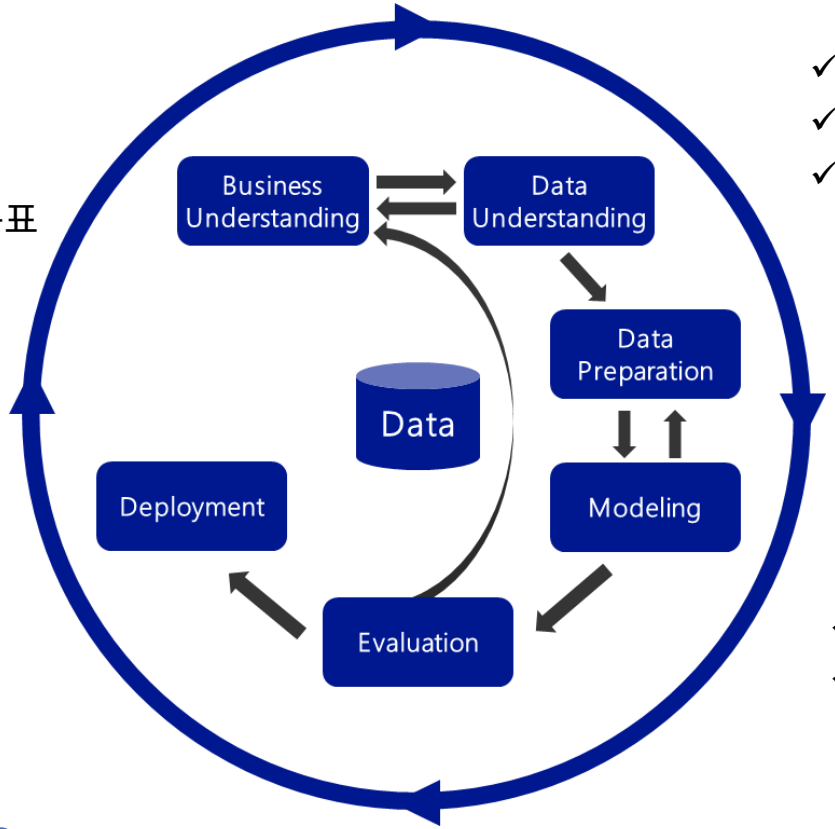
머신러닝 복습

전체 Process(CRISP-DM)

무엇이 문제인가?

✓ 비즈니스 **문제정의**
✓ 데이터분석 방향, 목표
✓ 초기 **가설 수립**

$$x \rightarrow y$$



✓ 원본식별
✓ **분석을 위한 구조 만들기**
✓ **데이터분석 EDA & CDA**

✓ 모델링을 위한
데이터 구조 만들기
- 모든 셀은 **값**이 있어야 한다.
- 모든 값은 **숫자** 여야 한다.
- (필요 시) 숫자의 **범위**가 일치

✓ 모델을 만들고
✓ 검증하기

✓ 모델 관리
✓ AI 서비스 구축

문제가 해결 되었는가?

✓ 기술적 관점 평가
✓ 비즈니스 관점 평가

알고리즘 한판 정리

	선형회귀	로지스틱회귀	KNN	SVM	Decision Tree	Random Forest	Gradient Boost (GBM, XGB, LGBM)
개념	✓오차를 최소화 하는 직선, 평면	✓오차를 최소화 하는 직선, 평면 ✓직선을 로지스틱 함수로 변환 (0~1 사이 값으로)	✓예측할 데이터와 train set과의 거리 계산 ✓가까운 [k개 이웃의 y]의 평균으로 예측	✓마진을 최대화 하는 초평면 찾기 ✓데이터 커널 변환	✓정보전달량 = 부모 불순도 - 자식 불순도 ✓정보 전달량이 가장 큰 변수를 기준으로 split	✓여러 개의 트리 ✓각각 예측 값의 평균 ✓행과 열에 대한 랜덤 : 조금씩 다른 트리들 생성	✓여러 개의 트리 ✓트리를 더해서 하나의 모델로 생성 ✓더해지는 트리는 오차를 줄이는 모델
전제 조건	✓NaN조치 ✓가변수화 ✓x들 간 독립	✓NaN조치 ✓가변수화 ✓x들 간 독립	✓NaN조치 ✓가변수화 ✓스케일링	✓NaN조치 ✓가변수화 ✓스케일링	✓NaN조치 ✓가변수화	✓NaN조치 ✓가변수화	✓NaN조치 ✓가변수화
성능	✓변수 선택 중요 ✓x가 많을 수록 복잡	✓변수 선택 중요 ✓x가 많을 수록 복잡	✓주요 hyper-parameter - n_neighbors : k 작을수록 복잡 - metric : 거리계산법	✓주요 hyper-parameter - C : 클수록 복잡 - gamma : 클수록 복잡	✓주요 hp - max_depth : 클수록 복잡 - min_samples_leaf : 작을수록 복잡	✓주요 hp 기본값으로도 충분! - n_estimators - max_features ✓기본값으로 생성된 모델 ==> 과적합 회피	✓주요 hp - n_estimators - learning_rate ✓XGB, LGBM : 과적합 회피를 위한 규제

회귀모델 평가

오차의 크기

$\hat{y}$: 예측값

오차 | 제곱 오차 | 절대값 오차 | 오차율

y	$\hat{y}$	$y-\hat{y}$	$(y-\hat{y})^2$	$\lvert y-\hat{y}\rvert$	$\left\lvert\frac{y-\hat{y}}{y}\right\rvert$
6	4				
5	6				
12	9				
2	2				

평균

MSE
RMSE

MAE

MAPE

평균 오차

평균 오차율

분류 모델 평가 : Confusion Matrix (복습)

✓일반적으로 이진 분류일때,

- 우리의 관심사 = 1, 양성 Positive
- 그 외 = 0, 음성 Negative

COVID 19		진단 결과	
		양성(Positive)	음성(Negative)
실제 감염여부	감염	True Positive	False Negative
	정상	False Positive	True Negative

✓ **정분류율**, **정확도**(Accuracy) = $\frac{(TP+TN)}{Total} \times 100$

✓ **민감도**(Sensitivity), **재현율**(Recall) = $\frac{TP}{(TP+FN)} \times 100$ (실제 독감에 감염된 자들 중 양성이라고 맞춘 비율)

✓ **정밀도**(Precision) = $\frac{TP}{(TP+FP)} \times 100$ (양성이라고 예측한 자들 중 실제 양성인(맞춘) 비율)

분류 모델 평가 : classification report (복습)

```
1 # confusion matrix
2 print(confusion_matrix(val_y, val_pred))
```

```
[[80 11]
 [19 33]]
```

```
1 # classification report
2 print(classification_report(val_y, val_pred))
```

	precision	recall	f1-score	support
0	0.81	0.88	0.84	91
1	0.75	0.63	0.69	52
accuracy			0.79	143
macro avg	0.78	0.76	0.76	143
weighted avg	0.79	0.79	0.79	143

✓ **f1-score**

- precision과 recall의 조화 평균
  - 예) 갈 때 60km/h, 올 때 80km/h로 주행했을 때, 평균 속력은?
- 수식 : 역수의 산술평균의 역수 = $\frac{2}{\frac{1}{\text{precision}}+\frac{1}{\text{recall}}}$

✓ **macro avg : 산술평균**

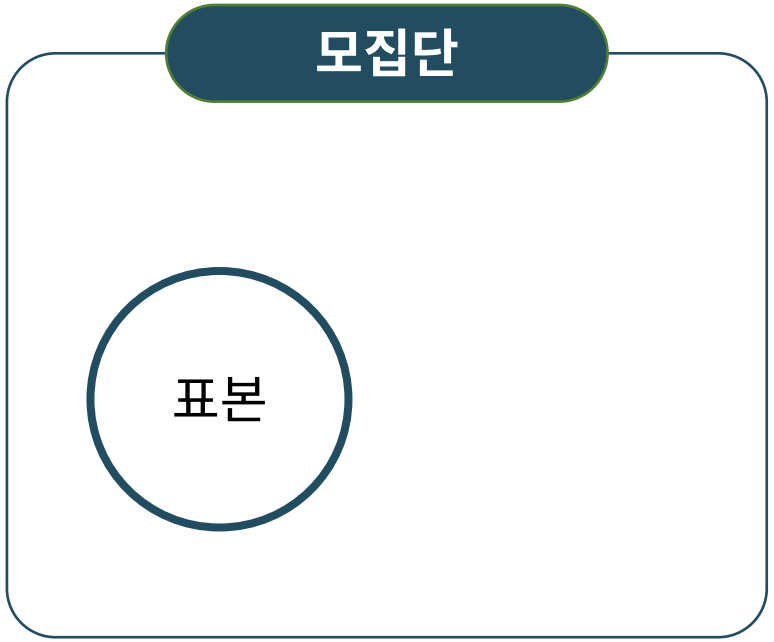
✓ **weighted avg : 가중평균**

✓ **support : 데이터 개수**

모델링의 목적

✓모델링의 목적

- 학습용 데이터에 있는 패턴으로,
  그 외 데이터(모집단 전체)를 적절히 예측
- 학습한 패턴(모델)은,
  - 학습용 데이터를 잘 설명할 뿐만 아니라,
  - 모집단의 **다른 데이터(val, test)도 잘 예측**해야 함

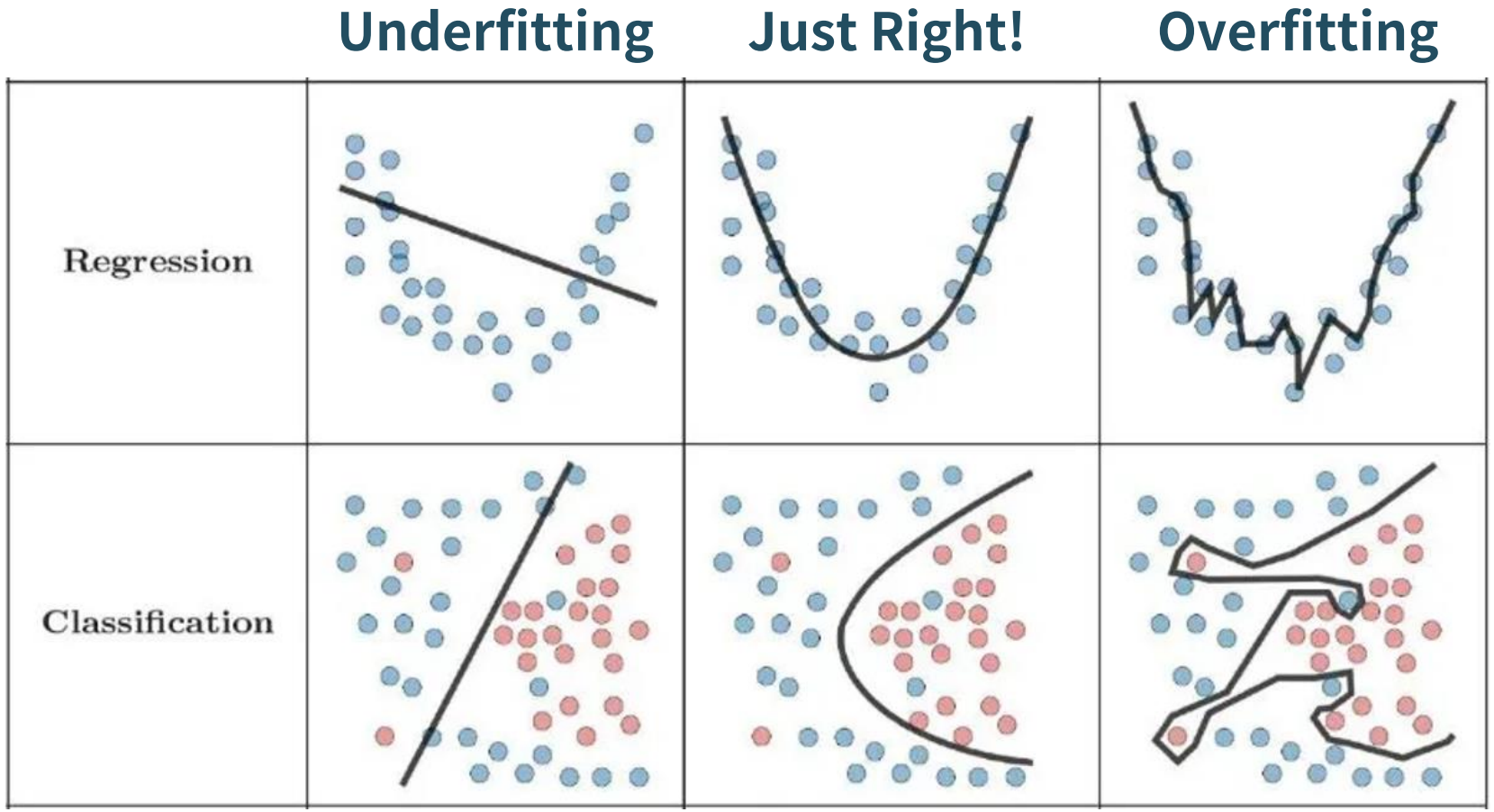


✓모델의 복잡도

- **너무 단순한** 모델 : train, val 성능이 떨어짐
- **적절히 복잡한** 모델 : 적절한 예측력
- **너무 복잡한** 모델 : train 성능 높고, val 성능 떨어짐

Underfitting과 Overfitting

✓ 모델(알고리즘)마다 복잡도를 결정하는 요인이 있음.



성능 최적화와 과적합의 관계

✓모델의 **복잡도** : 학습용 데이터의 패턴을 반영하는 정도

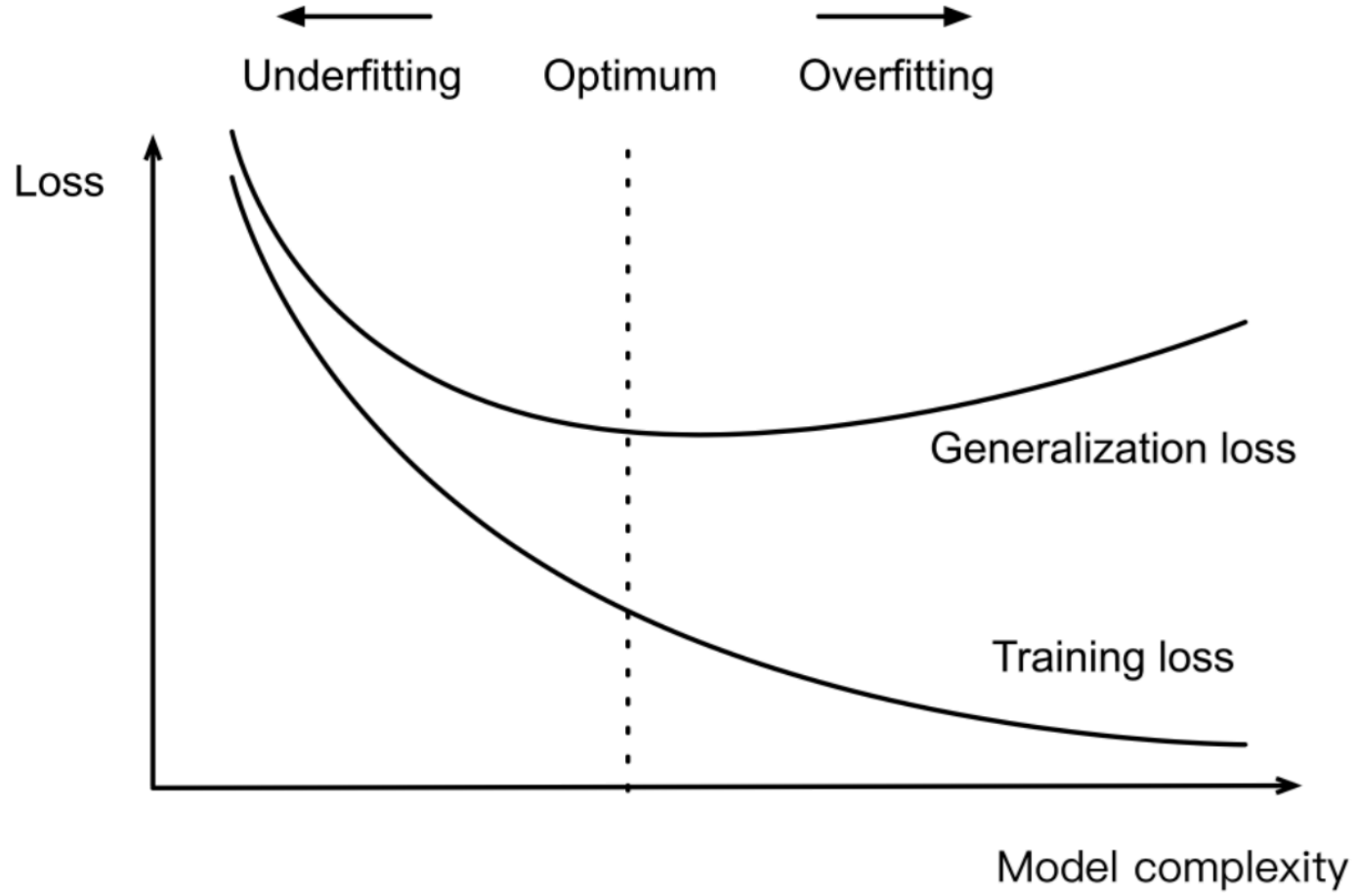


그림 출처 : https://vitalflux.com/model-complexity-overfitting-in-machine-learning/