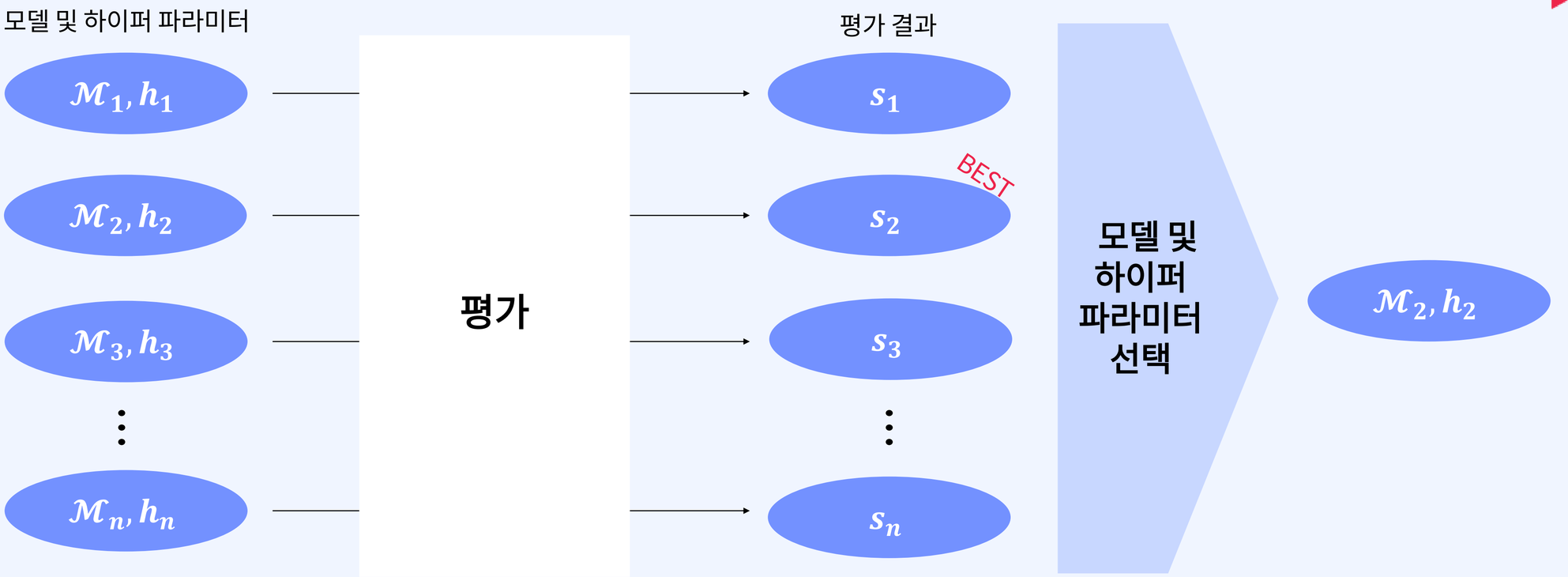
5. 모델 평가

1 머신러닝 자동화에서의 모델 평가의 필요성

모델 및 하이퍼 파라미터 선택의 기준

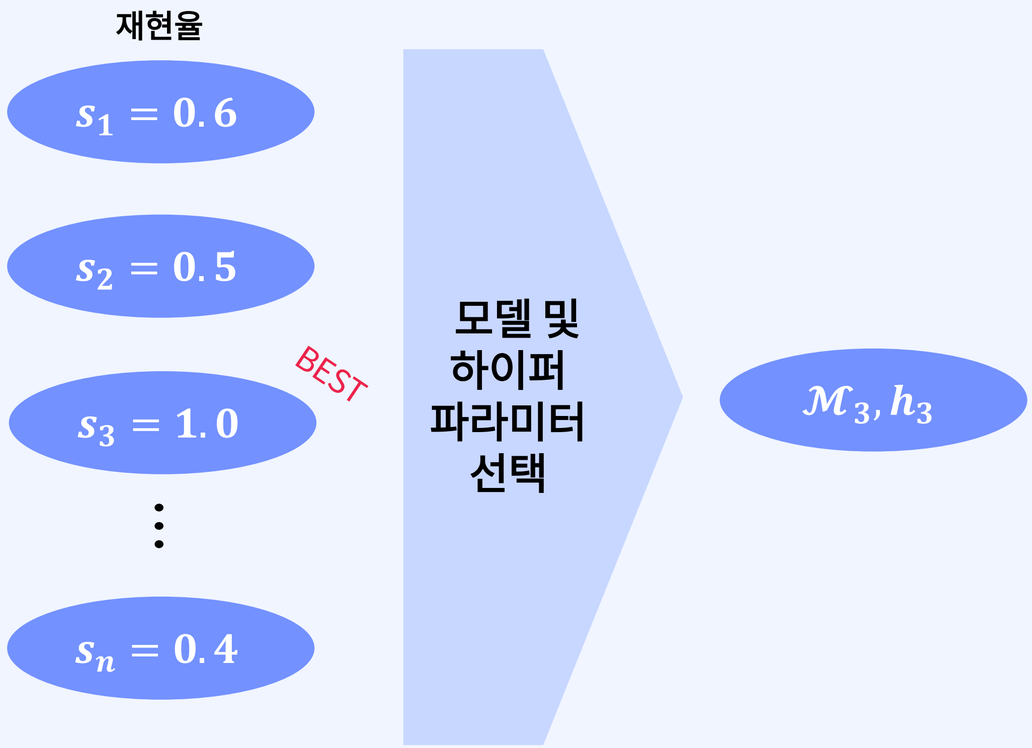
머신러닝 자동화 시스템은 다양한 모델 및 하이퍼 파라미터를 평가 지표를 기준으로 비교하여 최종 모델을 선택합니다.



모델과 하이퍼 파라미터를 객관적으로 평가하고 그 결과를 바탕으로 최종 모델 및 하이퍼 파라미터를 선택할 때 평가 지표가 필요함

평가 지표를 잘못선택하면?

부적절한 평가 지표를 사용하면 편향된 모델을 선택할 수도 있습니다.



$\mathcal{M}_3, h_3$

- 모든 샘플을 긍정으로 분류하는 모델과 하이퍼 파라미터
- 모든 긍정 샘플을 정 분류했으므로 재현율은 1
- 그러나 모든 부정 샘플을 오 분류했으므로 정확도와 정밀도는 매우 낮음

5. 모델 평가

2 분류 모델 평가

혼동 행렬

혼동 행렬은 분류 모델이 예측한 라벨과 실제 라벨을 비교해 보여주는 행렬로 분류 모델을 평가하는 지표를 계산할 때 자주 활용합니다.

이진 분류 모델 평가를 위한 혼동 행렬

		실제	
		긍정	부정
예측	긍정	참 긍정 (True Positive; TP)	거짓 긍정 (False Positive; FP)
	부정	거짓 부정 (False Negaitive; FN)	참 부정 (True Negative; TN)

- **참 긍정**: 긍정으로 예측한 샘플 가운데 실제 긍정인 샘플 수
- **거짓 긍정**: 긍정으로 예측한 샘플 가운데 실제 부정인 샘플 수
- **참 부정**: 부정으로 예측한 샘플 가운데 실제 부정인 샘플 수
- **거짓 부정**: 부정으로 예측한 샘플 가운데 실제 긍정인 샘플 수

"참"은 정 분류했다는 의미이며, "거짓"은 오 분류했다는 의미입니다. 뒤에 따라오는 "긍정(부정)"은 예측을 "긍정(부정)"으로 했다는 의미입니다.

정확도

정확도는 모든 샘플 가운데 정 분류된 샘플의 비율로 직관적이라는 장점 때문에 가장 널리 사용됩니다.

공식

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

한계

클래스 불균형 문제가 있는 모델을 좋게 평가할 위험이 있음

(예시)

		실제	
		긍정	부정
예측	긍정	0	0
	부정	1	9999

- 정확도: 99.99% (10,000개 가운데 9,999개를 정 분류함)
- 그러나 모든 샘플을 부정으로 분류한 것이므로 좋은 모델이라 보기 힘듦

정밀도

정밀도는 모델이 긍정이라 예측한 샘플 가운데 실제 긍정인 샘플의 비율로 계산하며, 사람의 공수와 관련이 있습니다.

공식

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

해석: 정밀도가 높을수록 사람의 공수가 줄어드는 경향이 있음

- 일반적으로 긍정이라 분류된 샘플에 대해서는 추가 작업을 함

(예시 1) 암 환자라 분류한 사람만 재검사함

(예시 2) 불량품이라 판단한 제품만 재검사, 재작업, 폐기 등을 수행함

- 모델의 정밀도가 낮다면 긍정이라 예측했는데 실제 긍정일 가능성이 작으므로 불필요한 작업을 하게 될 가능성이 큼

(예시) 불량이라 판단했는데 실제로는 양품이었다면 재검사, 재작업 등 불필요한 작업을 하게 됨

한계: 긍정 클래스에 대해 보수적인 모델을 좋게 평가함

- 예측하기 어려운 샘플 대부분을 부정으로 분류하면 정밀도가 커짐
- 따라서 정밀도만 사용하면, 확실히 긍정으로 분류할 수 있는 샘플만 긍정으로 분류하는 모델을 좋게 평가함

재현율

재현율은 실제 긍정인 샘플 가운데 긍정이라 분류된 샘플의 비율로 계산하며, 검출력과 관련이 있습니다.

공식

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

해석: 실제 긍정인 샘플을 얼마나 잘 검출해내는지를 나타냄

- 실제 긍정 가운데 긍정이라 예측된 샘플의 비율임
- 긍정의 비율이 낮을 때(클래스 불균형 문제가 있을 때) 사용하기 적합함

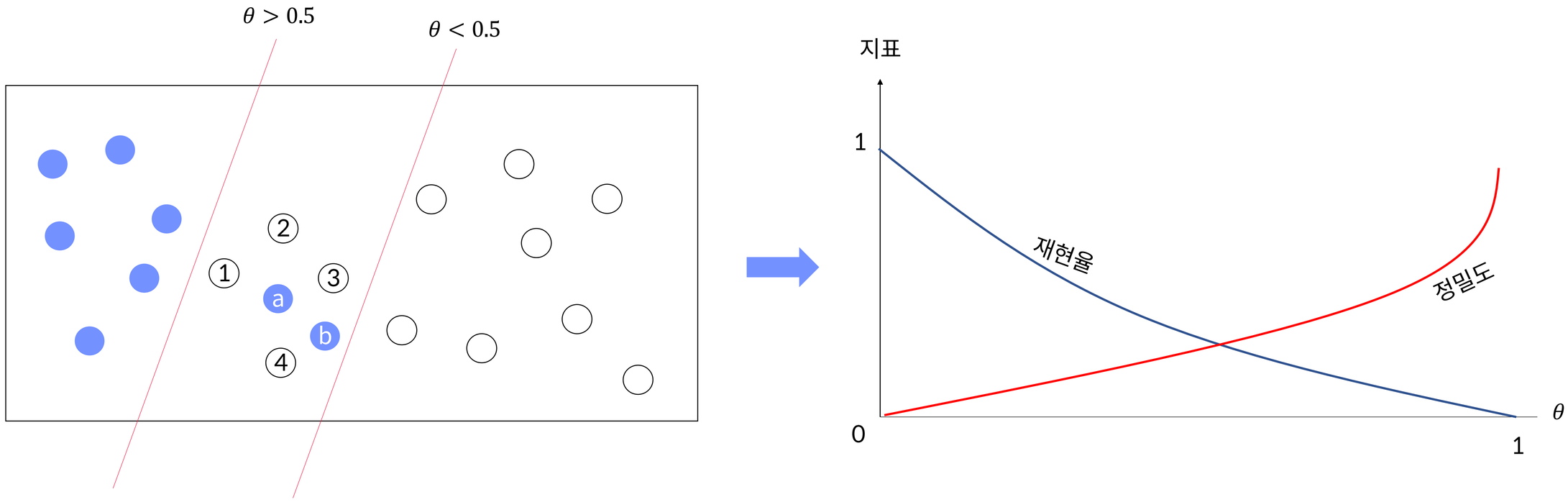
한계

- 실제 긍정만 고려하고 실제 부정과 예측 긍정 등을 고려하지 않음
- 모든 샘플을 긍정이라 분류하면 당연히 실제 긍정 샘플 모두가 완벽하게 검출돼 재현율이 1이 되지만, 사실 아무런 의미 없는 분류임

정밀도와 재현율 간 관계

일반적으로 긍정이라 판단하기 위한 임계치가 커질수록 정밀도는 증가하나 재현율은 감소하고, 임계치가 작아질수록 정밀도는 감소하나 재현율은 증가하는 반비례 관계가 있습니다.

$$\hat{y} = \begin{cases} 1, & \text{if } \Pr(y|\boldsymbol{x}) > \theta \\ 0, & otherwise \end{cases}$$



F1-점수

F1-점수는 정밀도와 재현율의 조화 평균으로 계산하며, 크게 편향되지 않는 지표입니다.

공식

$$F_1 = 2 \times \frac{Pre \times Rec}{Pre + Rec}$$

장단점

- 정밀도와 재현율은 단독으로 사용했을 때 문제가 있고 두 지표는 반비례한다는 점을 잘 고려한 지표임
- 다른 지표에 비해 잘못된 분류를 좋게 평가할 위험이 적음
- 직관적인 해석이 어려움
- 정밀도와 재현율 가운데 어느 지표가 커서 F1-점수가 큰 지 알 수 없음

다중 분류로 확장
: 혼동 행렬

다중 분류란 라벨이 세 개 이상의 값을 갖는 분류를 의미하며, 보통은 긍정 클래스와 부정 클래스 개념이 없습니다.

		실제		
		A	B	C
예측	A	$n_{A,A}$	$n_{A,B}$	$n_{A,C}$
	B	$n_{B,A}$	$n_{B,B}$	$n_{B,C}$
	C	$n_{C,A}$	$n_{C,B}$	$n_{C,C}$

- $n_{x,y}$: 실제 라벨이 x인 샘플 가운데 y로 예측된 샘플 수
- 정확도: $\frac{n_{A,A}+n_{B,B}+n_{C,C}}{n_{A,A}+n_{A,B}+n_{A,C}+n_{B,A}+n_{B,B}+n_{B,C}+n_{C,A}+n_{C,B}+n_{C,C}}$
- 긍정과 부정 클래스의 개념이 없으므로 정밀도, 재현율, F1 점수를 직접 계산할 수 없음

다중 분류로 확장
: 지표 계산

다중 분류란 라벨이 세 개 이상의 값을 갖는 분류를 의미하며, 보통은 긍정 클래스와 부정 클래스 개념이 없습니다.

한 클래스에 대한 지표 계산

		실제		
		A	B	C
예측	A	$n_{A,A}$	$n_{A,B}$	$n_{A,C}$
	B	$n_{B,A}$	$n_{B,B}$	$n_{B,C}$
	C	$n_{C,A}$	$n_{C,B}$	$n_{C,C}$

- 다중 분류 결과를 평가할 때는 각 클래스를 긍정으로 간주했을 때의 정밀도, 재현율, F1 점수를 계산함
(초록색: TP, 주황색: FP, 노랑색: FN, 회색: TN)
- 클래스 A에 대한 재현율: $rec_A = \frac{n_{A,A}}{n_{A,A}+n_{B,A}+n_{C,A}}$

평균지표 계산

- 그런데 클래스마다 지표가 있으면 평가 결과를 직관적으로 이해하기 어려움
- 이러한 이유로 각 클래스를 긍정으로 간주하여 계산한 지표의 평균을 사용하는데, 평균을 계산하는 방법에 따라 매크로(macro)와 가중(weighted) 평균로 구분됨

매크로: 산술 평균 $macro - pre = \frac{pre_A + pre_B + pre_C}{3}$

가중 평균
(가중치: 클래스별 샘플 수) $\text{weighted} - pre = \frac{n_A \times pre_A + n_B \times pre_B + n_C \times pre_C}{n_A + n_B + n_C}$ ($n_X = n_{A,X} + n_{B,X} + n_{C,X}$는 실제 클래스가 X(X = A, B, C)인 샘플 수)

5. 모델 평가

3 회귀 모델 평가

Mean Absolute Error

MAE는 절대 오차의 평균으로 가장 직관적인 지표 중 하나지만, 라벨의 스케일을 반드시 고려해야 합니다.

공식

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- y_i: i번째 샘플의 실제 라벨
- $\hat{y}_i$: i번째 샘플의 예측 라벨
- n: 샘플 수

해석

- 평균적으로 어느 정도의 오차를 내는지를 나타내므로 직관적으로 해석할 수 있음

(예시) MAE가 10인 모델은 예측 값과 실제 값의 평균적인 차이가 10이라는 뜻이므로, 예측 값이 40이라면 실제 값은 평균적으로 30과 50 사이에 있다고 할 수 있음

- MAE를 기준으로 모델을 평가할 때는 라벨의 스케일을 반드시 고려해야 함

(예시) 코스피 지수를 예측하는 모델의 MAE가 1,000이라면 매우 형편없는 모델이지만, 연도별 전 세계 인구수를 예측하는 모델의 MAE가 1,000이라면 완벽에 가까운 모델임

Root Mean Squared Error

RMSE는 평균 제곱근 오차로 이상치에 강건합니다.

공식

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- y_i: i번째 샘플의 실제 라벨
- $\hat{y}_i$: i번째 샘플의 예측 라벨
- n: 샘플 수

장단점

- RMSE는 평균 오차 제곱합에 루트를 씌운 것으로, 이 수치를 바탕으로 평균 어느 정도의 오차를 낸다고 해석할 수 없음
- RMSE는 MAE에 비해 매우 이상치가 강건함. 즉, 오차가 매우 큰 샘플이 있으면 MAE가 매우 커지지만, RMSE는 상대적으로 그렇지 않음

Mean Absolute Percentage Error

MAPE는 절대 퍼센트 오차의 평균으로 직관적으로 이해하기 좋지만, 몇몇 문제가 있어 잘 사용하지 않습니다.

공식

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

- y_i: i번째 샘플의 실제 라벨
- $\hat{y}_i$: i번째 샘플의 예측 라벨
- n: 샘플 수

장단점

- 직관적으로 이해할 수 있음
- 실제 라벨이 0인 샘플이 있으면 정의되지 않음
- 실제 값이 1 미만이라면 MAPE가 매우 커질 위험이 있음
- 실제 값보다 적게 예측했을 때 오차의 상한은 100%이지만, 크게 예측했을 때 오차의 상한은 없음