

mlads 3주차 정리

Evaluation techniques

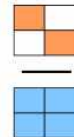
1. classification evaluation

Confusion Matrix

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN



Accuracy
(정확도)



Precision
(정밀도)



Recall
(재현도)

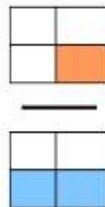


F1 Score

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

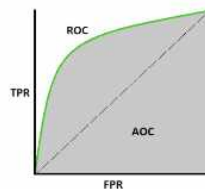
TPR = Recall

FPR = 1 - TPR =



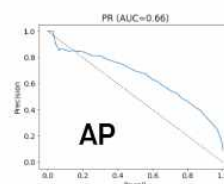
ROC

X축 : FPR
Y축 : TPR



PR

X축 : Recall
Y축 : Precision



Matthews Correlation Coefficient

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

데이터가 불균형해서 accuracy 값이 이상한 경우에 사용함

micro/macro precision, recall, F1 등이 있다.

F1의 경우엔 micro-average, weighted 도 존재

Kohen's Kappa

$$k = \frac{p_0 - p_e}{1 - p_e} = 1 - \frac{1 - p_0}{1 - p_e}$$

p0 : accuracy

pe : 우연히 맞혔을 확률

= (대각선에 있는 값 / 전체 개수)의 평균

2. Regression evaluation

(1)

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - f(\mathbf{x}_i)|$$

(2) 아웃라이어에 약함

$$MSE = \frac{\sum_{i=1}^n (y - \hat{y})^2}{n}$$

(3)

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

(4)

$$MPE = \frac{100}{n} \sum_{i=1}^n \frac{y - \hat{y}}{y}$$

(5)

$$SMAPE = \frac{100}{n} \times \sum_{i=1}^n \frac{|Y_i - \hat{Y}_i|}{(|Y_i| + |\hat{Y}_i|) / 2}$$

(6) feature가 많아질수록 커지는 경향이 있어 더 이상 사용 잘 안함

$$R^2 = 1 - \frac{SSE}{SST}$$

(7) 위 문제점 해결을 위해 도입

$$adj.R^2 = 1 - \frac{(n-1)}{(n-p)} (1 - R^2)$$

3. Clustering evaluation

실루엣 점수 - 클러스터 개수 판단에 사용, 데이터 개수가 많아지면 연산량 너무 높음

던 지수 - 간략하고 직관적

랜드 지수 - 가능한 모든 데이터 쌍의 개수에 대해 정답인 데이터 쌍의 개수의 비율 다만 무작위로 해도 좋은 점수가 나올 수 있음.

adj 랜드 지수 - 위 문제 해결을 위해 무작위 군집화에서 생기는 랜드지수의 기댓값을 원래의 값에서 빼서 기댓값과 분산을 재조정한 것이 조정 랜드지수다.

파울쿠스-말로우 지수 - 군집 간 유사성 확인에 사용