



5. Evaluation

KUBIG 학술부



Index

1. Bias-Variance trade off
2. Underfitting & Overfitting
3. Evaluation Metric

How to evaluate a model?

with **Error!**

$$\textit{Error}(\theta) = \underbrace{\textit{Noise}(\theta)}_{\text{irreducible error}} + \underbrace{\textit{Var}(\theta) + \textit{Bias}(\theta)}_{\text{reducible error}}$$

Source : <https://bit.ly/2ILIdBq>

1. Bias-Variance trade off

- irreducible error

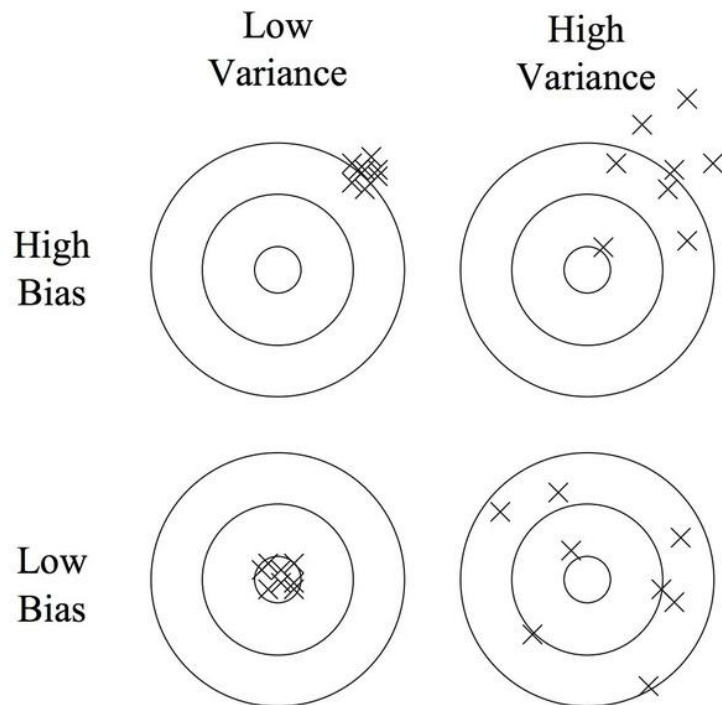
Noise(잡음) : 데이터가 가지는 본질적인 한계치

- reducible error

1. Bias(편향) : 참 값과 추정 값의 차이

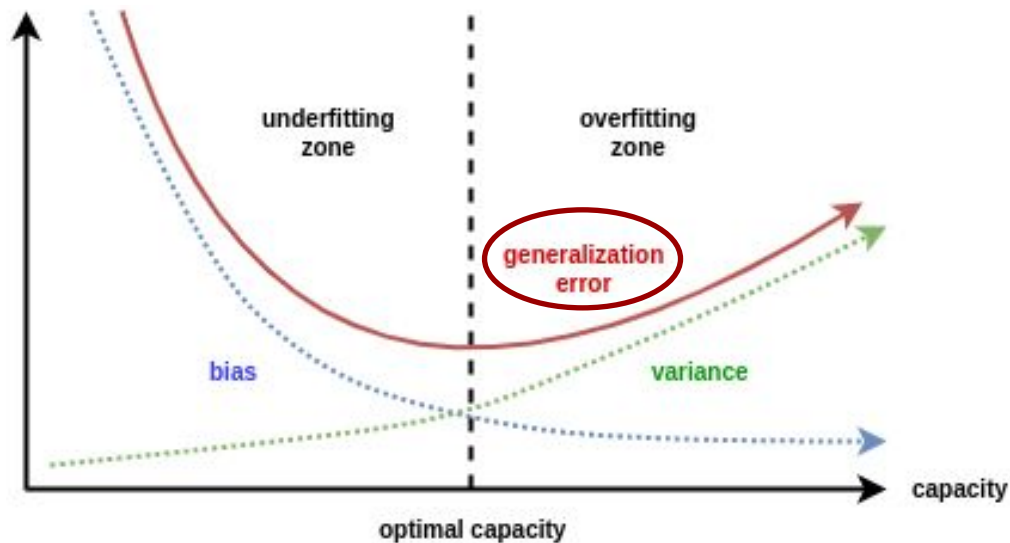
2. Variance(분산) : 추정 값의 산포도

오른쪽의 그림에서 보면 알 수 있듯이,
Low Bias, Low Variance를 가지는
model을 만드는 것이 우리의 목표이다.



Source : <https://bit.ly/2ILIdBq>

1. Bias-Variance trade off



지도학습에서 bias variance trade-off를 잘 조정하는 것은 가장 중요한 부분이다.

Source : <https://djsaunde.wordpress.com/2017/07/17/the-bias-variance-tradeoff/>

$$MSE(\theta) = Var(\theta) + Bias(\theta)^2$$

Variance와 Bias를 잘 조정하여 가장 작은 MSE를 만들어 내는 것이 최종 목표이다.

Source : <https://bywords.tistory.com/entry/번역-유치원생도-이해할-수-있는-biasvariance-tradeoff>

2. Underfitting & Overfitting

- Bias?

데이터의 모든 정보를 고려하지 않아서 지속적으로 잘못된 것을 학습하는 알고리즘의 경향이다.

Train Data가 변경될 때 알고리즘의 평균 정확도가 얼마나 변하는지 보여준다.

- Underfitting

작은 Variance, 큰 Bias를 가진 상태
Train Data를 잘 설명하지 못한다.
새로운 데이터들을 막 예측해버린다.

- Variance?

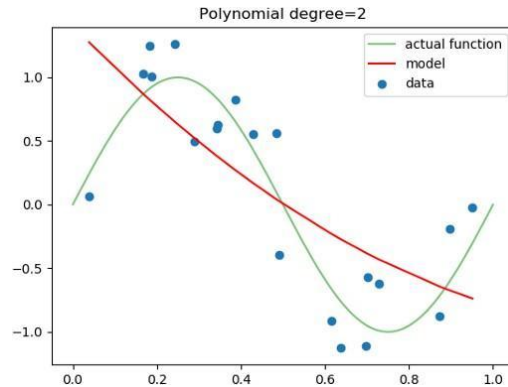
데이터의 Noise를 너무 반영한 유연한 모델을 적합하여 참값과 상관없는 것을 학습하는 알고리즘의 경향이다.

특정 Input Data에 알고리즘이 얼마나 민감한지 나타낸다.

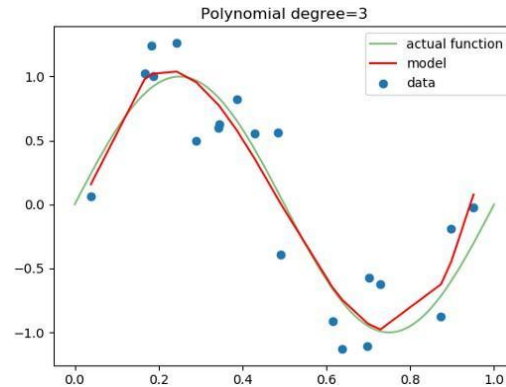
- Overfitting

큰 Variance, 작은 Bias를 가진 상태 Train Data만 매우 잘 예측한다.
새로운 데이터들을 잘 예측하지 못한다.

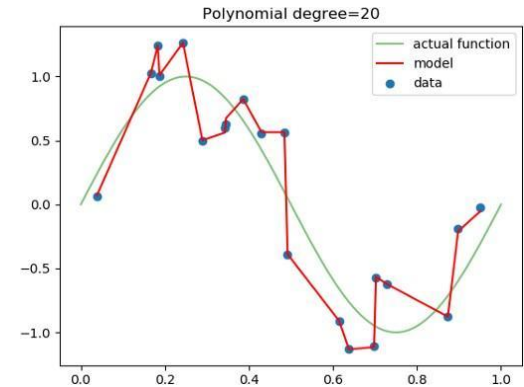
2. Underfitting & Overfitting



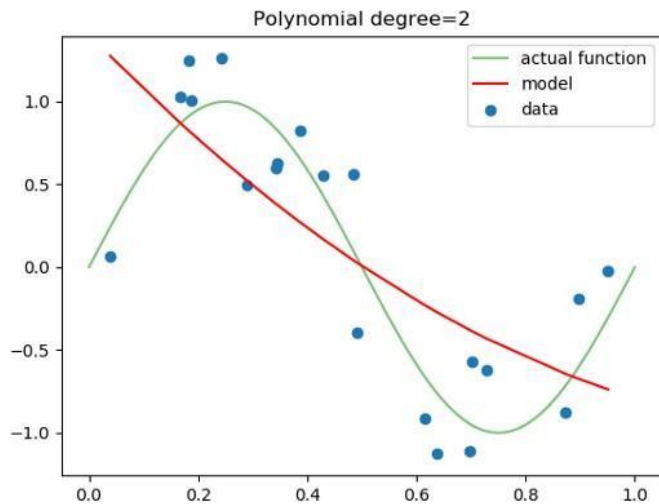
Underfitting



Desirable fit

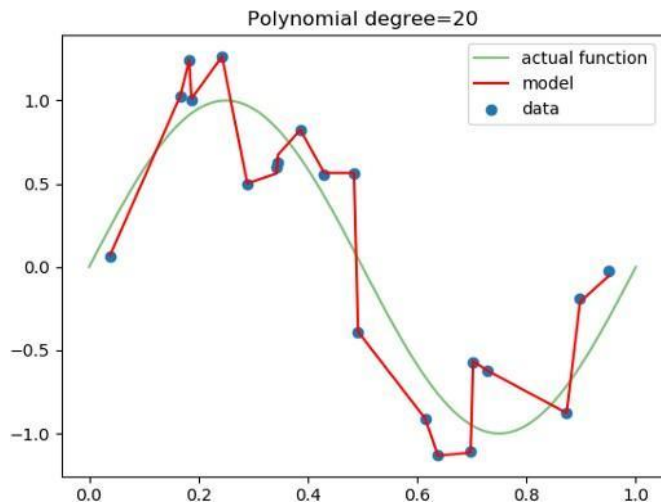


Overfitting



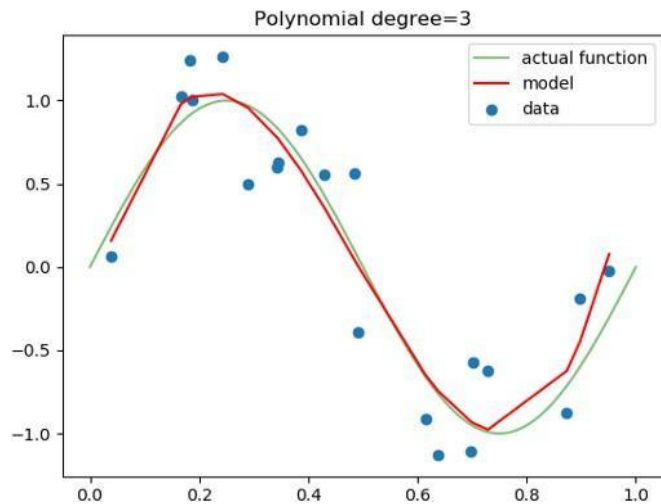
- Underfitting

1. Training Data의 많은 정보를 놓친다. (High Bias)
2. 새로운 데이터가 들어와도 모형의 형태가 크게 변하지 않을 것이다. (Low Variance)
3. Bias가 크기에 예측치가 참값을 예측한다고 믿을 수 없다.



- Overfitting

1. Training Data의 많은 부분을 설명한다. (Low Bias)
2. 새로운 데이터가 들어왔을 때 모형이 완전히 다른 형태로 변하게 된다. (High Variance)
3. Variance가 크기에 예측치가 참값에 가까운 값이라고 믿을 수 없다.



- Desirable fit

1. 데이터의 규칙성을 잘 잡아낸다. (Low Bias)
2. 새로운 데이터가 들어왔을 때 형태가 변하지 않아 일반화가 가능하다. (Low Variance)
3. Bias-Variance trade off의 optimal point에 해당된다.

다다익선은
Machine Learning에서는
통하지 않습니다.

Overfitting

3. Evaluation Metric

3.1 Classification

3.1.1 Confusion Matrix

3.1.2 ROC Curve & AUC

3.1.3 Precision-Recall Curve

3.2 Regression

3.1.1 Confusion Matrix

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

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

$$\text{Accuracy} = \frac{a + d}{a + b + c + d} = \frac{TP + TN}{TP + FN + FP + TN}$$

Source : JunGeol Baek, 2019 1st semester Data mining chapter 3. pp.79.

cf) Accuracy 변형 metric

- $\text{Accuracy} = \frac{n_c}{n}$

n is the number of instances **covered** by the rule / path.

n_c is the number of **positive (correct) instances** covered by rule.

m is the **total number of classes**

p_c is the **prior probability** for the positive class.

- $\text{Laplace} = \frac{n_c + 1}{n + k}$

- $\text{M - estimate} = \frac{n_c + mp_c}{n + m}$

⇒ Naïve Bayes Classifier, Decision Tree 등에서 사용

Source : JunGeol Baek, 2019 1st semester Data mining chapter 4. pp.22.

3.1.1 Confusion Matrix

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

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

$$\text{TPR (Sensitivity)} = \frac{a}{a + b} = \frac{TP}{TP + FN}$$

$$\text{FPR (1 - Specificity)} = \frac{c}{c + d} = \frac{FP}{FP + TN}$$

Source : JunGeol Baek, 2019 1st semester Data mining chapter 3. pp.79.

3.1.1 Confusion Matrix

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

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

$$\text{Recall (r)} = \frac{a}{a + b} = \frac{TP}{TP + FN}$$

$$\text{Precision (p)} = \frac{a}{a + c} = \frac{TP}{TP + FP}$$

$$\text{F - measure (F)} = \frac{2rp}{r + p}$$

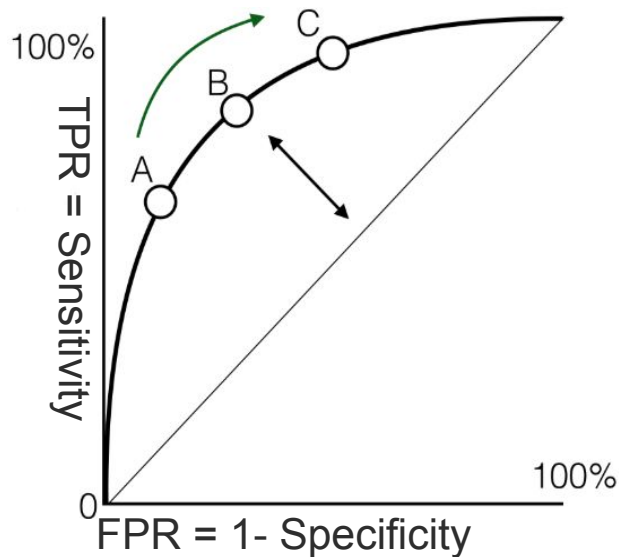
Source : JunGeol Baek, 2019 1st semester Data mining chapter 3. pp.84.

3.1.2 ROC Curve & AUC

- 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



3.1.2 ROC Curve & AUC

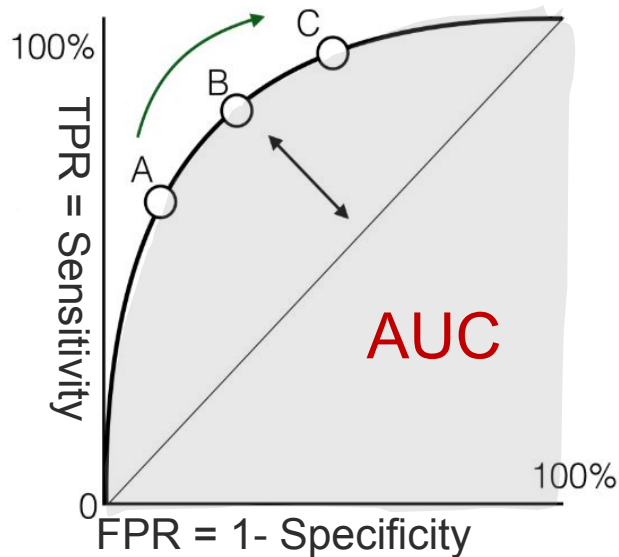
Area Under CURVE = AUC

AUC Range : [0, 1]

Random Guessing AUC = 0.5

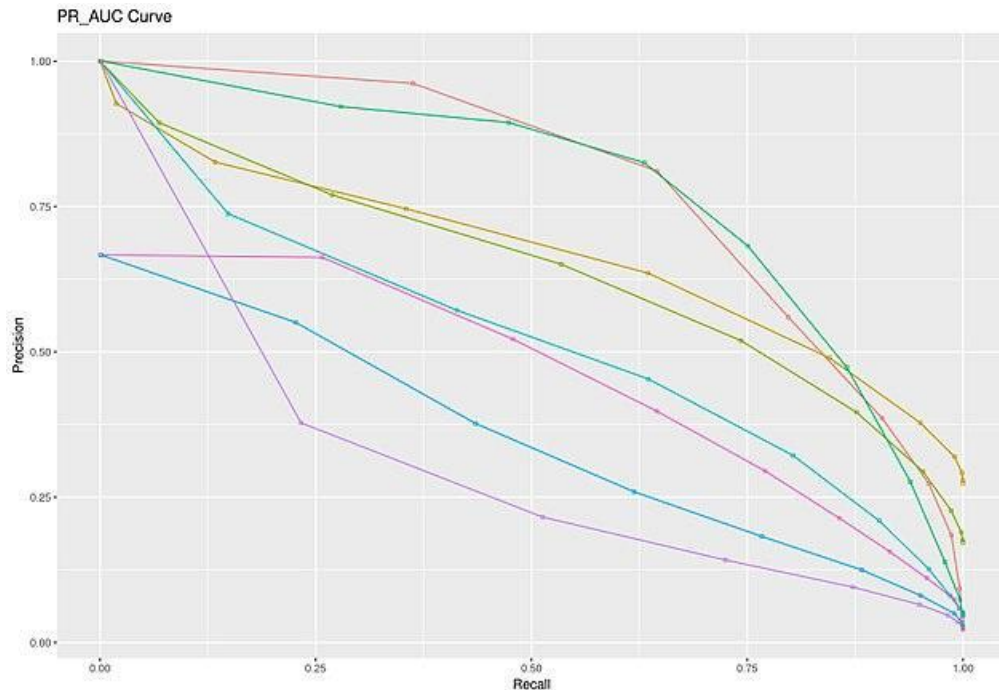
100% 정확도 예측 모델 AUC = 1.

⇒ AUC가 1에 가까울수록 좋은
모델이라고 볼 수 있다.



3.1.3 Precision - Recall Curve

- X축 : Recall = $TP / (TP + FN)$
- Y축 : Precision = $TP / (TP + FP)$
- In Case of imbalanced-Data
⇒ Precision이 FPR에 비해 False Positive를 더 민감하게 잡아낼 수 있다.
⇒ Imbalanced data에서 효과적인 metric!



3.2 Evaluation Metric - Regression

$$SST = SSR + SSE$$

- Sum of Squared Error (SSE) = $\sum_{i=1}^N (y_i - \bar{y})^2$
- Regression Sum of Squares (SSR) = $\sum_{i=1}^N (\hat{y}_i - \bar{y})^2$
- Total Sum of Squares (SST) = $\sum_{i=1}^N (y_i - \hat{y}_i)^2$

Proof : <https://stats.stackexchange.com/questions/207841/why-is-sst-sse-ssr-one-variable-linear-regression>

3.2 Evaluation Metric - Regression

- Sum of Squared Error : $SSE = \sum_{i=1}^N (y_i - \hat{y}_i)^2$
- 데이터의 개수가 많아질수록 커지는 경향이 있다.

- Mean Squared Error : $MSE = SSE/df$

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

- Root Mean Squared Error : $RMSE = \sqrt{MSE}$

3.2 Evaluation Metric - Regression

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

모델이 복잡할수록 값이 커지는
경향이 있다.

- $Adjusted\ R^2 = 1 - \frac{SSE/(N-K)}{SST/(N-1)}$

N : sample size

K : the number of variable

기존의 R^2 에 모델의 복잡도를
페널티로 주었다.

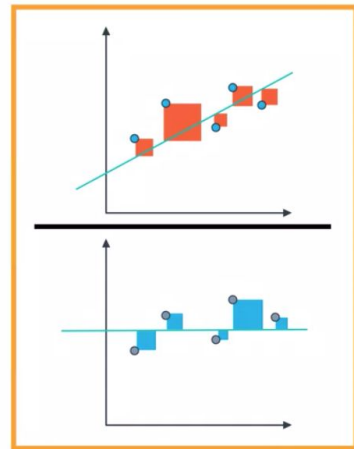
BAD MODEL

The errors should be similar.
R2 score should be close to 0.

GOOD MODEL

The mean squared error for the
linear regression model should be
a lot smaller than the mean squared
error for the simple model.

$$R^2 = 1 -$$



Source : <https://medium.com/@Aj.Cheng/mse-r2-score-f64bb2f84d54>

Reference

1. Bias-Variance trade off

<https://www.quora.com/What-is-the-best-way-to-explain-the-bias-variance-trade-off-in-layman%E2%80%99s-terms>

<https://djsaunde.wordpress.com/2017/07/17/the-bias-variance-tradeoff/>

https://en.wikipedia.org/wiki/Bias-variance_tradeoff

- Noise vs. Bias vs. Variance

<https://medium.com/autonomous-agents/mathematical-foundation-for-noise-bias-and-variance-in-neuralnetworks-4f79ee801850>

https://ko.wikipedia.org/wiki/일반화_오차

2. Underfitting vs. overfitting

<https://nittaku.tistory.com/289>

3.1.1 Precision and Recall

https://en.wikipedia.org/wiki/Precision_and_recall#Precision

- m-estimate, Laplace

<https://pdfs.semanticscholar.org/bc02/6fa01bb32e3e2c4fd1bf8c005319a0daf4ea.pdf>

3.1.3 Precision-Recall Curve

https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html

3.2 Evaluation of Regression

<https://stats.stackexchange.com/questions/207841/why-is-sst-sse-ssr-one-variable-linear-regression>

<https://partrita.github.io/posts/regression-error/>

Thank you!

- Q&A Time!