



부산대학교

PUSAN NATIONAL UNIVERSITY

Evaluation Metrics

Counting the Mistakes

Which Number Counts Which Error

BX3600087-140 / DS3600087-140 / AB3600087-140

기계학습개론

2026학년도 2학기

오용경

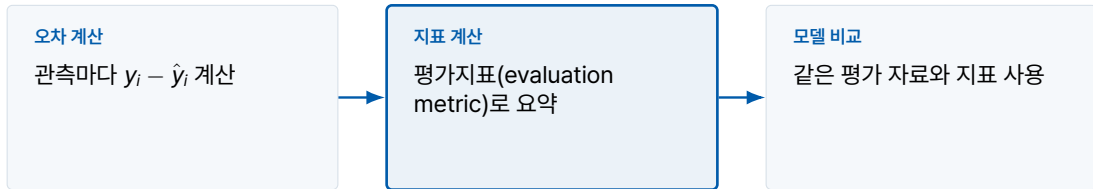
데이터사이언스전공

yongkyung.oh@pusan.ac.kr

Regression Metrics

수치 예측의 오차를 하나의 수로 적는다

예측 오차를 지표로 요약한다



회귀 지표가 요약하는 값

지표	계산	단위	해석
MSE	오차 제곱의 평균	관측값 단위의 제곱	큰 오차에 큰 벌점
RMSE	MSE의 제곱근	관측값과 같음	오차 크기를 원래 단위로
MAE	절대 오차의 평균	관측값과 같음	오차 크기에 비례한 벌점
R^2	결정계수, 평균 예측 대비 오차 감소	없음	같은 자료의 기준선과 비교

Source: scikit-learn MOOC (2022), regression metrics.

같은 절대 오차 합도 MSE는 다르다

10개 관측의 절대 오차	MSE	MAE
10, 나머지 9개는 0	$100/10 = 10$	$10/10 = 1$
10개 모두 1	$10/10 = 1$	$10/10 = 1$

MSE는 오차가 일부 관측에 집중된 경우 더 커진다

평가 목적에 맞는 회귀 지표

평가 목적	지표	주의할 점
큰 오차를 강하게 반영	MSE, RMSE	제곱 때문에 큰 오차의 영향이 큼
오차를 선형적으로 반영	MAE	이상치의 영향이 없어지는 것은 아님
평균 예측과 비교	R^2	관측값 분포가 다르면 단순 비교 금지
원래 단위로 오차 보고	RMSE, MAE	두 지표의 계산 방식은 다름

무엇을 큰 오류로 볼지에 따라 지표를 먼저 정한다

R^2 의 기준선은 평가 자료의 평균

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

값	평가 자료에서의 의미
1	예측 오차가 0
0	평가 자료의 평균을 항상 예측한 것과 같은 제곱오차
음수	그 평균을 항상 예측한 것보다 제곱오차가 큼

관측값이 모두 같아 분모가 0인 경우는 이 식으로 정의되지 않는다

같은 예측 결과를 네 지표로 계산한다

가상 관측값 $y = (2, 4, 6, 8)$, 예측값 $\hat{y} = (2, 5, 5, 10)$

MSE	RMSE	MAE	R^2
$6/4 = 1.50$	$\sqrt{1.5} \approx 1.22$	$4/4 = 1.00$	$1 - 6/20 = 0.70$

절대 오차는 0, 1, 1, 2다. RMSE는 항상 MAE 이상이고 모든 절대 오차가 같을 때만 같으므로, RMSE가 더 크다는 사실만으로 이상치를 단정할 수 없다

Checkpoint

Regression Metrics

- MSE와 MAE는 크게 빛나간 관측을 어떻게 다르게 세는가
- RMSE를 보고에 쓰는 까닭은 무엇인가
- R^2 가 음수이면 무슨 뜻인가

The Confusion Matrix

분류의 실수를 네 칸으로 나눈다

이진분류의 두 가지 오류

DEFINITION

거짓 양성(FP)

실제 음성인 관측을 양성으로 예측한 경우

DEFINITION

거짓 음성(FN)

실제 양성인 관측을 음성으로 예측한 경우

FP는 false positive, FN은 false negative를 뜻한다

네 칸으로 적는다

	실제 양성	실제 음성	
양성 예측	TP: 양성을 맞힘	FP: 음성을 양성으로	→ $TP / (TP + FP)$
음성 예측	FN: 양성을 음성으로	TN: 음성을 맞힘	

↓
 $TP / (TP + FN)$

혼동행렬의 관측 수로 정확도·정밀도·재현율을 계산한다

Source: James et al. (2023), Section 4.4.

양성으로 둘 범주를 먼저 정한다

문제	양성으로 두는 쪽	까닭
질환 검사	환자	찾아내는 것이 목적
불량품 검사	불량	놓치면 안 되는 쪽
스팸 분류	스팸	걸러 내려는 쪽

Caution 양성을 어느 쪽으로 두느냐에 따라 정밀도와 재현율의 의미가 달라지지만 정확도는 같다

정확도는 전체 중 맞힌 비율이다

$$\text{정확도} = \frac{TP + TN}{TP + FP + FN + TN}$$

반영하는 것

반영하지 않는 구분

전체 중 맞힌 비율

FP와 FN의 상대적인 비용

전체 관측 수가 0보다 클 때 정의된다

혼동행렬은 먼저 축을 확인한다

이 표의 방향	합의 의미	계산할 수 있는 비율
양성 예측 행	$TP + FP$: 양성 예측 수	정밀도 = $TP / (TP + FP)$
실제 양성 열	$TP + FN$: 실제 양성 수	재현율 = $TP / (TP + FN)$
대각선	$TP + TN$: 맞힌 수	정확도의 분자

도구마다 행·열 방향이 다르므로 축 이름을 먼저 확인한다. scikit-learn의 `confusion_matrix`는 행이 실제, 열이 예측이며 라벨 오름차순이라 0/1이면 음성(0)이 먼저다

Default 자료의 실제 혼동행렬

로지스틱 회귀(잔액·소득·학생 여부)를 고객 7,000명으로 학습하고 평가 자료 3,000명에 임계값 0.5 적용

예측 / 실제	연체(양성)	연체 아님(음성)
연체	$TP = 39$	$FP = 9$
연체 아님	$FN = 74$	$TN = 2,878$

앞의 네 칸 표와 같은 방향(행 예측, 열 실제, 양성 먼저)이다. 정확도는 $(39 + 2,878)/3,000 \approx 0.972$ 지만 실제 연체 113명 중 39명만 찾았다

Data: James et al. (2023), Default.

Checkpoint

The Confusion Matrix

- 두 종류의 실수는 각각 무엇인가
- 양성을 어느 쪽으로 돌지가 왜 먼저인가
- 정확도는 네 칸 중 어느 칸을 세는가

Precision and Recall

네 칸에서 두 비율을 뽑는다

DEFINITION

정밀도(precision)

양성으로 예측한 관측 중 실제 양성의 비율

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

DEFINITION

재현율(recall)

실제 양성 중 양성으로 예측한 비율

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

각 분모가 0보다 클 때 정의되며, 분모가 0이면 별도의 처리 규칙이 필요하다

혼동행렬에서 직접 계산한다

가상 관측 1,000개 중 실제 양성은 20개, 양성 예측은 15개

예측 / 실제	양성	음성
양성	$TP = 12$	$FP = 3$
음성	$FN = 8$	$TN = 977$

정밀도 $12/15 = 0.80$, 재현율 $12/20 = 0.60$

민감도와 특이도

지표	계산	조건 집단
재현율 = 민감도(sensitivity) = TPR(true positive rate)	$TP / (TP + FN)$	실제 양성
특이도(specificity) = TNR(true negative rate)	$TN / (TN + FP)$	실제 음성
FPR(false positive rate) = 1 - 특이도	$FP / (FP + TN)$	실제 음성

이름이 달라도 분자와 분모가 같으면 같은 지표다

정밀도와 재현율의 조화평균

$$F_1 = \frac{2PR}{P+R} \quad (P+R > 0)$$

정밀도 P	재현율 R	F_1
0.80	0.60	$24/35 \approx 0.69$
0.80	0.80	0.80

F_1 은 두 값이 불균형하면 산술평균보다 작지만, 한쪽만 개선해도 높아질 수 있다

하나의 지표만으로 판단하지 않는다

가상 관측 1,000개, 양성 20개	정밀도	재현율	F_1
모두 양성으로 예측	0.02	1.00	0.04
실제 양성 1개만 양성으로 예측	1.00	0.05	0.10
앞의 혼동행렬	0.80	0.60	0.69

Caution 한쪽 지표가 1이어도 다른 오류와 그 비용을 함께 확인해야 한다

scikit-learn으로 같은 지표를 계산한다

앞의 Default 표를 만든 코드다. X는 잔액·소득·학생 여부(학생=1), y는 연체=1이다.
분할·파이프라인·모델의 import 줄은 생략했다

```
1 from sklearn.metrics import (confusion_matrix, precision_score,
2                               recall_score, f1_score)
3 X_train, X_test, y_train, y_test = train_test_split(
4     X, y, test_size=0.3, random_state=0)
5 model = make_pipeline(StandardScaler(), LogisticRegression())
6 y_pred = model.fit(X_train, y_train).predict(X_test)
7 print(confusion_matrix(y_test, y_pred))
8 # [[2878    9]
9 #    [  74   39]]
10 for score in (precision_score, recall_score, f1_score):
11     print(score.__name__, round(score(y_test, y_pred), 4))
12 # precision_score 0.8125
13 # recall_score 0.3451
14 # f1_score 0.4845
```

출력 행렬은 행이 실제, 열이 예측이고 라벨 오름차순(0, 1)이라 앞의 표를 전치하고 양성·음성 순서를 바꾼 모양이다. 세 함수는 pos_label=1을 양성으로 둔다. 손으로 계산하면 39/48, 39/113, 78/161이다

Adapted from scikit-learn MOOC (2022), classification metrics · data: James et al. (2023), Default.

Checkpoint

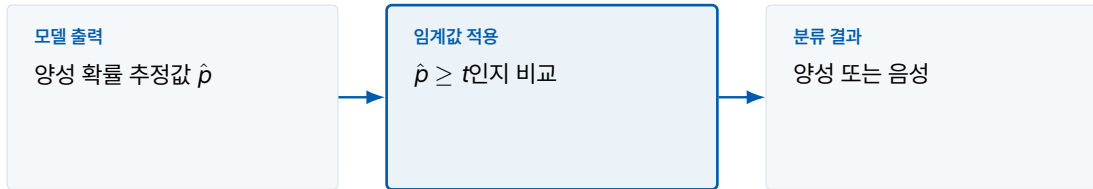
Precision and Recall

- 두 비율의 분모는 각각 무엇인가
- 재현율과 민감도는 어떤 관계인가
- F_1 이 산술평균이 아닌 까닭은 무엇인가

Threshold and ROC

한 모델에서 여러 분류기를 만든다

확률 추정값을 분류 결과로 바꾼다



같은 모델도 임계값에 따라 혼동행렬이 달라질 수 있다

임계값을 낮추면 양성 예측이 늘어난다

앞의 가상 자료에서 실제 양성 20개, 음성 980개는 그대로 유지

임계값	TP	FP	FN
0.8	6	1	14
0.5	12	3	8
0.2	18	25	2

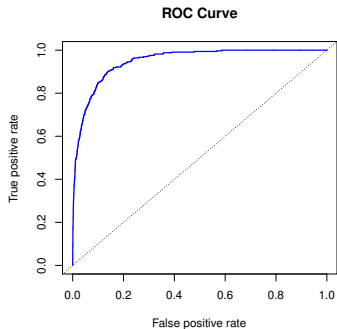
고정된 예측값에서 임계값을 낮추면 FP 는 늘거나 같고 FN 은 줄거나 같다

ROC 곡선의 두 축

축	지표	계산
세로축	TPR, 재현율	$TP/(TP + FN)$
가로축	FPR	$FP/(FP + TN)$

실제 양성과 음성이 모두 있는 자료에서 임계값마다 두 비율을 계산해 점을 잇는다. 앞 표의 세 임계값은 (FPR, TPR) = (0.001, 0.30), (0.003, 0.60), (0.026, 0.90)이다

ROC 곡선



Default 자료에서 연체 여부를 예측한 분류기의 ROC 곡선(AUC 0.95). 왼쪽 위는 TPR이 높고 FPR이 낮은 영역이며 대각선은 무작위 순위의 기준선이다

Source: James et al. (2023), Figure 4.8.

ROC 곡선 아래 면적(AUC)은 점수의 순위를 평가한다

AUC	평가 자료에서의 해석
1.0	모든 양성의 점수가 모든 음성의 점수보다 높음
0.5	양성·음성 쌍의 평균 순위 성능이 무작위 수준
0.5 미만	점수 방향과 양성 범주의 지정부터 확인

점수는 양성 확률 추정값 $\hat{p}$ 다. AUC는 양성의 점수가 음성보다 높을 확률에 동점 확률의 절반을 더한 값과 같다

AUC가 0.5라는 사실만으로 모델이 아무것도 학습하지 않았다고 단정할 수 없다

같은 FPR도 오류 건수는 다르다

실제 음성 수	FPR	FP
1,000	1%	10
10,000	1%	100

Caution ROC는 범주별 비율을 보여 주므로 실제 FP 건수는 혼동행렬에서 함께 확인한다

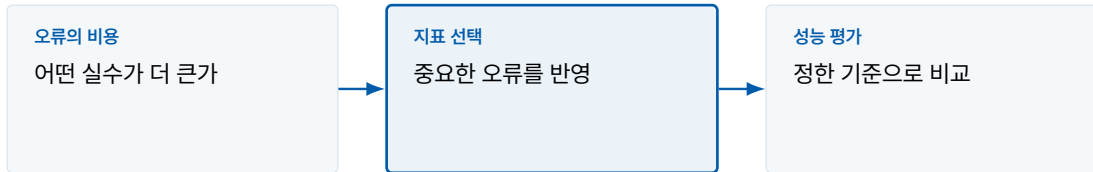
지표와 함께 보고할 정보

함께 적는 것	왜
혼동행렬 그대로	네 칸의 실제 수를 보인다
사용한 임계값	같은 모델도 임계값에 따라 결과가 달라진다
양성으로 둔 쪽	지표의 방향이 정해진다
범주별 비율	정확도를 읽는 기준이 된다

수 하나만 적으면 읽는 사람이 어느 실수를 봤는지 알 수 없다

용어	무엇을 가리키는가
혼동행렬	두 종류의 실수를 나눠 적은 네 칸 표
정밀도	양성이라 답한 것 중 실제 양성의 비율
재현율	실제 양성 중 찾아낸 비율
F_1	정밀도와 재현율의 조화평균
ROC 곡선	임계값을 옮기며 TPR과 FPR을 이은 곡선

평가지표는 문제의 목적에 맞춰 고른다



결과를 본 뒤 유리한 지표로 바꾸지 않도록 평가 기준을 먼저 정한다

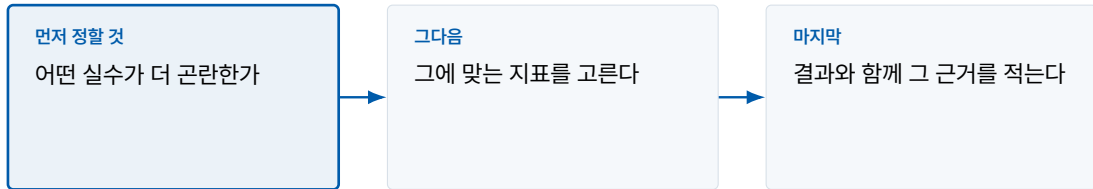
한 혼동행렬에서 여러 지표를 보고한다

가상 자료, 양성 20개, 음성 980개, 임계값 0.5

지표	계산
정확도	$(12 + 977)/1,000 = 0.989$
정밀도	$12/(12 + 3) = 0.80$
재현율	$12/(12 + 8) = 0.60$
F_1	$24/(24 + 3 + 8) \approx 0.69$

임계값 하나의 혼동행렬만으로 ROC AUC를 계산할 수는 없다

모델의 평가 기준을 먼저 정한다



모델을 평가할 지표는 결과를 보기 전에 정한다

Checkpoint

Threshold and ROC

- 임계값을 내리면 FN과 FP는 각각 어떻게 되는가
- 곡선 아래 면적이 0.5면 무슨 뜻인가
- 같은 FPR에서 실제 음성 수가 늘면 FP 건수는 어떻게 되는가

Why Would You Fit the Data Less Well on Purpose?

Next Lecture: Regularization and Bias-Variance

- 예측 오차를 편향·분산·잡음으로 분해하기
- 계수의 크기에 벌점 매기기
- 벌점의 세기를 검증으로 고르기