




Validation & Evaluation

KUBIG 학술부



1. Validation은 왜 필요할까?

- 모델에는 parameter / hyperparameter이 존재.
- Parameter은 데이터로 estimate 할 수 있는 값
- Hyperparameter은 연구자가 직접 설정하는 값

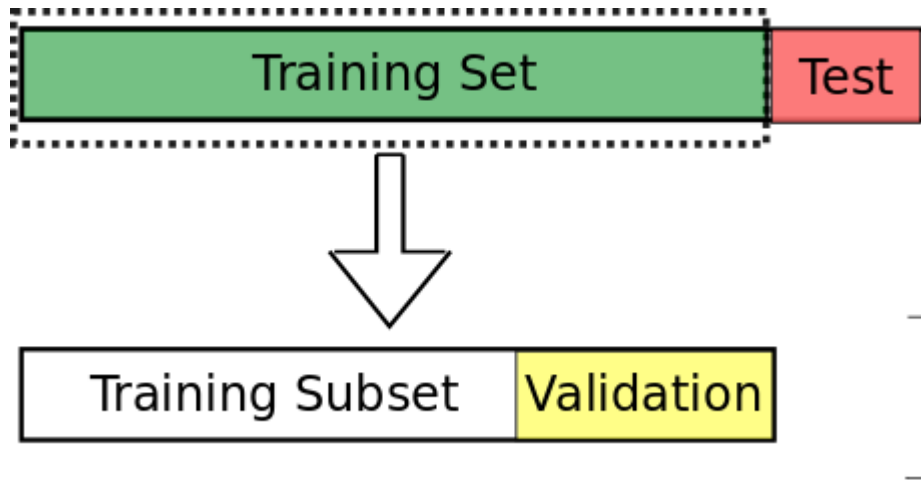


Hyperparameter을 적절하게 설정해 주기 위해
validation의 과정이 필요함!

2. Validation의 종류

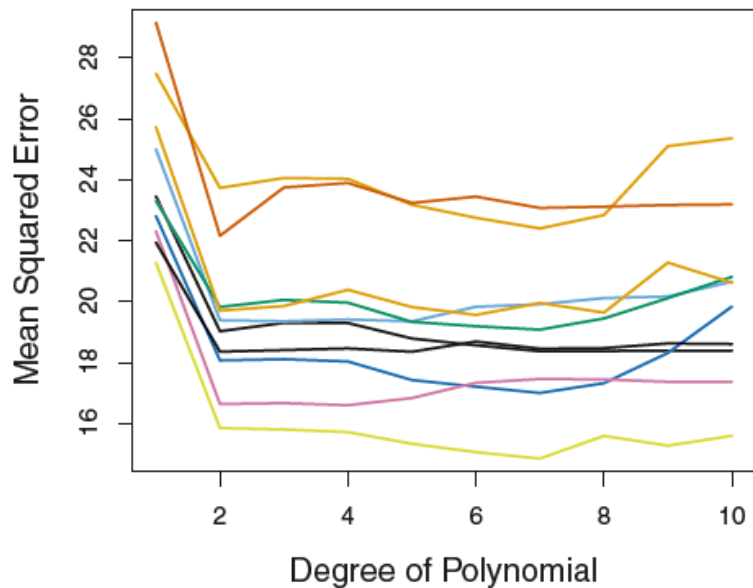
1. Validation set approach
2. LOOCV
3. K-Fold Cross-Validation

2-1. Validation set approach



0. 데이터를 Training set과 Test set으로 나눈다.
1. Training set을 다시 Training Subset과 Validation set으로 한번 나눈다.
2. Training Subset으로 학습한 모델이 Validation set에서 가장 좋은 성능을 내는 hyperparameter를 선택

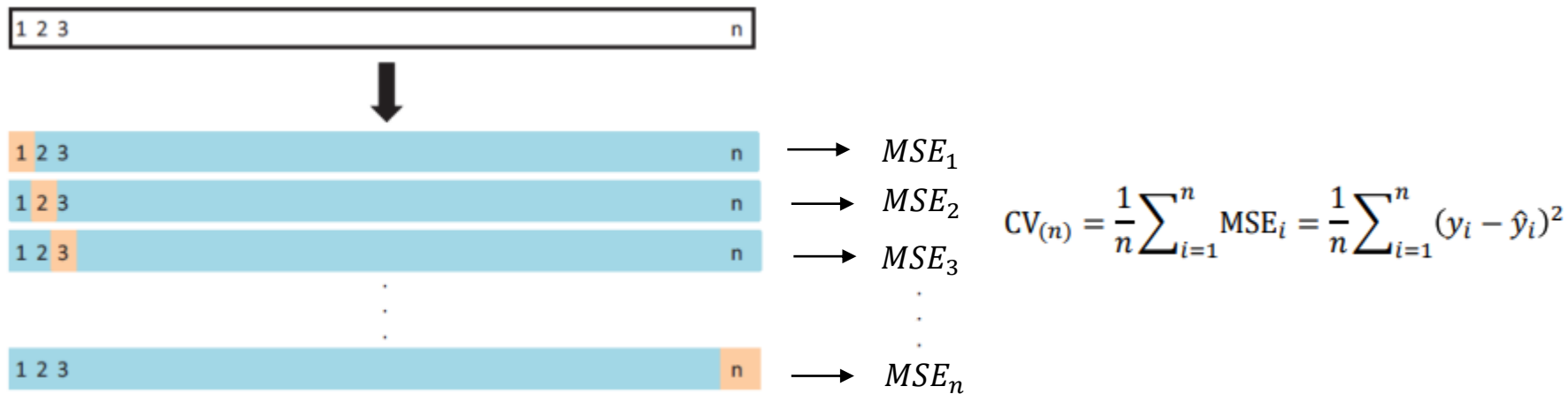
2-1. Validation set approach의 단점



Validation을 한 번만 진행함
→ 다른 방법에 비해 신뢰도가 떨어짐,
→ 분류 방법에 따라 다양한 mse 추정량이 나옴.

출처 : An introduction to statistical learning with Applications in R

2-2. LOOCV(leave-one-out-cross validation)



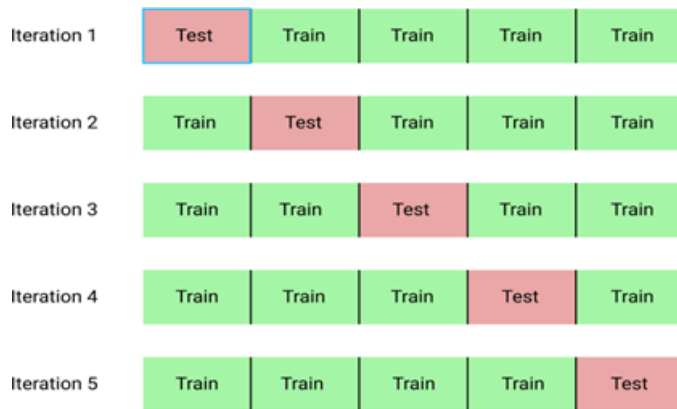
1. 관측치가 n 개인 training set을 관측치가 1개인 validation set과 관측치가 $(n-1)$ 개인 training subset으로 나눈다.
2. 각각의 관측치들이 한번씩 validation set이 되도록 1번 과정을 n 번 반복하며 MSE를 계산한다.
3. 이 n 개 metric의 평균을 기준으로, 가장 좋은 성능을 보여주는 hyper parameter를 선택한다.

2-2. LOOCV의 장점

[Validation Set approach에 비해]

- 훨씬 작은 bias를 갖는다. → overfitting의 위험성이 줄어든다.
- 항상 같은 validation error를 내기 때문에, 신뢰성이 높다.

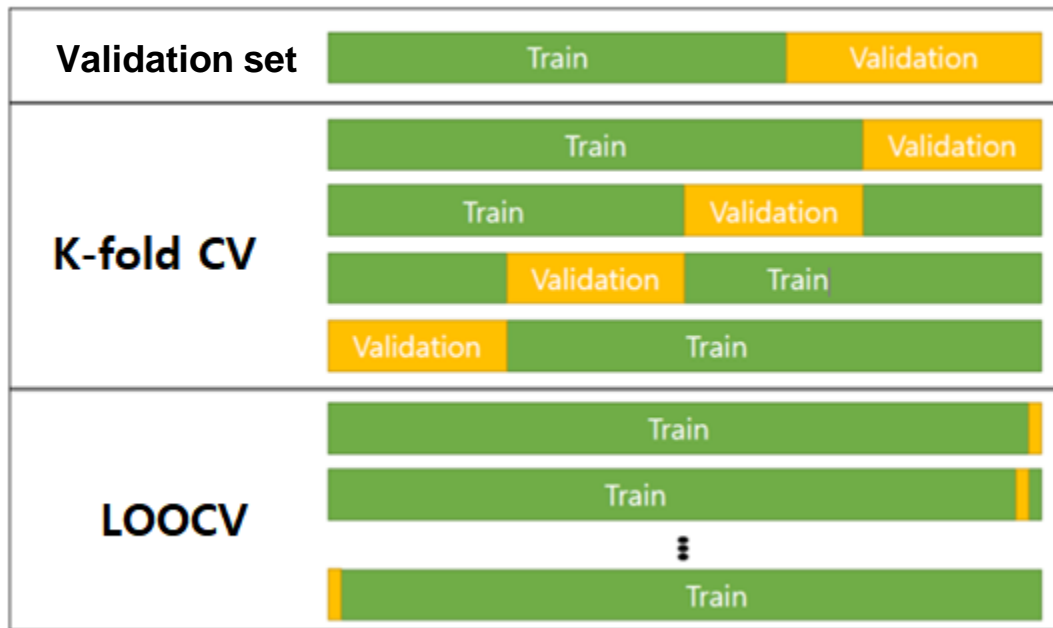
3. k-fold cross validation



$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^k MSE_i.$$

1. Training set을 동일한 크기의 k개의 그룹으로 나눈다.
2. K개의 그룹 중 (k-1)개의 그룹에 해당하는 data들을 training set, 나머지 1개의 그룹에 해당하는 data들을 validation set으로 한다.
3. 모든 그룹이 한번씩 validation set이 되도록 2번 과정을 k번 반복하며 K개의 MSE를 구한 후 평균을 내서 가장 좋은 성능을 보여주는 hyperparameter를 선택한다.

2-1, 2-2, 2-3 비교



3. Bias & Variance

- 정확도를 측정하기 위한 가장 일반적인 방법: MSE $MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2$
- 이 때, MSE는 크게 세 가지 부분으로 나눌 수 있다.

$$MSE = E[(y - \hat{y})^2] = \boxed{Var(\hat{y})} + \boxed{Bias(\hat{y})^2} + \boxed{Var(\epsilon)}$$

Reducible error

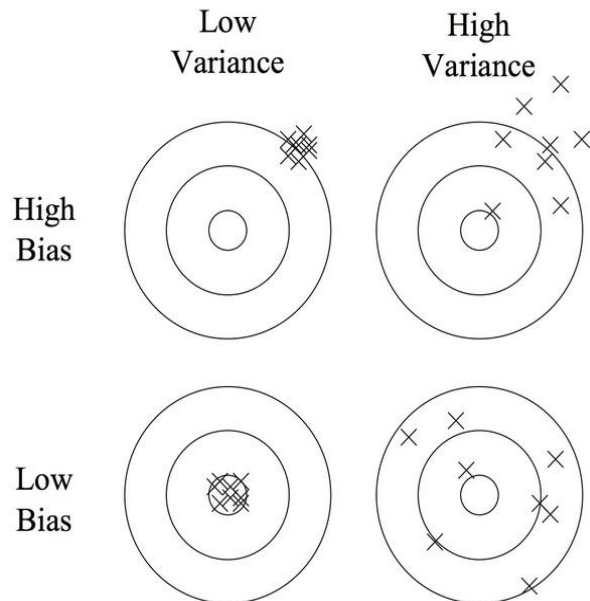
Noise

3. Bias & Variance

- reducible error

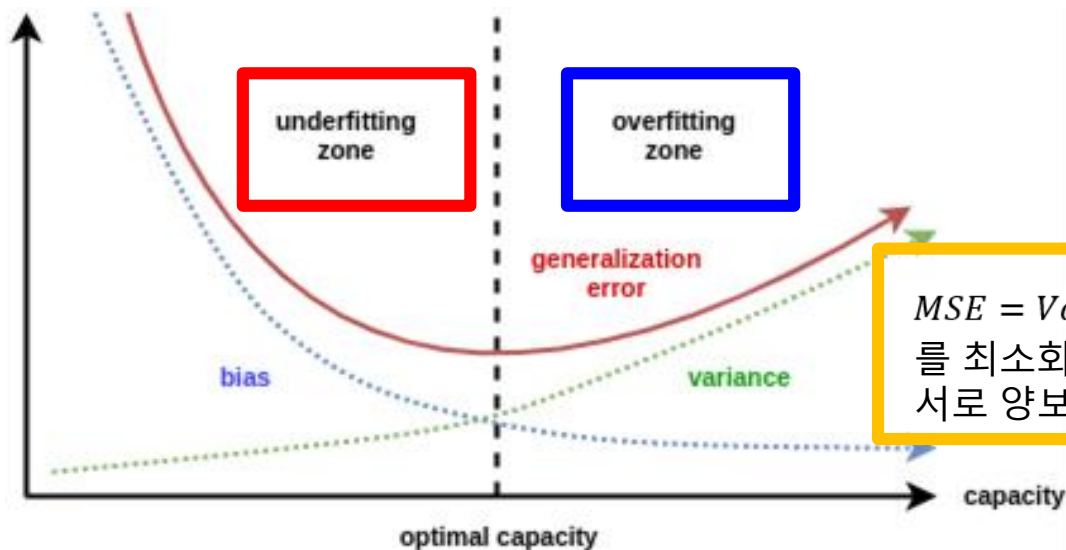
1. Bias(편향) : 참 값과 추정 값의 차이
2. Variance(분산) : 추정 값의 산포도

오른쪽의 그림에서 알 수 있듯이, Low Bias, Low Variance를 가지는 model을 만드는 것이 가장 좋음.



3-2. Bias & Variance Trade off

하지만, bias-variance는 tradeoff 관계에 있다.



$$MSE = Variance + Bias^2$$

를 최소화하는 선에서 분산/편향이
서로 양보하며 최적화된 값을 찾아야!

4. Evaluation metric – 4.1 Classification

	PREDICTED CLASS		
		Class=Yes	Class=No
	ACTUAL CLASS	Class=Yes	Class=No
		Class=No	Class=No
		a	b
		c	d

a : TP (True Positive)
b : FN (False Negative)
c : FP (False Positive)
d : TN (True Negative)

- Classification에서 가장 많이 쓰이는 metric 중 하나이다.
- True positive (TP) : 실제값이 positive 일때 모델이 positive로 예측한 데이터 개수
- False positive (FP) : 실제값이 positive이나 모델이 negative로 예측한 데이터 개수
- False negative (FN) : 실제값이 negative이나 모델이 positive로 예측한 데이터 개수
- False positive (FP) : 실제값이 negative이고 모델이 negative로 예측한 데이터 개수

4. Evaluation metric – 4.1 Classification

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error
	Negative	False Positive (FP) Type I Error	True Negative (TN)

Accuracy	모델이 결과값을 맞게 검출한 비율. 가장 자주 쓰이는 metric	$(TP+TN)/(TP+TN+FP+FN)$
Sensitivity / Recall / True positive rate (TPR)	실제값이 positive인 데이터 중 맞게 검출한 비율	$TP/(TP+FN)$
Specificity / Selectivity / True negative rate (TNR)	실제값이 negative인 데이터 중 맞게 검출한 비율	$TN/(TN+FP)$

4. Evaluation metric – 4.1 Classification

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error
	Negative	False Positive (FP) Type I Error	True Negative (TN)

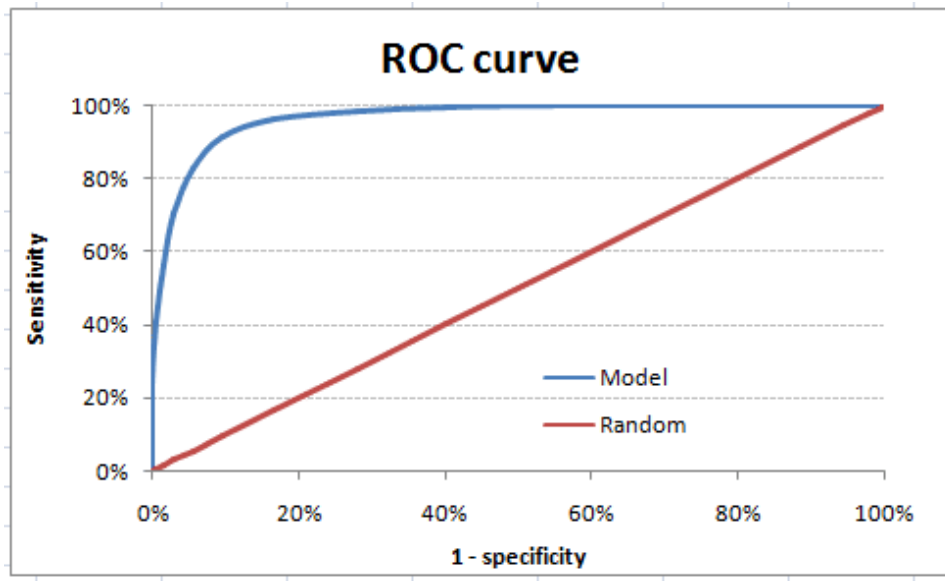
False positive rate (FPR) (=Type I error)	실제값이 negative인 데이터 중 positive로 검출된 비율.	$FP/(TN+FP)$
False negative rate (FNR) (=Type II error)	실제값이 positive인 데이터 중 negative로 검출된 비율.	$FN/(TP+FN)$
Positive predictive value (PPV) /Precision	positive로 검출된 데이터 중 실제 positive일 확률	$TP/(TP+FP)$

4-1. classification – ROC Curve

- X축 : $FPR = FP / (FP + TN)$
- Y축 : $TPR = TP / (TP + FN)$

좋은 모델일 수록 ROC Curve의 elbow point는 (0, 1)에 가까워진다.

- Diagonal line = Random Guessing 일 때를 의미



4-1. classification – ROC Curve

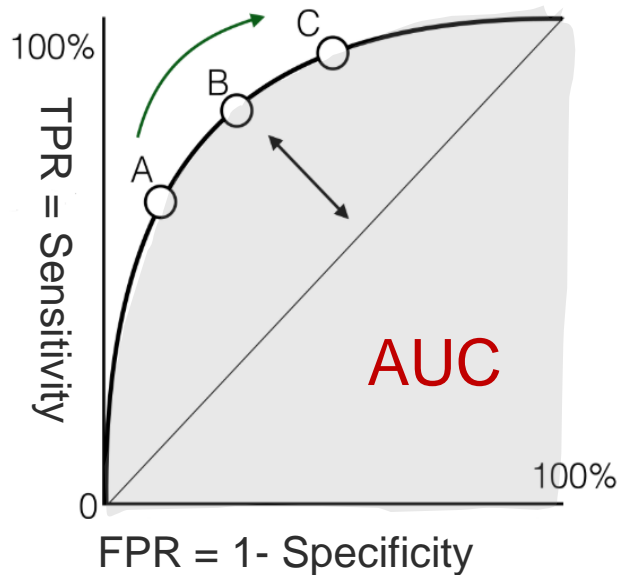
Area Under CURVE = AUC

AUC Range : [0, 1]

Random Guessing AUC = 0.5

100% 정확도 예측 모델 AUC = 1.

⇒ AUC가 1에 가까울수록 좋은 모델이라고 볼 수 있다.



4-1. classification – F1 score

- 분류 모델을 평가할 때 두 번째로 많이 사용하는 precision & recall을 활용,
- Precision과 Recall의 조화평균.
- Precision과 recall이 비슷할수록 F1 score는 높게 계산되며, 데이터 label이 불균형 구조일때 자주 사용된다.

$$F1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} =$$

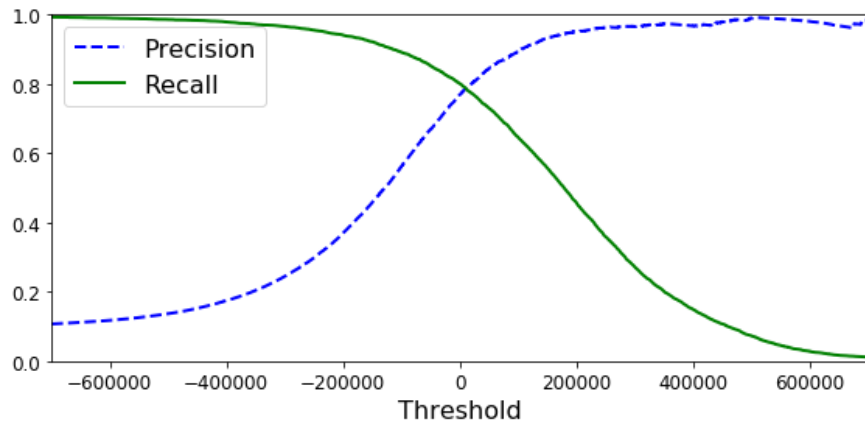
$$\frac{2 * (precision * recall)}{precision + recall}$$

4-1. classification – Precision & Recall trade-off

- precision과 recall사이에는 trade-off가 존재한다.

Ex. 생성한 모델이 모든 label을 positiv로 예측하는 경우 – recal은 100%가 되나, precision은 줄어든다.

- 따라서 상황에 따라 precision, recall 각각의 중요도를 매기고 threshold를 정해주는 것이 필요!



4-2. Regression

1. SSE, MSE

- Regression에서의 evaluation metric으로 SSE를 활용할 수 있다.
- $SSE/df = MSE$

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

2. RMSE

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

3. R squared

$$R^2 = 1 - \frac{SSE}{SST}, \quad SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

Thank you!