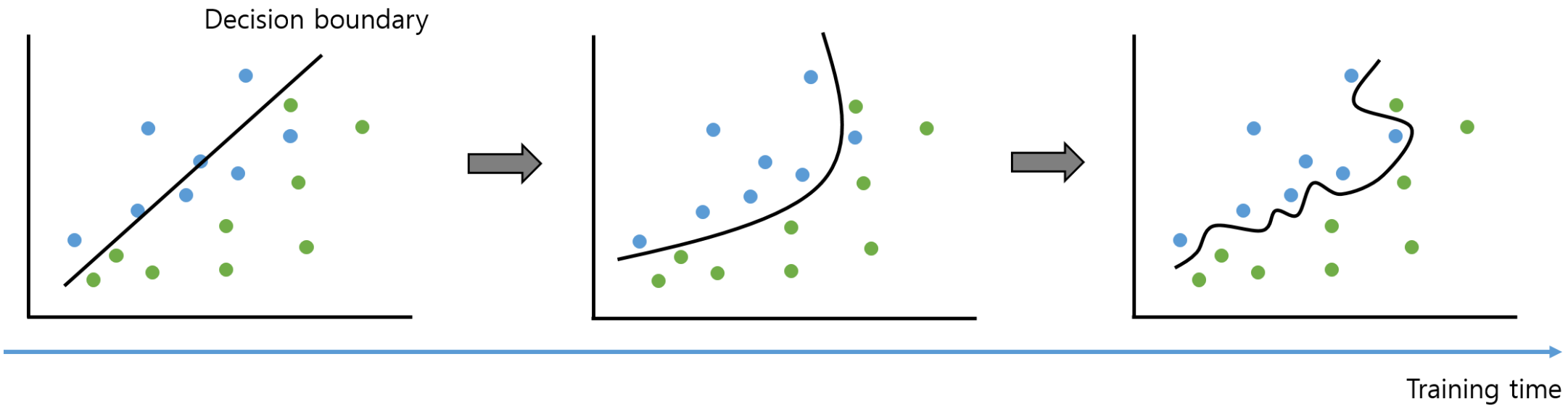
20 Popular Machine Learning Metrics.

Part 1: Classification & Regression Evaluation Metrics

contents

기초 개념 : 성능 평가 지표



- 실제값과 모델에 의한 **예측된 값을 비교**하여 두 값의 차이를 계산
- (실제값 - 예측값) = 0 일 경우 오차 = 0
- **과적합을 방지**하고 **최적의 모델**을 찾기 위해 성능 평가 실시
- **지도학습**에서 사용

Introduction : 평가 지표의 종류

- *Classification Metrics (accuracy, precision, recall, F1-score, ROC, AUC, …)*
- *Regression Metrics (MSE, MAE)*
- *Ranking Metrics (MRR, DCG, NDCG)*
- *Statistical Metrics (Correlation)*
- *Computer Vision Metrics (PSNR, SSIM, IoU)*
- *NLP Metrics (Perplexity, BLEU score)*
- *Deep Learning Related Metrics (Inception score, Frechet Inception distance)*

- 분류, 회귀, 랭킹, 통계, 컴퓨터 비전, NLP, 딥러닝으로 **평가 지표 범주화** 가능
- **모델의 상황**마다 적절한 평가 지표 필요

Shervin Minaee, 20 Popular Machine Learning Metrics. Part 1: Classification & Regression Evaluation Metrics, **2019**

Introduction : 평가 지표의 종류

- *Classification Metrics (accuracy, precision, recall, F1-score, ROC, AUC, …)*
- *Regression Metrics (MSE, MAE)*

part1

- *Ranking Metrics (MRR, DCG, NDCG)*
- *Statistical Metrics (Correlation)*
- *Computer Vision Metrics (PSNR, SSIM, IoU)*
- *NLP Metrics (Perplexity, BLEU score)*
- *Deep Learning Related Metrics (Inception score, Frechet Inception distance)*

- 모델이 **예측하는 분야**에 따라 분류와 회귀 구별
- 예측하고자 하는 결과가 범주인 경우 : **Classification**
- 숫자인 경우 : **Regression**

Shervin Minaee, 20 Popular Machine Learning Metrics. Part 1: Classification & Regression Evaluation Metrics, **2019**

Introduction : 평가 지표는 손실 함수와 다르다

- 손실 함수
  - 모델 성능을 **측정**하고 **훈련하는데 사용**(최적화)
  - 일반적으로 모델의 매개변수가 **미분 가능**
- 평가 지표
  - 모델 훈련 시 **성능 측정**
  - 미분할 필요가 없으나, 일부 미분 가능한 평가 지표의 경우 손실함수로 활용 가능(MSE 등)

Classification

- 과거의 데이터를 통해 입력 **데이터의 범주** 예측
  - 독립변수 : 공부 시간
  - 종속변수 : 합격 여부(합격/불합격)
- 얼굴 인식, 유튜브 비디오 분류, SNS 감정표현 탐지 등 다양한 산업 분야에서 활용되는 머신러닝 문제
- 분류 모델 종류 : SVM, 로지스틱 회귀, 결정 트리, 랜덤 포레스트, CNN, RNN 등

Classification : Confusion Matrix

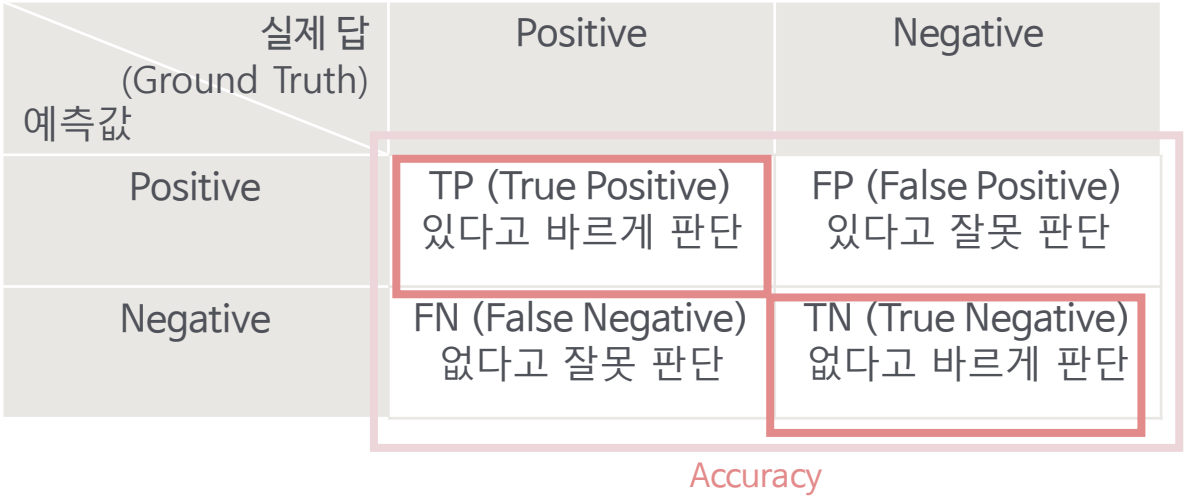
		Actual Class	
		Cat	Non-Cat
Predicted Class	Cat	90	60
	Non-Cat	10	940

실제 답 (Ground Truth) / 예측값	Positive	Negative
Positive	TP (True Positive) 있다고 바르게 판단	FP (False Positive) 있다고 잘못 판단
Negative	FN (False Negative) 없다고 잘못 판단	TN (True Negative) 없다고 바르게 판단

- 정오분류표의 행은 **예측값**, 각 열은 **정답값** 출력
- 고양이 이미지 100장
  - 90장을 정확하게 예측 : true positive
  - 10장을 잘못 예측 : false negative
- 고양이가 아닌 이미지 1000장
  - 940장을 정확하게 예측 : true negative
  - 60장을 잘못 예측 : false positive
- 대각선은 바르게 판단, 비대각선은 잘못 판단

Classification : 정확도

		Actual Class	
		Cat	Non-Cat
Predicted Class	Cat	90	60
	Non-Cat	10	940



- $\frac{\text{올바르게 예측한 물체의 수}}{\text{예측한 모든 물체}}$
- 1100개의 샘플 중 1030개가 올바르게 예측
- 정확도 : $\frac{90+940}{1000+10}$ = 93.6%
- 데이터의 분포가 불균형한 경우 **정확도의 신뢰도 하락**
  - 모델이 모든 샘플을 고양이가 아닌 것으로 예측하더라도 정확도 90.9%

Shervin Minaee, 20 Popular Machine Learning Metrics. Part 1: Classification & Regression Evaluation Metrics, **2019**

Classification : 정밀도

		Actual Class	
		Cat	Non-Cat
Predicted Class	Cat	90	60
	Non-Cat	10	940

실제 답 (Ground Truth) / 예측값	Positive	Negative
Positive	TP (True Positive) 있다고 바르게 판단	FP (False Positive) 있다고 잘못 판단
Negative	FN (False Negative) 없다고 잘못 판단	TN (True Negative) 없다고 바르게 판단

precision

- 모델이 **긍정으로 예측**한 물체 중에서 **실제로 정답**인 경우
- $\frac{\text{올바르게 예측한 물체의 수(TP)}}{\text{긍정으로 예측한 모든 물체 (TP+FP)}}$
- cat : $\frac{90}{90+60}$ = 60%
- NonCat : $\frac{940}{940+10}$ = 98.9%

Classification : 재현율

		Actual Class	
		Cat	Non-Cat
Predicted Class	Cat	90	60
	Non-Cat	10	940

실제 답 (Ground Truth) / 예측값	Positive	Negative
Positive	TP (True Positive) 있다고 바르게 판단	FP (False Positive) 있다고 잘못 판단
Negative	FN (False Negative) 없다고 잘못 판단	TN (True Negative) 없다고 바르게 판단

recall

- 실제 **정답값 중에서** 모델이 **올바르게 예측**한 비율
- $\frac{\text{올바르게 예측한 물체의 수(TP)}}{\text{실제 정답(TP+FN)}}$
- cat : $\frac{90}{90+10}$ = 90%
- NonCat : $\frac{940}{940+60}$ = 94%

Classification : 민감도/특이도

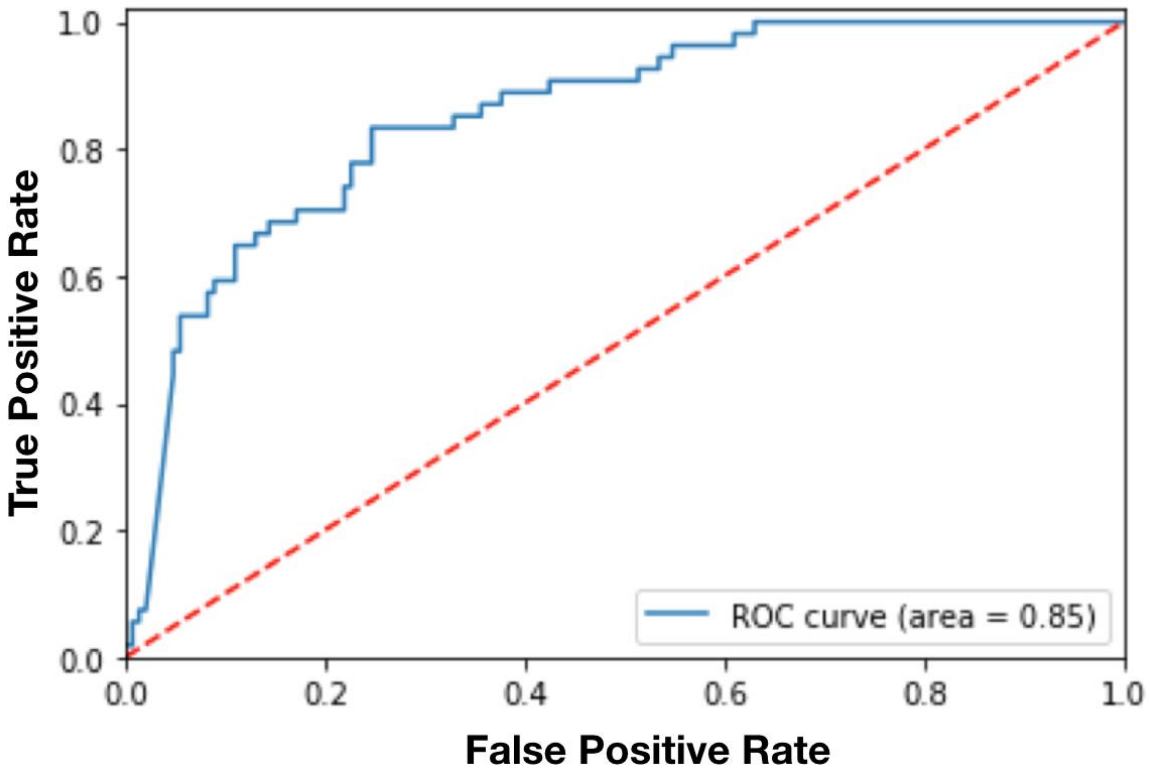
		Actual Class	
		Cat	Non-Cat
Predicted Class	Cat	90	60
	Non-Cat	10	940

실제 답 (Ground Truth) / 예측값	Positive	Negative
Positive	TP (True Positive) 있다고 바르게 판단	FP (False Positive) 있다고 잘못 판단
Negative	FN (False Negative) 없다고 잘못 판단	TN (True Negative) 없다고 바르게 판단

Specificity

- 의학 및 생물학 분야에서 주로 사용
- 민감도 = $\frac{\text{올바르게 예측한 물체의 수}(\text{TP})}{\text{실제 정답}(\text{TP}+\text{FN})}$ = 재현율
  - 실제로 심장 질환이 있는 사람 중 올바르게 분류
- 특이도 = $\frac{\text{올바르게 예측한 물체의 수}(\text{TN})}{\text{실제 오답}(\text{TN}+\text{FP})}$
  - 실제로 심장 질환이 없는 사람들 중 올바르게 분류

Classification : ROC

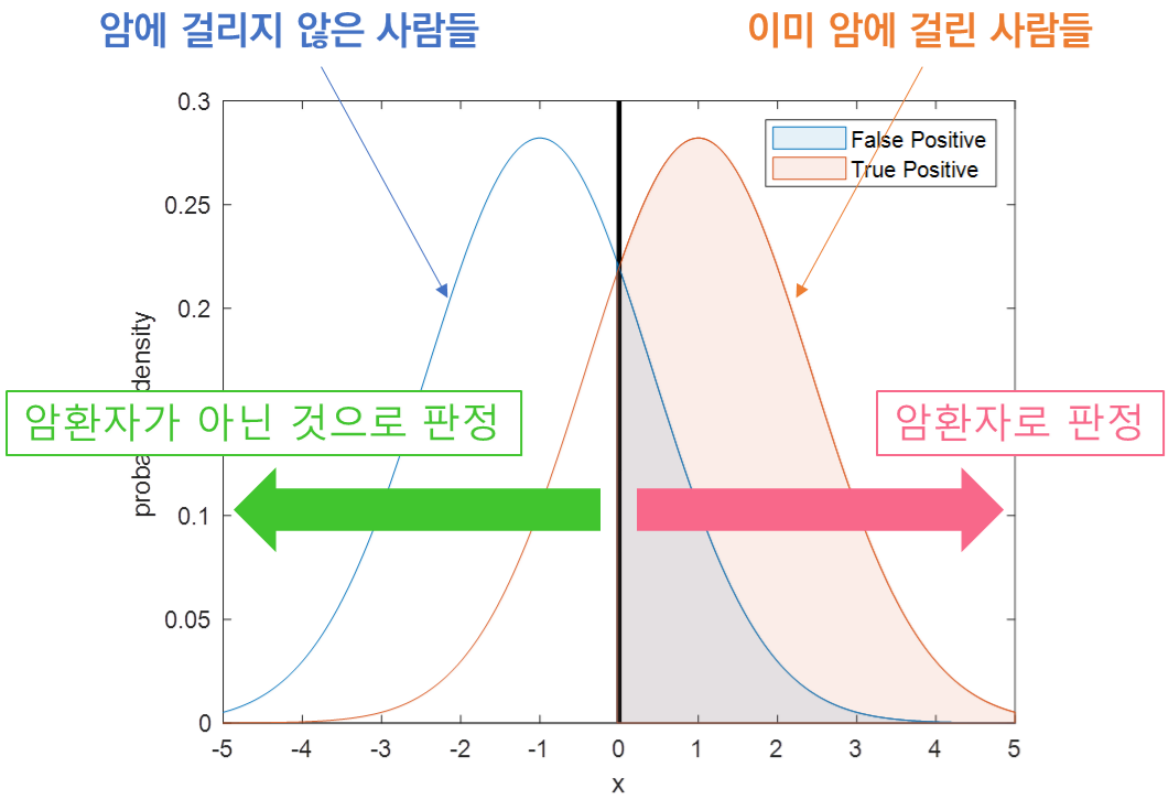


- True Positive와 False Positive를 이용한 곡선
- 곡선이 좌상단에 붙어있을수록 좋은 분류기 의미
  - 물체가 있다고 바르게 판단할 확률 높음
  - 물체가 있다고 잘못 판단할 확률 낮음

- decision threshold
  - 분류기가 판정한 값이 정답인지 아닌지 판정
  - decision threshold가 0.5인 경우, 분류기가 0.7을 산출했다면 정답, 0.4를 산출했다면 오답으로 처리
  - ROC 곡선을 통해 최적의 기준을 찾을 수 있음

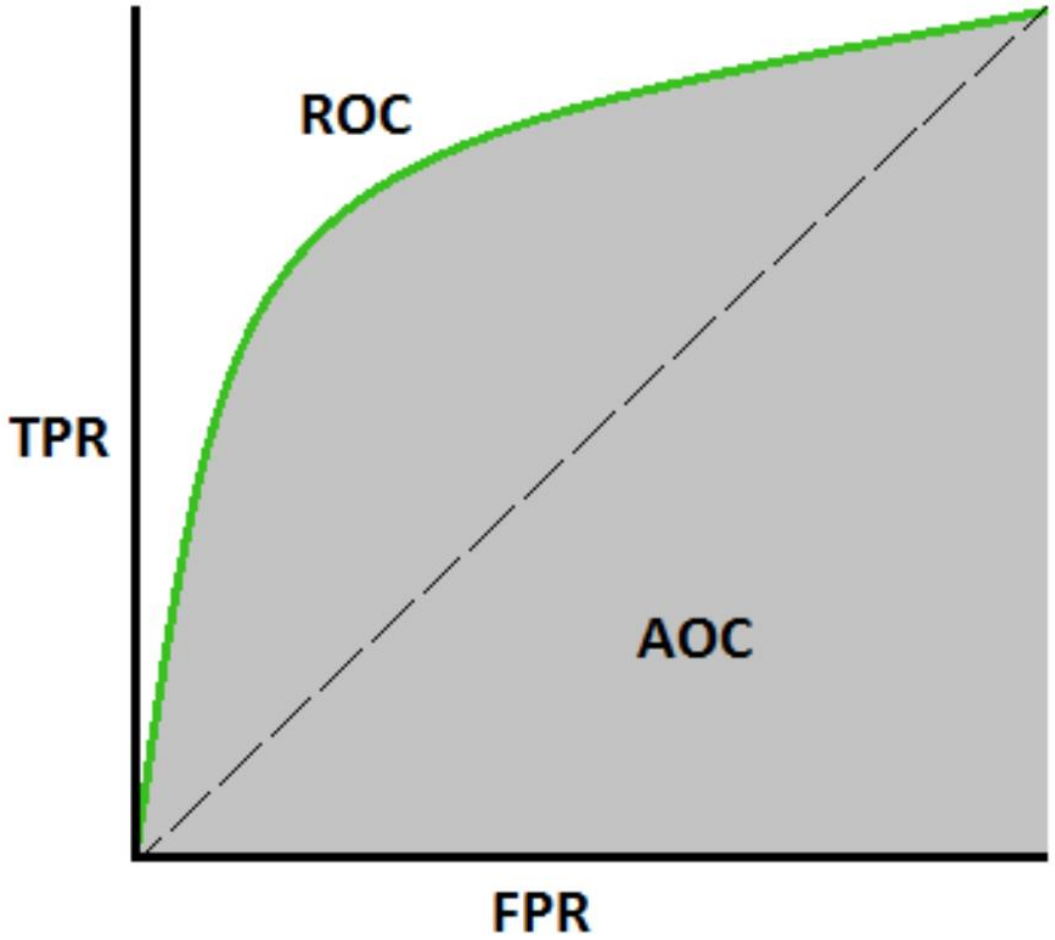
Shervin Minaee, 20 Popular Machine Learning Metrics. Part 1: Classification & Regression Evaluation Metrics, **2019**

Classification : ROC



- True Positive와 False Positive를 이용한 곡선
- 곡선이 좌상단에 붙어있을수록 좋은 분류기 의미
  - 물체가 있다고 바르게 판단할 확률 높음
  - 물체가 있다고 잘못 판단할 확률 낮음

Classification : AUC



- ROC 곡선의 밑넓이
- 예측이 얼마나 잘 평가되었는지 측정
  - 판정선이 얼마나 민감한지 나타냄
  - AUC가 높을 경우 클래스를 구별하는 모델의 성능이 훌륭하다는 것을 의미
- AUC의 면적이 80% 이상이면 훌륭한 성능을 가진 분류기라고 판단
- AUC의 면적이 50% 이하라면 쓸모없는 분류기라고 판단

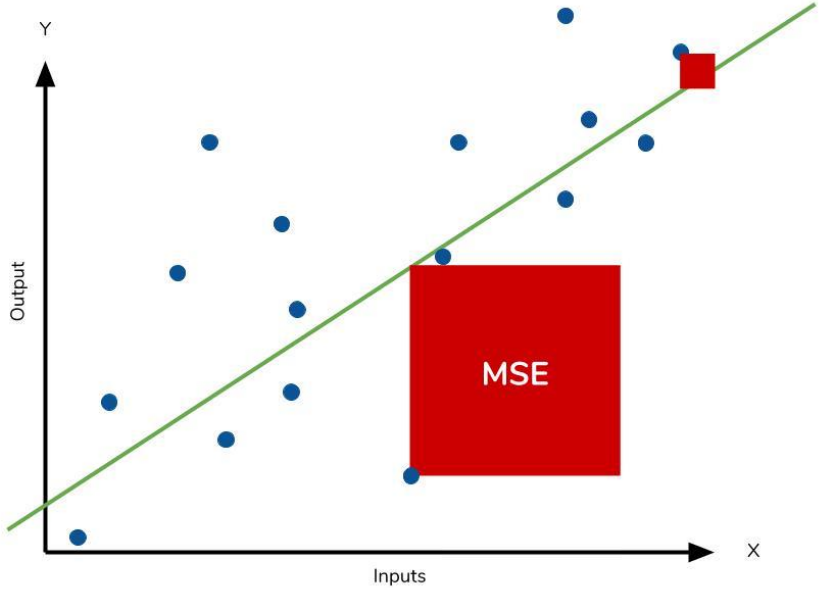
Shervin Minaee, 20 Popular Machine Learning Metrics. Part 1: Classification & Regression Evaluation Metrics, **2019**

Regression

- 과거의 데이터를 통해 입력 **데이터의 연속값** 예측
  - 독립변수 : 공부시간
  - 종속변수 : 시험점수
- 주택 가격 예측, 일기 예보, 주식 예측, 이미지 해상도 향상, 이미지 압축 등
- 회귀 모델 종류 : 선형회귀, CNN, RNN 등

Regression : MSE

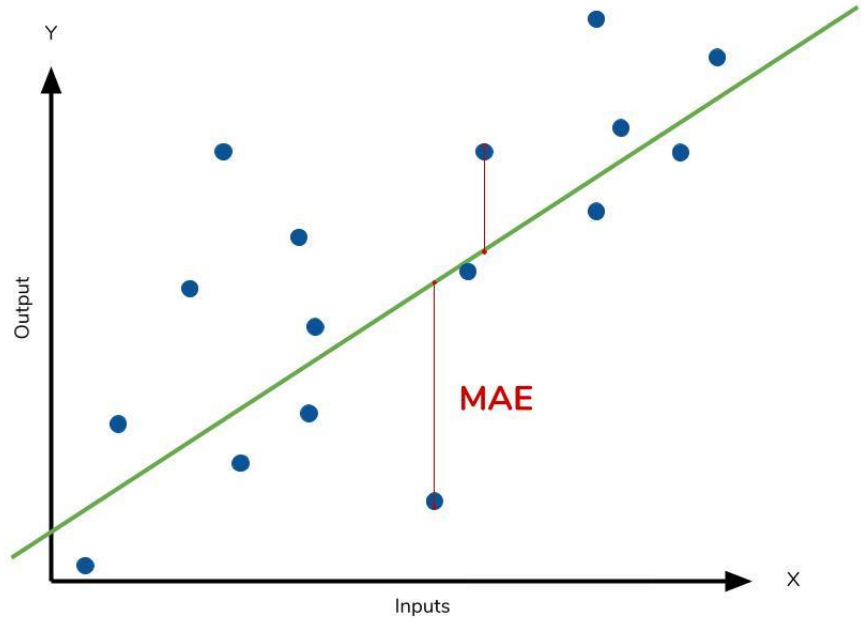
$$\mathrm{MSE} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})^2$$



- 예측 값과 실제 값 사이의 **평균 제곱 오차** 계산
  - 실제 값과 예측값의 차이를 제곱해 평균
  - 실제 값과 예측값 차이의 면적의 합
- MSE에 **제곱근을 사용**하여 모델을 평가하기도 함
- 특이값이 존재하면 수치가 많이 늘어남

Regression : MAE

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



- 예측값과 목표값 사이의 **평균 절대 거리**
  - 오차의 절대값의 평균
  - 상대적으로 특이값에 민감도 낮음

감사합니다