Chapter 05. 분류- 은지

- 지도 학습
- 이진분류나 다중분류로 예측하는 것이 목표
  - 이진분류 : 데이터가 1인지, 0인지
  - 다중분류 : 여러 카테고리 중 어디에 속할지
- 일반적인 접근 방법
  1. 컷오프 확률 정하기 (threshold)
  2. 레코드가 관심 클래스에 속할 확률 추정
  3. 그 확률이 컷오프 확률 이상이면, 관심 클래스에 레코드를 할당
- 범주 항목이 두가지 이상이라면
  → 조건부 확률을 사용해 여러개의 이진 문제로 돌려서 생각해보기

  예) Y=0, Y=1, Y=2로 분류 시 → Y=0 or Y>0 → Y>0이라면, Y=1 or Y=2로 예측

▼ 5.1 나이브 베이즈

참고 : 나이브 베이즈 분류기

> ✏️ 용어 정리
> - 조건부 확률 : 어떤 사건($Y = i$)이 주어졌을 때, 해당 사건($X = i$) 을 관찰할 확률 $P(X_i|Y_i)$
> - 사후확률 : 예측 정보를 통합한 후 결과의 확률

- 베이즈 정리

  $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$

  → $P(A|B)$를 구할 수 있으면, 위 식을 통해 $P(B|A)$ 계산 가능

  대부분의 사회 통계는 전수조사가 불가능해 기존 사건들의 확률을 알지 못하기 때문에 베이즈 정리는 쓸모없는 것이 되어버리는 한계가 있었으나, ***빅데이터를 통해 기존 사건들의 확률을 대략적으로 알 수 있게 되면서 베이즈 정리 활용이 필수적이게 됨*** *(나무위키:베이즈정리)*
- 완전한/정확한 베이즈 분류 방법
  1. 예측변수의 값이 동일한 모든 레코드를 찾는다.
  2. 해당 레코드들이 가장 많이 속한(즉, 가능성이 가장 많은) 클래스를 정한다.
  3. 새 레코드에 해당 클래스를 지정한다.

  → 모든 예측변수들이 동일하다면 같은 클래스에 할당될 가능성이 높기 때문에, **표본에서 새로 들어온 레코드와 정확히 일치하는 데이터를 찾는 것에 무게**를 두는 방식

▼ 5.1.1 나이브하지 않은 베이즈 분류는 왜 현실성이 없을까?

- 예측변수값이 정확히 일치하지 않는 데이터가 존재할 수 있기 때문

> ❗ CAUTION : **나이브베이즈는 베이즈 통계의 방법으로 간주되지 않음.**
> 나이브베이즈는 상대적으로 통계 지식이 거의 필요 없는 데이터 중심의 경험적 방법
> 다만, 베이즈 규칙과 비슷한 예측 계산 때문에 붙여진 이름. (결과가 주어졌을 때, 예측변수의 확률을 계산하는 부분과 결과 확률을 최종적으로 계산하는 부분에서 비슷)

▼ 5.1.2 나이브한 해법

$$P(Y=i|X_1,X_2,...,X_p)=\frac{P(Y=i)P(X_1,...,X_p|Y=i)}{P(Y=0)(P(X_1,...,X_p|Y=0)+P(Y=1)(P(X_1,...,X_p|Y=1)}$$

"

예측변수 벡터의 정확한 조건부확률은 각 조건부확률 $P(X_j|Y=i)$의 곱으로 충분히 잘 추정할 수 있다는 단순한 가정을 기초로 함
→ 즉, $P(X_1,X_2,...,X_p|Y=i)$ 대신 $P(X_j|Y=i)$를 추정하면서 X_j가 $k\neq j$ 인 모든 X_k와 서로 독립이라고 가정

1. 이진 응답 $Y=i(i=0또는1)$ 에 대해, 각 예측변수에 대한 조건부확률 $P(X_j|Y=i)$를 구한다. ($Y=i$일 때, 예측변수의 값이 나올 확률)
2. 각 확률값을 곱한다음, $Y=i$에 속한 레코드들의 비율을 곱한다.
3. 모든 클래스에 대해 1-2단계를 반복한다.
4. 2단계에서 모든 클래스에 대해 구한 확률값을 모두 더한 값으로 클래스 i의 확률을 나누면 결과 i의 확률을 구할 수 있다.
5. 이 예측변수에 대해 가장 높은 확률을 갖는 클래스를 해당 레코드에 할당한다.

`sklearn.naive_bayes.MultinomialNB`

▼ 5.1.3 수치형 예측변수

베이즈 분류기는 예측변수들이 범주형인 경우에 적합

→ 수치형 변수는 binning(구간화)하여 범주형으로 변환하거나 정규분포 같은 확률모형 사용

> ❗ CAUTION : **나이브베이즈는 특정 카테고리에 해당하는 데이터가 없을 때, 확률을 0으로 할당.**
> → 이를 방지하기 위해 평활화 인수(라플라스 평활화) 사용
> → *평활화 인수 : 학습 데이터에 없던 데이터가 출현해도 빈도수에 a를 더해, 확률이 0이 되는 것을 방지*

▼ 5.2 판별분석

- 선형판별분석(Linear Discriminanat Analysis : LDA)가 주로 사용됨
- LDA는 결정경계(Decision boundary)를 만들어 데이터를 분류하는 방법

> ✏️ 용어 정리
> - 공분산 : 하나의 변수가 다른 변수와 함께 변화하는 정도를 측정하는 지표
> - 판별함수 : 예측변수에 적용했을 때, 클래스 구분을 최대화하는 함수
> - 판별가중치 : 판별함수를 적용하여 얻은 점수로, 클래스에 속할 확률 추정에 사용

▼ 5.2.1 공분산행렬

- 두 변수 x와 z 사이의 공분산(두 변수 사이의 관계를 의미하는 지표)

$$S_{x,z} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z})}{n-1}$$

상관계수와 마찬가지로 양수는 양의 관계를, 음수는 음의 관계를 나타냄
단, 상관관계는 -1에서 1로 정의되지만 공분산의 척도는 변수 x와 z에서 사용하는 척도에 따라 달라짐

- 공분산 행렬

$$\hat{\Sigma} = \begin{bmatrix} s_x^2 & s_{x,z} \\ s_{z,x} & s_z^2 \end{bmatrix}$$

각 변수의 분산을 대각원소로 놓고, 변수간 공분산을 비대각원소에 위치시킨 행렬

- *선형판별분석 가정*
  - 독립변수가 다변량정규분포를 이룬다 (다변량정규분포 : 정규분포를 다차원 공간에 확장한 분포)
  - 종속변수에 의해 범주화되는 집단의 분산-공분산행렬이 동일해야 한다

▼ 5.2.2 피셔의 선형판별

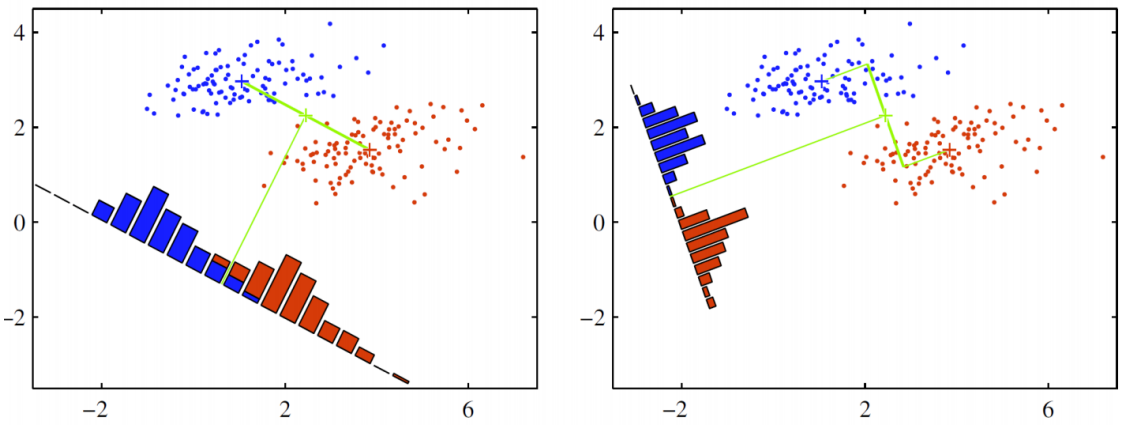
(수식 참고 : 선형판별분석)

- '내부' 제곱합(그룹 내 변동)에 대한 '사이' 제곱합(그룹 간 편차)의 비율을 최대화하는 직선을 찾는 것이 목표

$$\frac{SS_{\text{사이}}}{SS_{\text{내부}}}$$

- 사이 제곱합 : 두 그룹 평균 사이 거리 제곱
- 내부 제곱합 : 공분산행렬에 의해 가중치가 적용된, 각 그룹 내 평균이 주변으로 퍼져 있는 정도

→ 사이 제곱합을 최대화하고, 내부 제곱화를 최소화하는 것이 두 그룹 사이를 가장 명확하게 나누는 방법



- 오른쪽 그림의 분류가 더 잘 됨

`sklearn.discriminat_analysis.LinearDiscriminantAnalysis`

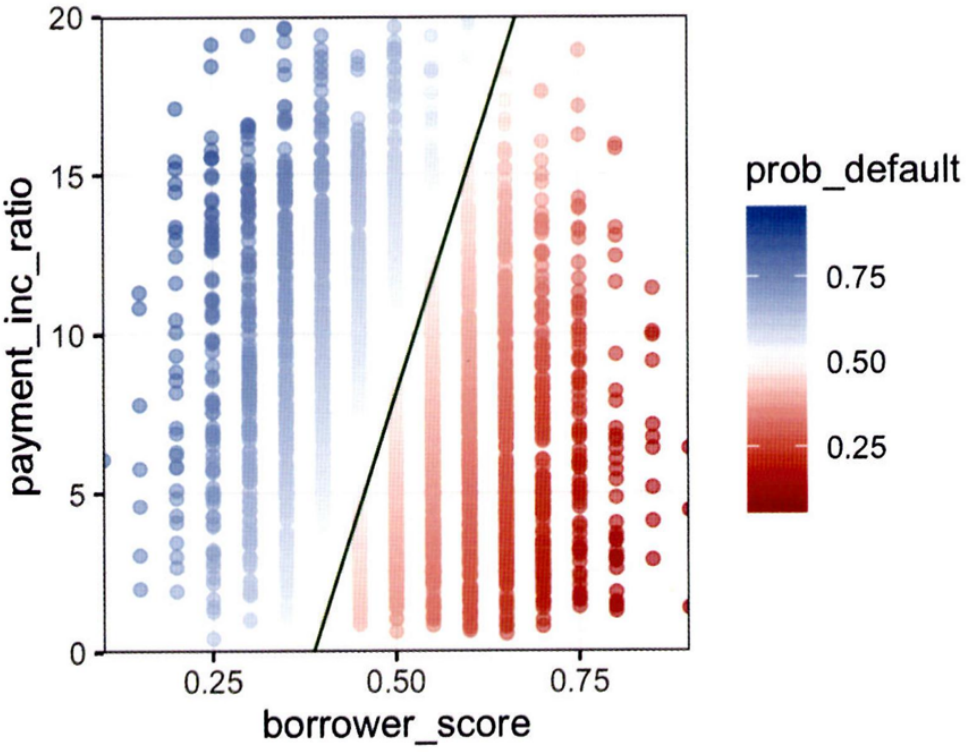


그림 5-1 두 변수(채무자의 신용점수와 소득에 대한 지급 비율)를 사용한 연체에 대한 LDA 예측 결과

- 실선으로부터 양방향으로 멀리 떨어진 예측결과일수록 신뢰도가 높음(즉, 확률이 0.5로부터 멀어짐)

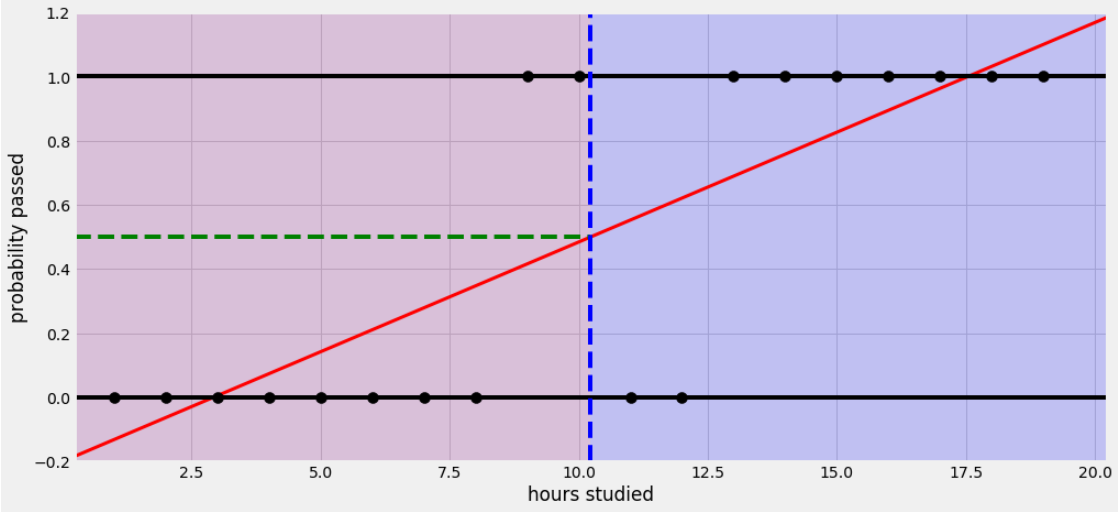
▼ 5.3 로지스틱 회귀

- 결과가 이진형 변수라는 점만 빼면 다중선형회귀와 유사

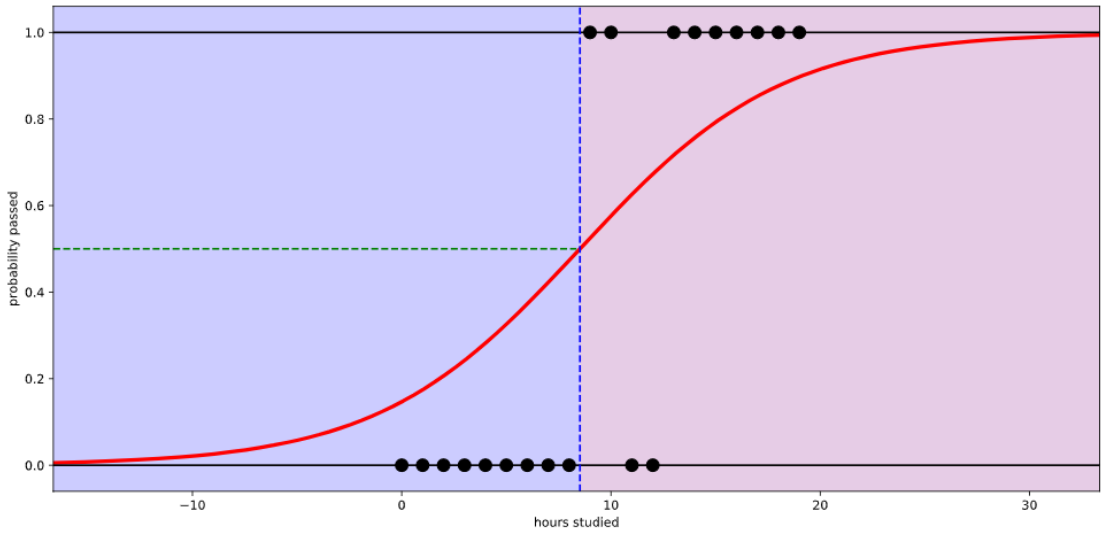
> ✏️ 용어정리
> - 로짓(logit) : (0~1이 아니라) $\pm\infty$ 범위에서 어떤 클래스에 속할 확률을 결정하는 함수
> - 오즈(odds) : '실패(0)'에 대한 '성공(1)'의 비율
> - 로그 오즈(log odds) : 변환모델(선형)의 응답변수. 이 값을 통해 확률을 구한다.

▼ 5.3.1 로지스틱 반응 함수와 로짓

(참고 : 로지스틱회귀 쉽게 이해하기)



- 공부시간에 따른 합격 확률 회귀 직선 그래프 : 불합격 확률이 0 이하



- 로지스틱 회귀 그래프 : 좀 더 말이 된다.

→ 로지스틱 회귀를 사용해 확률이 0과 1사이 값으로 그려짐

- 로지스틱 회귀 단계 (참고 : 시그모이드 함수 정의, Logistic)

1. 라벨이 '1'이 될 확률 p의 선형함수로 모델링

$$p = \beta_0 + \beta_1 x_1 + ... + \beta_q x_q$$

하지만, 선형함수를 이용해서 예측하는 것은 의미 없음 → Odds 이용

$$Odds(p) = \frac{p}{1-p}$$

확률 p의 범위가 (0,1)이라면, $Odds(p)$의 범위는 (0, ∞)가 되고, $Odds$에 로그함수를 취한 $log(Odds(p))$은 범위가 ($-\infty$, ∞)가 됨 → 이 값에 대해 선형 회귀 분석하는 것은 의미가 있음

2. 각 속성들의 값(value)에 계수(coefficient)를 곱해 log-odds를 구함

$$log(odds(p)) = \beta_0 + \beta_1 x_1 + ... + \beta_q x_q$$

3. 예측변수에 로지스틱 반응 혹은 역로짓 함수를 이용해 모델링
   → log-odds를 sigmoid 함수에 넣어 0-1 범위의 확률을 구함

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + ... + \beta_q x_q)}}$$

`sklearn.linear_model.LogisticRegression`

▼ 5.3.5 계수와 오즈비 해석하기

- 로지스틱 회귀분석에서 계수 B_j는 X_j에 대한 오즈비의 로그값이기 때문에 오즈비를 사용

$$\text{오즈비} = \frac{\text{오즈}(Y = 1|X = 1)}{\text{오즈}(Y = 1|X = 0)}$$

→ $X = 1$일 때 $Y = 1$인 경우와 $X = 0$일 때 $Y = 1$인 경우의 오즈를 비교한 것

예) 만약 오즈비가 2이면, $X = 1$일 때 $Y = 1$의 오즈가 $X = 0$일 때보다 두 배 더 높다는 것을 의미

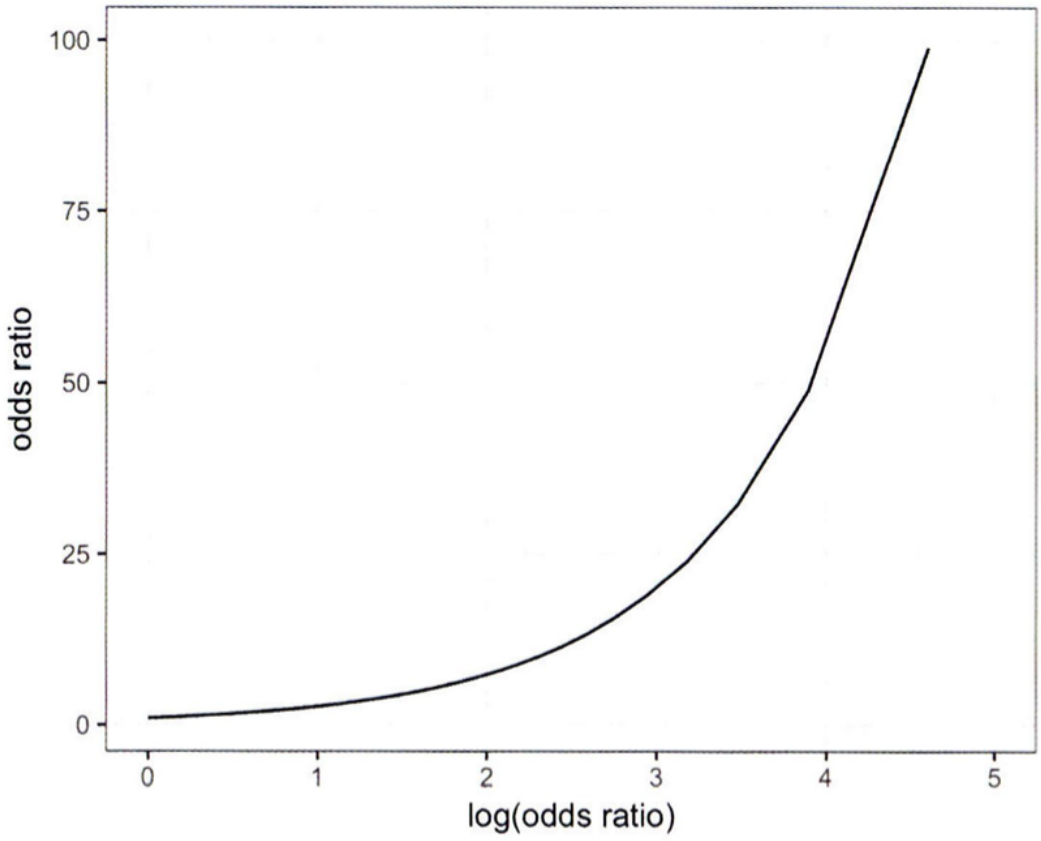


그림 5-3 오즈비와 로그 오즈비 사이의 관계

→ 계수가 로그 스케일이라서 1 증가할수록 오즈비는 $exp(1) \approx 2.72$만큼 증가함

→ 수치형 변수 X에서 단위 크기만큼 변화할 때도 오즈비에서의 변화를 생각할 수 있음

▼ 5.3.6 선형회귀와 로지스틱 회귀:유사점과 차이점

- 모델을 피팅하는 방식
  - 선형 회귀 : 최소제곱 사용
  - 로지스틱 회귀 : 최대우도추정(Maximum Likelihood Estimation : MLE) 사용

    - 우도 (Likelihood) :결과에 따라 가능한 가설을 평가할 수 있는 측도
    - 최대우도추정 : 결과에 해당하는 각 가설마다 계산된 우도 값 중 가장 큰 값을 추정 (일어날 가능성이 가장 큰 것)
- 잔차분석

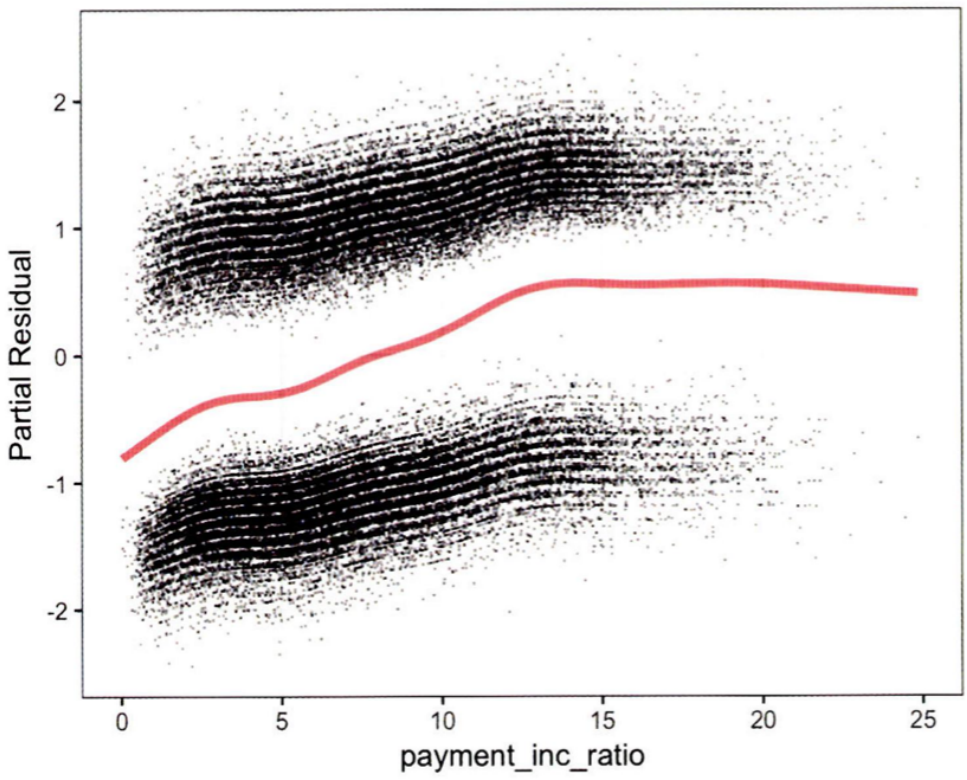


그림 5-4 로지스틱 회귀에서 얻은 편잔차

결과변수가 이진형이기 때문에 로지스틱 회귀에서 얻은 잔차는 구름 같은 모양이 두군데 있고, 추정결과로 얻은 회귀선이 그사이를 지나가는 형태 (위쪽구름은 1, 아래쪽구름은 0 의미)

▼ 5.4 분류 모델 평가하기

(참고 : ROC Curve란 무엇이고 어떻게 그리는가, 이진분류의 성능평가지표)

- 컷오프(threshold) 기준값
    - 가장 기본적인 기준 값은 0.5
    - 실제 데이터에서 1이 차지하는 비율을 컷오프로 사용하는 방법

▼ 5.4.1 혼동행렬

- 분류 결과를 나타내는 가장 대표적인 행렬

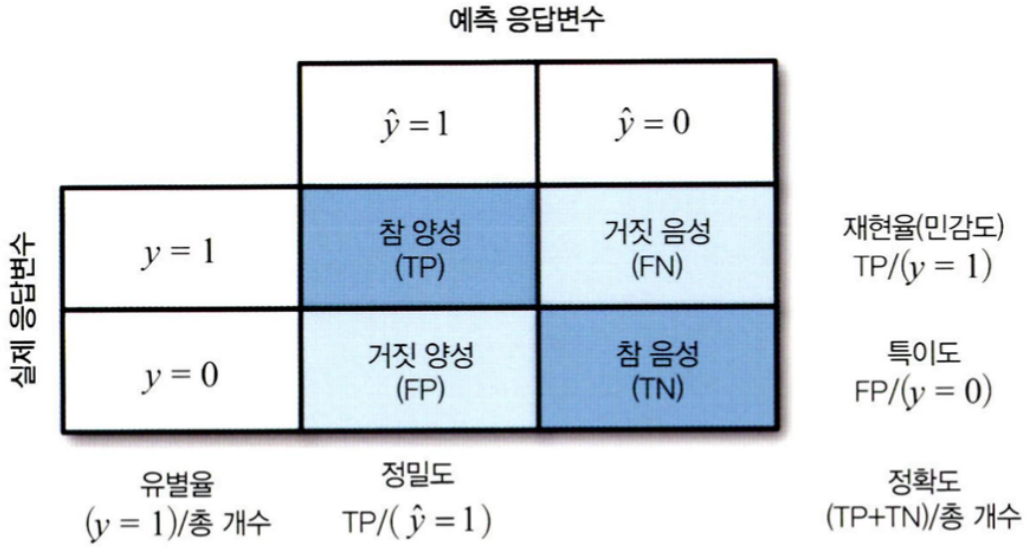
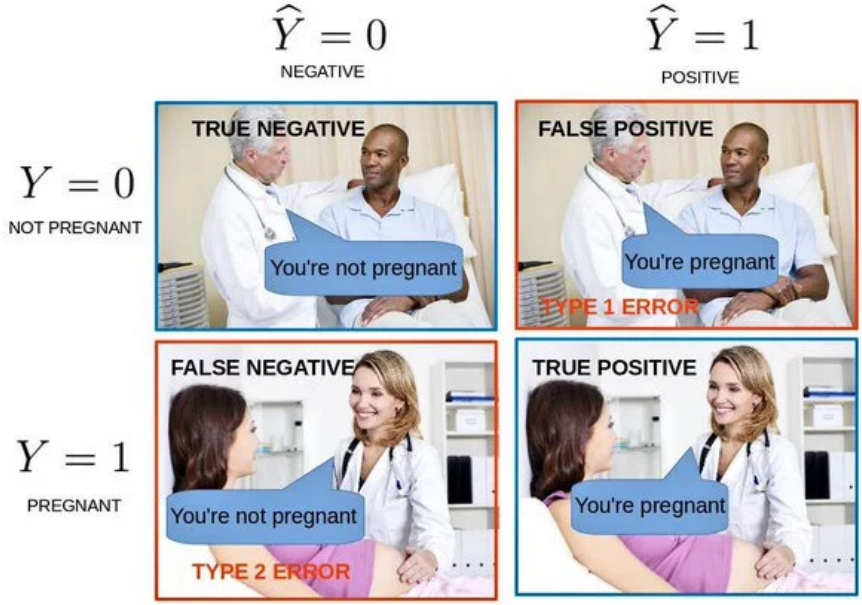


그림 5-5 이진 응답변수에 대한 혼동행렬과 그에 관련된 다른 지표들



- 보통 데이터 수가 상대적으로 작은(희귀한) 클래스가 관심의 대상이 되므로 1로 지정, 반대를 0으로 지정
  → 즉, 일반적인 경우에는 1이 더 중요한 사건을 의미
- 중요한 지표 중 하나는 거짓양성비율(FPR)

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

결과가 1인 데이터수가 희박할 때, 모든 예측 응답변수에 대해 거짓 양성 값의 비율이 높아져 예측결과는 1이지만 실제로는 0일 가능성이 높은 상황이 됨
예) 유방 조영술 검사결과가 양성이라고 해서 그것이 바로 유방암을 의미하는 것은 아님

- 정밀도, precision : 예측된 1에 대해 실제 1의 비율 → 예측된 양성 결과의 정확도

$$Precision = \frac{TP}{FP + TP} = \frac{\text{예측, 실제1}}{\text{예측1}}$$

→ FP(실제 0, 예측 1)을 낮추는데 초점

- 재현율(민감도), recall : 실제 1에 대해 1이라고 예측한 결과의 비율 → 양성 결과를 예측하는 모델 능력 평가

$$Recall = \frac{TP}{FN + TP} = \frac{\text{예측, 실제1}}{\text{실제1}}$$

→ FN(실제 1, 예측 0)을 낮추는데 초점

- 특이도, specificity : 실제 0에 대해 0이라고 예측한 결과의 비율 → 음성결과를 예측하는 모델 능력 평가

$$Specificity = \frac{TN}{FP + TN} = \frac{\text{예측, 실제0}}{\text{실제0}}$$

- 재현율과 특이도 사이에는 트레이드오프 관계가 있음 → ROC 곡선 이용

▼ 5.4.4 ROC 곡선

- 수신자 조작 특성 (Receiver Operating Characteristic, ROC curve) 표시 방법

  y축에 민감도를 표시하면서,

  - x축 왼쪽에 1부터 오른쪽에 0까지 특이도를 표시한다.
  - x축 왼쪽에 0부터 오른쪽에 1까지 1-특이도를 표시한다.

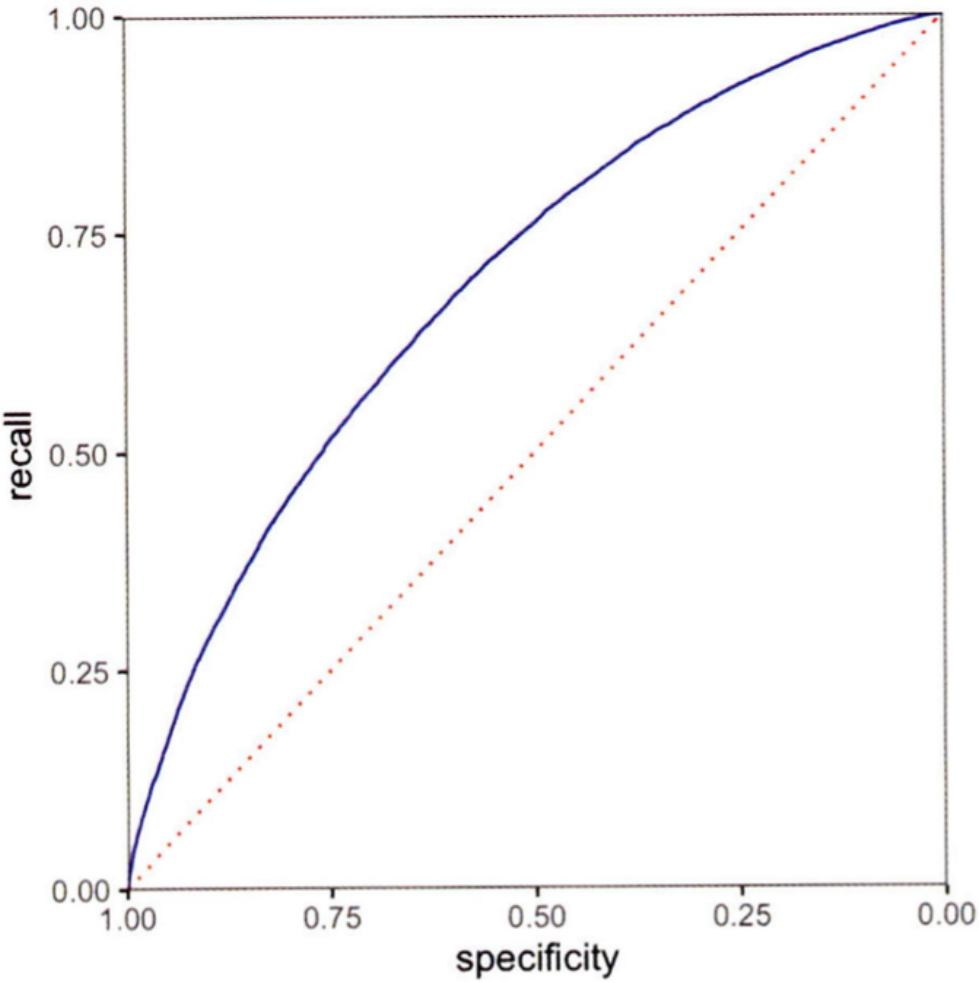
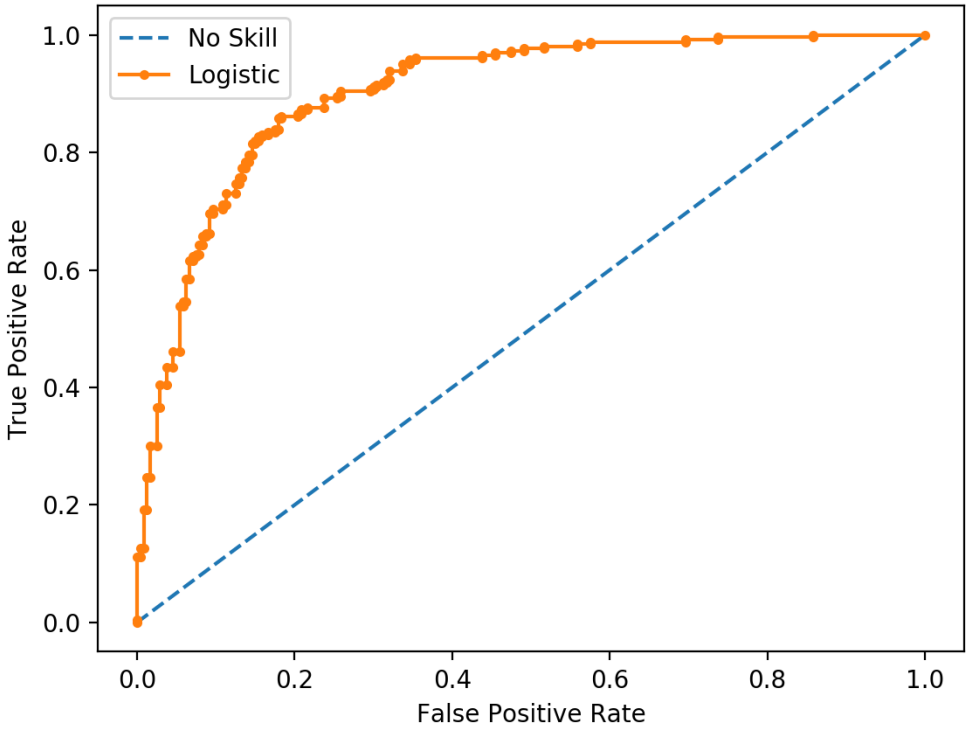


그림 5-6 대출 데이터에 대한 ROC 곡선

- y축 : 재현율, x축 = 특이도



- y축 : TPR = 재현율, x축 : FPR

$$1 - Specificity = 1 - \frac{TN}{FP + TN} = \frac{(FP + TN) - TN}{FP + TN} = \frac{FP}{FP + TN} = FPR$$

- 점선은 랜덤으로 예측했을 때의 결과이고, 좋은 분류기는 ROC 곡선이 왼쪽 상단에 가까운 형태

`sklearn.metrics.roc_curve`

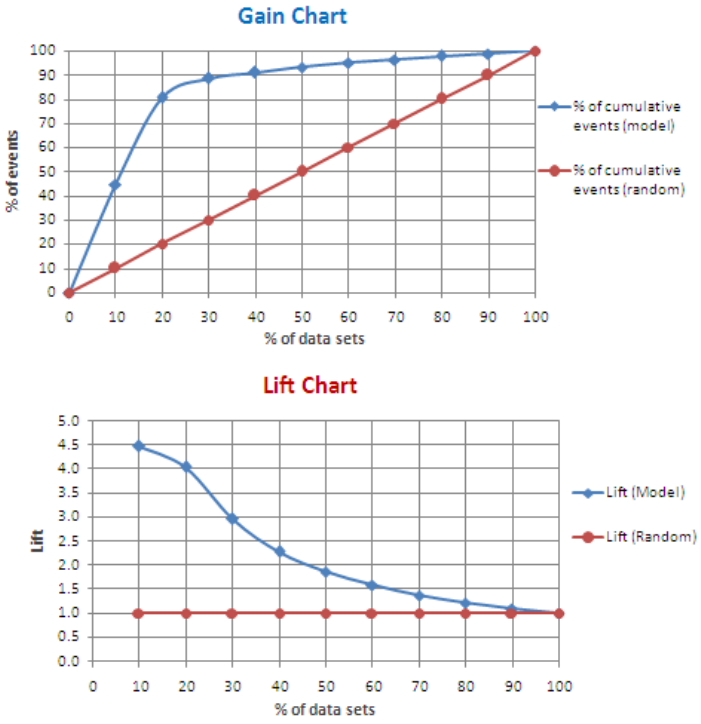
▼ 5.4.5 AUC

- ROC 곡선이 훌륭한 시각화 도구이지만, 하나의 값을 주지 않음 → ROC 곡선 아래 면적인 AUC(Area Underneath the Curve, AUC) 지표 이용
- AUC 값이 높을 수록 더 좋은 분류 성능을 나타냄
- 최악은 ROC 곡선이 가운데를 지나가는 직선인 경우 → AUC 0.5

▼ 5.4.6 리프트

- 최적의 컷오프(threshold)를 찾기 위한 중간 단계로 활용 가능
- 모델이나 알고리즘의 확률 예측 능력을 평가하는 방법

예) 1로 예측될 확률이 있는 레코드를 정렬 → '상위 10%의 레코드를 1로 분류하는 알고리즘 vs 아무거나 선택하는 경우' 비교 → 무작위로 선택했을 때 0.1%의 정확도를 얻었고, 상위 10%에서 0.3%의 결과를 얻었다면 이 알고리즘은 상위 10%에서 3의 리프트를 갖는다.



(참고 : Understand gain and lift charts)

▼ 5.5 불균형 데이터 다루기

(참고 : 클래스 불균형이란?)

▼ 5.5.1 과소표본추출(다운샘플링)

- 다수의 클래스에 속한 데이터들 중에 중복된 레코드가 많을 것이라는 사실에서 출발
- 다수의 데이터에 해당하는 클래스에서 과소표본추출(다운샘플링)
- 단점 : 데이터의 일부가 버려지기 때문에 모든 정보 활용 불가

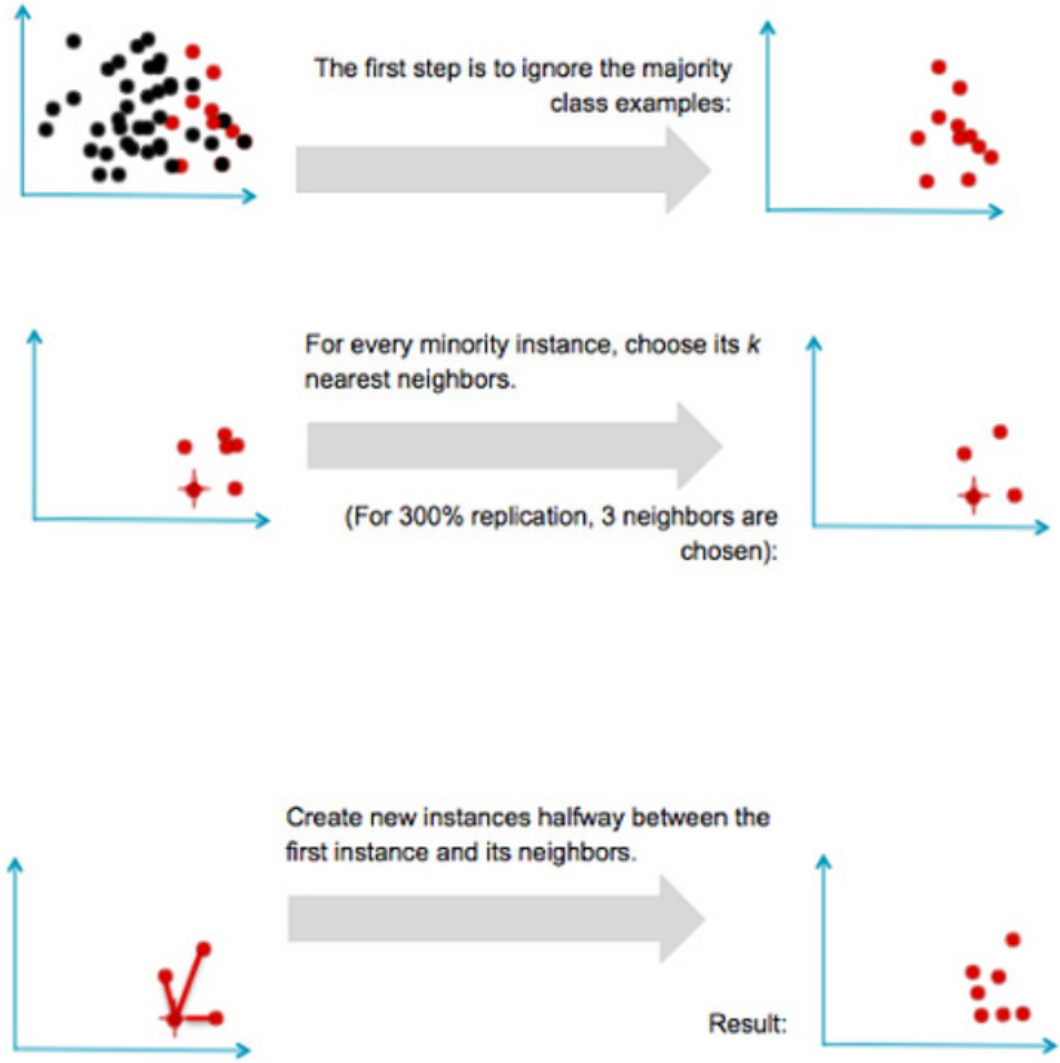
▼ 5.5.2 과잉표본추출과 상향/하향 가중치

- 복원추출방식(부트스트래핑)으로 희귀 클래스 데이터를 과잉표본추출(업샘플링)
- 가중치 : 1/다수클래스 확률

예) 클래스 수가 2:8이라면 loss에 대한 가중치를 8:2로 설정

▼ 5.5.3 데이터 생성

- 업샘플링 방식의 변형으로 기존에 존재하는 데이터를 살짝 바꿔 새로운 데이터를 만드는 방법
- 합성 소수 과잉표본 기법(Synthetic Minority Oversampling Technique, SMOTE) : 업샘플링된 레코드와 비슷한 레코드를 찾고 원래 레코드와의 랜덤 가중평균으로 새로운 합성 레코드 생성



소수 데이터 중 관측치 X를 잡아 X와 가장 가까운 이웃 K개를 선택→K개의 X와 X사이 임의의 새로운 데이터 X' 생성 (참고 : SMOTE를 통한 데이터 불균형 처리)