승자의 이익이 상쇄됨.

Logistic Regression

· 오즈(Odds)에 대해 알아보자.

· 상금지급 기준 변경 2안

· 경마장의 예제 : 전과 동일

- A말은 50%의 확률로 승리; B말은 30%의 확률로 승리; C말은 20%의 확률로 승리
- 만약, 참가자에게 1000원의 마권을 사게 하고, 오즈의 역수배로 상금 지불
- 오즈(Odds) : 확률 ÷ (1−확률)을 의미
- A의 오즈 = 50% ÷ (1−50%) = 1; 기댓값 = 1000원 X 1배 X 50% = 500원
- B의 오즈 = 30% ÷ (1−30%) = 0.43; 기댓값 = 1000원 X2.33배X30% ≈ 700원
- C의 오즈 = 20% ÷ (1−20%) = 0.25; 기댓값 = 1000원 X4X20% = 800원

$$\text{ODDS} = \frac{\text{Probability of winning}}{\text{Probability of losing}} = \frac{p}{1-p}$$

Where p = probability of winning (event occurring)

Logistic Regression

· 오즈 비 (Odds ratio)

· 오즈비의 예제 : 극단적인 경우의 이항변수

· 희귀질병의 발병여부

· 종속변수 : 희귀질병의 발병여부

· 독립변수 : 거주지역구분(도시지역 or 비도시지역)

	발병 Yes	발병 NO	합계
도심지역	1명	1,999명	2,000명
비도심지역	1명	7,999명	8,000명
합 계	2명	9,998명	10,000명

Logistic Regression

· 오즈 비 (Odds ratio)

· 희귀질병의 발병여부

	발병 Yes	발병 NO	합계
도심지역	1명	1,999명	2,000명
비도심지역	1명	7,999명	8,000명
합 계	2명	9,998명	10,000명

· 도심지역 주민의 발병 오즈(odds) $\frac{\frac{1}{2000}}{1-\frac{1}{2000}} = \frac{1}{1999}$

· 비도심지역 주민의 발병 오즈(odds) $\frac{\frac{1}{8000}}{1-\frac{1}{8000}} = \frac{1}{7999}$

· 비도심지역 대비 도심 지역 발병 오즈비(odds ratio) $\frac{\frac{1}{1999}}{\frac{1}{7999}} \approx 4.0015$

Logistic Regression

· 오즈 비 (Odds ratio)

· 왜 이런 복잡한 걸 쓸까?

· 도심지역 주민의 발병 오즈(odds) : $\frac{\frac{1}{2000}}{1-\frac{1}{2000}} = \frac{1}{1999}$

· 비도심지역 주민의 발병 오즈(odds) : $\frac{\frac{1}{8000}}{1-\frac{1}{8000}} = \frac{1}{7999}$

· 비도심지역 대비 도심 지역 발병 오즈비(odds ratio) : $\frac{\frac{1}{1999}}{\frac{1}{7999}} \approx 4.0015$

· 도심지역 발병 확률 = $\frac{1}{2000}$: 비도심지역 발병 확률 = $\frac{1}{8000}$: 거의 차이 없음

Logistic Regression

· 오즈 비 (Odds ratio)

· 왜 이런 복잡한 걸 쓸까?

- 오즈(odds)와 단순 확률(P)는 거의 같음
- 오즈를 오즈로 나누면 비교가 가능

- 비도심지역 대비 도심지역 발병 오즈비(Odds ratio) = $\frac{\frac{1}{1999}}{\frac{1}{7999}} \approx 4.0015$
- 비도심지역에 1명의 희귀질병 환자가 발병할 경우, 도심지역에는 4명 발생

- 여기에 log를 붙이면 더 활용범위가 높아진다.
- 0〈p〈1(여기서 p는 단순확률) & 0〈1−p 〈1
- p가 '0'에 가까울 수록 오즈비(odds ratio)는 '0'을 향해 접근
- p가 '1'에 가까울 수록 오즈비(odds ratio)는 '무한대'을 향해 접근
- 0〈오즈비(odds ration)〈 무한대 −〉−무한대 〈log(오즈비) = 로짓 〈+무한대

Logistic Regression

· 로짓(logit) = log(오즈비)

·종속변수가 0과 1인 경우 : 오즈비를 구하면 최소값은 0(모든값이 0)

· 로그 밑수를 Exponential 로 할 경우 log_e 0은 마이너스 무한대가 됨.

· 왜냐하면, 2.718의 수를 계속 곱하더라도 2.718보다 작은 수가 되지는 않음

· 그러나 여러 차례 계속 나누게 되면 1보다 점점 작은 값이 되어 0을 향해 움직임

· 나눈다는 것은 마이너스 제곱의 의미이므로 마이너스 무한제곱이 되면 거의 0

· 그래서 log 0 = $-\infty$

· 종속변수가 0과 1인 경우 : 오즈비를 구하면 최대값은 1(모든 값이 1)

·오즈비는 $1 \div (1-1)$이라 계산이 불가능 하지만 분모를 0에 가까운 0이 아닌 가장 작은 수라 가정하면, 1을 0에 가까운 가장 작은 수로 나누었으므로, 무한대가 됨.

· 이때 $log_e \infty = \infty$

Logistic Regression

· 결론)

– 로짓의 특징

$-\infty <$ 로짓(logit) = log(오즈비) $< +\infty$

– 이렇게 되면 우리의 종속변수는 0과 1로만 이루어진 문제에서 음의 무한대에서 양의 무한대까지

일반 연속변수처럼 바뀜

– 그러므로, 우리는 종속변수를 로짓으로 바꿔서 사용하면 우리가 아는 OLS로 회귀분석을 할 수 있음.

– 단점 : 해석방법이 기존의 OLS와 다소 다르다.

연습문제

4. 교차분석은 2개 이상의 변수를 결합하여 자료의 빈도를 살펴보는 기법이다. 다음 중 교차분석에 대한 설명으로 부적절한 것은 무엇인가?

① 범주의 관찰도수에 비교될 수 있는 기대도수를 기대한다.

② 교차분석은 두 문항 모두 범주형 변수가 아니어도 사용할 수 있으며, 두 변수 간 관계를 보기 위해 실시한다.

③ 교차분석은 교차표를 작성하여 교차빈도를 집계할 뿐 아니라 두 변수들 간의 독립성 검정을 할 수 있다.

④ 기대빈도가 5미만인 셀의 비율이 20%를 넘으면 카이제곱분포에 근사하지 않으며, 이런 경우 표본의 크기를 늘리거나 변수의 수준을 합쳐 셀의 수를 줄이는 방법 등을 사용한다.

연습문제

7. 아래 주성분분석의 결과에서 두 개의 주성분을 사용할 때 설명 가능한 전체 분산의 비율은?

```
> model<-princomp(Car)
> summary(model)
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
Standard deviation        1.503     1.075     0.840     0.752     0.555
Proportion of Variance    0.453     0.231     0.141     0.113     0.061
Cumulative Proportion     0.453     0.684     0.825     0.938     1.000
```


Thank you.
ADSP / 류영표 강사
ryp1662@gmail.com