Chapter07

로지스틱회귀

I로지스틱회귀 평가지표

FASTCAMPUS
ONLINE

금융공학/퀀트 I

강사. 장순용

키포인트

- 혼동행렬.
- 정확도, 민감도, 특이도, 정밀도.
- ROC 곡선.
- McNemar 검정과 예측 모형의 비교.

I 혼동행렬 (Confusion matrix)

	Actual 0	*Actual 1*
Predicted 0	134	42
Predicted 1	10	14

I 혼동행렬과 정확도 (Accuracy)

	Actual 0	*Actual 1*
Predicted 0	134	42
Predicted 1	10	14

⇒ 정확도는 행렬의 대각선의 합과 전체의 합 사이의 비율.

⇒ 예측된 유형이 실제 유형과 일치하는 비율.

I 맹점 발견!

- 그런데 정확도 만으로 테스트가 불가능한 상황이 종종 발생한다.

예). 은행 대출고객 중에서 3% ~ 5%만이 향후 신용불량인 경우.

⇒ 목표는 소수인 신용불량 고객을 사전에 검출하는 것.

⇒ 만약에 모두를 신용양호로 예측한다면 정확도는 **매우 높다!**

⇒ 하지만 신용불량 고객은 한명도 예측하지 못한다.

I 성능의 척도

- Accuracy (정확도) = $\frac{\text{정확하게 예측된 개수}}{\text{전체 개수}}$
- Sensitivity (민감도) = $\frac{\text{정확하게 예측된 1의 개수}}{\text{실제 1의 개수}}$
- Specificity (특이도) = $\frac{\text{정확하게 예측된 0의 개수}}{\text{실제 0의 개수}}$

I 성능의 척도

- Precision (정밀도) = $\frac{\text{정확하게 예측된 1의 개수}}{\text{1로 예측된 개수}}$
- Recall (재현율) = 민감도와 같은 의미
- Cohen의 카파 $\kappa = \frac{Accuracy - p_e}{1 - p_e}$ ⇐ p_e는 우연으로 맞을 확률.

I 혼동행렬과 민감도 (Sensitivity)

	Actual 0	*Actual 1*
Predicted 0	134	42
Predicted 1	10	14

⇒ 민감도는 실제 1 중에서 정확하게 1로 예측된 비율.

I 혼동행렬과 특이도 (Specificity)

	Actual 0	*Actual 1*
Predicted 0	134	42
Predicted 1	10	14

⇒ 특이도는 실제 0 중에서 정확하게 0으로 예측된 비율.

I 혼동행렬과 정밀도 (Precision)

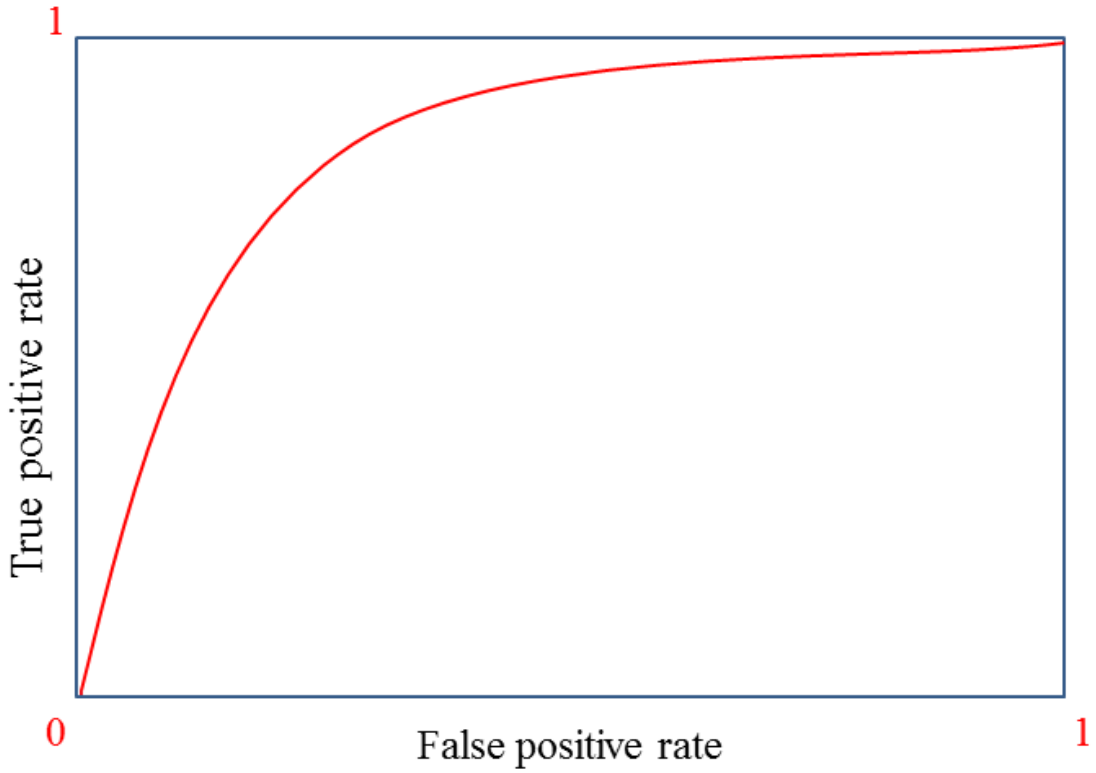
	Actual 0	*Actual 1*
Predicted 0	134	42
Predicted 1	10	14

⇒ 정밀도는 1로 예측된 경우 중에서 정확하게 1로 예측된 비율.

I 성능의 척도

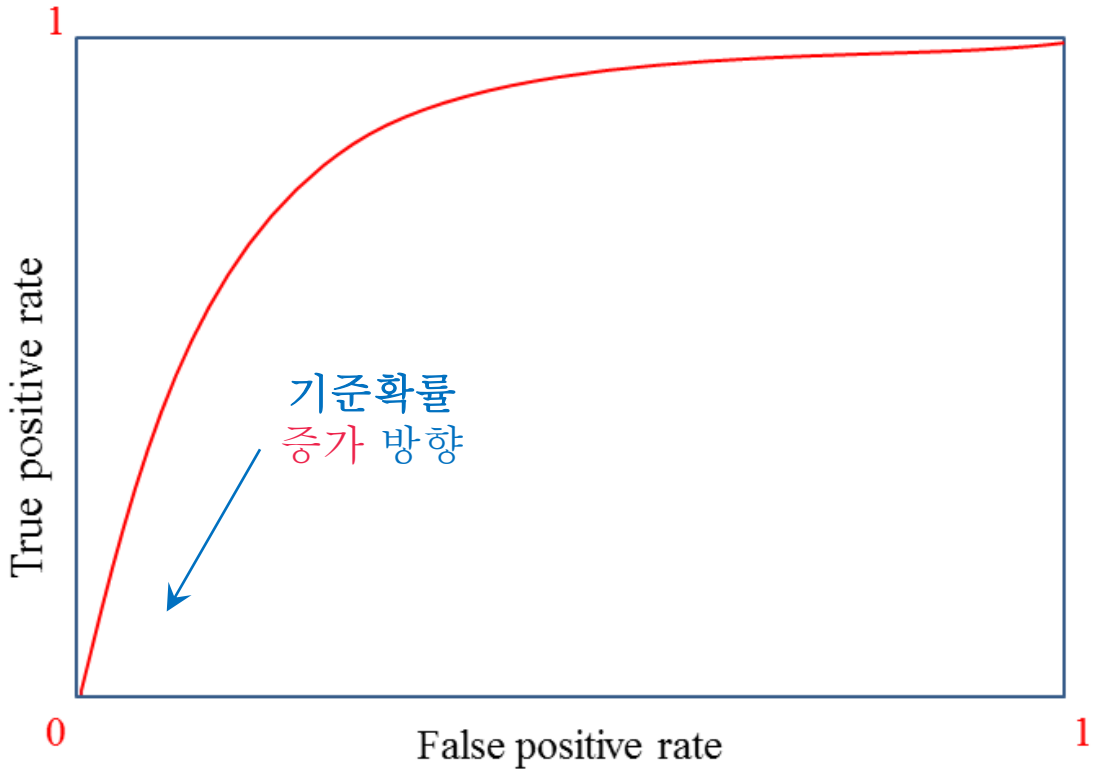
- True Positive Rate = Sensitivity
- True Negative Rate = Specificity
- False Positive Rate = $\frac{\text{실제는 0이지만 1로 인식된 개수}}{\text{실제 0의 개수}}$ = 1- Specificity
- False Negative Rate = $\frac{\text{실제는 1이지만 0로 인식된 개수}}{\text{실제 1의 개수}}$ = 1- Sensitivity
- Positive Predicted Value = Precision

I ROC 곡선



⇒ ROC 곡선은 기준확률에 대한 parametric plot.

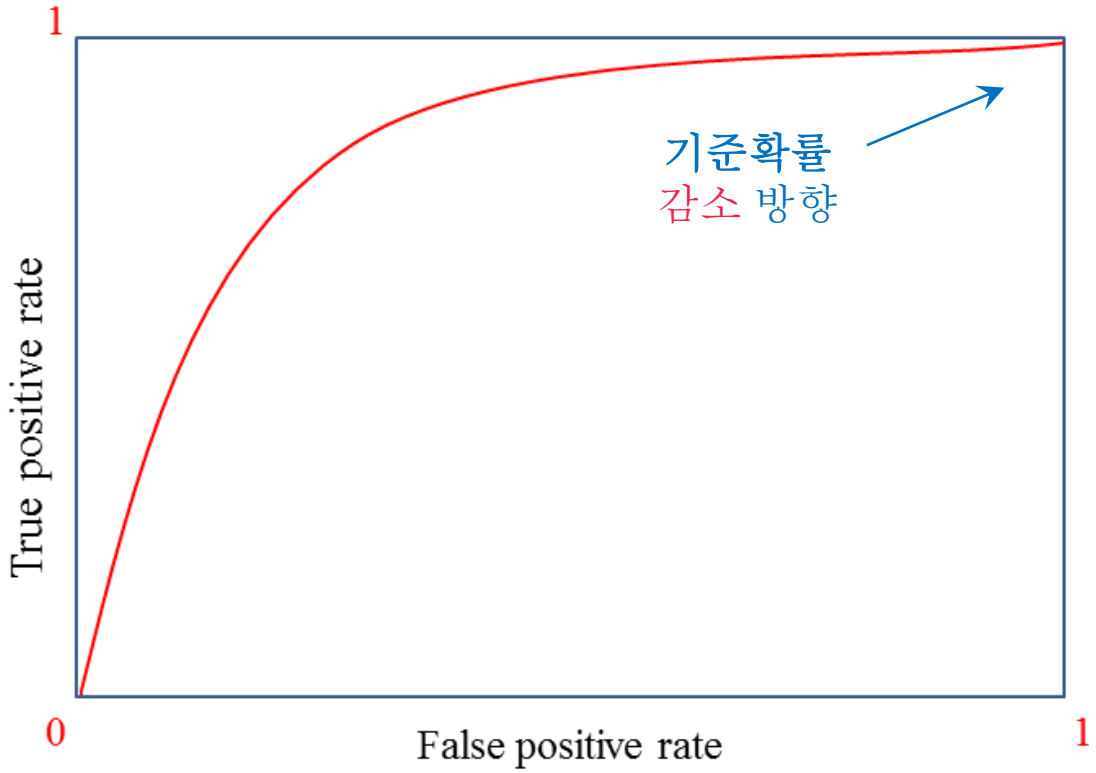
I ROC 곡선



기준확률이 증가 (1에 가까워 진다)	
성능척도	방향
민감도 (True Positive)	⇓
특이도	⇑
1-특이도 (False Positive)	⇓
정밀도	⇑

⇒ ROC 곡선은 기준확률에 대한 parametric plot.

I ROC 곡선



기준확률이 감소 (0에 가까워 진다)	
성능척도	방향
민감도 (True Positive)	⇑
특이도	⇓
1-특이도 (False Positive)	⇑
정밀도	⇓

⇒ ROC 곡선은 기준확률에 대한 parametric plot.

I ROC 곡선과 AUC

⇒ AUC는 곡선 아래의 면적 (Area Under the Curve)을 의미함.

⇒ AUC가 클수록 (1에 가까울 수록) 예측성능이 좋은 것이다.

I 예측모형의 비교검정 (McNemar)

- 2×2 테이블에 적용 가능 $\Rightarrow$ 두 모델의 예측 결과를 비교.
- 가로 방향의 합 (도수분포표)과 세로 방향의 합 (도수분포표) 비교 검정.

	Predicted′ 0	Predicted′ 1	
Predicted 0	a	b	$a+b$
Predicted 1	c	d	$c+d$
	$a+c$	$b+d$	

- $a+b=a+c$ 와 $c+d=b+d$를 검정 $\Rightarrow$ $b=c$ 검정으로 대체.

귀무가설 $\boldsymbol{H_0}$: $b=c$이다. 두 모델 사이에 차이가 없다.

대립가설 $\boldsymbol{H_1}$: $b \neq c$이다. 두 모델 사이에 차이가 있다.

I 예측모형의 비교검정 (McNemar)

- 다음 검정통계량과 자유도 1인 카이제곱확률을 사용하여 검정한다.

$$\text{검정통계량} = \frac{(b-c)^2}{b+c}$$

- p 값으로 귀무가설의 유지/기각 여부를 결정.

I 로지스틱회귀 해석 (베이즈 정리)

문제: 500개의 관측치가 있다. 이 중에서 종속변수의 값이 1인 경우는 30회이고 0인 경우는 나머지 470회이다. 그런데 로지스틱회귀 모형의 민감도는 0.92이고 특이도는 0.90이다. 만약에 이 모형을 가지고 한 예측결과가 1이라면 어느정도 믿을 수 있겠는가?

I 로지스틱회귀 해석 (베이즈 정리)

문제: 500개의 관측치가 있다. 이 중에서 종속변수의 값이 1인 경우는 30회이고 0인 경우는 나머지 470회이다. 그런데 로지스틱회귀 모형의 민감도는 0.92이고 특이도는 0.90이다. 만약에 이 모형을 가지고 한 예측결과가 1이라면 어느정도 믿을 수 있겠는가?

이 문제의 조건을 정리해 보면 다음과 같다.

$$P(\text{예측 } 1 | \text{실제 } 1) = 0.92 \qquad \text{“민감도”}$$

$$P(\text{예측 } 0 | \text{실제 } 0) = 0.90 \qquad \text{“특이도”}$$

$$\Rightarrow P(\text{예측 } 1 | \text{실제 } 0) = 1 - P(\text{예측 } 0 | \text{실제 } 0) = 0.10$$

$$P(1) = 30/500 = 0.06$$

I 로지스틱회귀 해석 (베이즈 정리)

문제: 500개의 관측치가 있다. 이 중에서 종속변수의 값이 1인 경우는 30회이고 0인 경우는 나머지 470회이다. 그런데 로지스틱회귀 모형의 민감도는 0.92이고 특이도는 0.90이다. 만약에 이 모형을 가지고 한 예측결과가 1이라면 어느정도 믿을 수 있겠는가?

그러므로 이 문제가 요구하는 답은 $P(\text{실제 } 1|\text{예측 } 1)$이다.

이것을 베이즈 정리를 적용하여 계산해 보면 다음과 같다.

$$P(\text{실제 } 1|\text{예측 } 1) = \frac{P(\text{예측 } 1|\text{실제 } 1)P(1)}{P(\text{예측 } 1|\text{실제 } 1)P(1) + P(\text{예측 } 1|\text{실제 } 0)P(0)}$$

$$= \frac{0.92 \times 0.06}{0.92 \times 0.06 + 0.1 \times 0.94} \cong \mathbf{0.37}$$

I 끝.

감사합니다.

FAST CAMPUS
ONLINE
장순용 강사.