Ensemble Learning: Bias-Variance Decomposition

Pilsung Kang

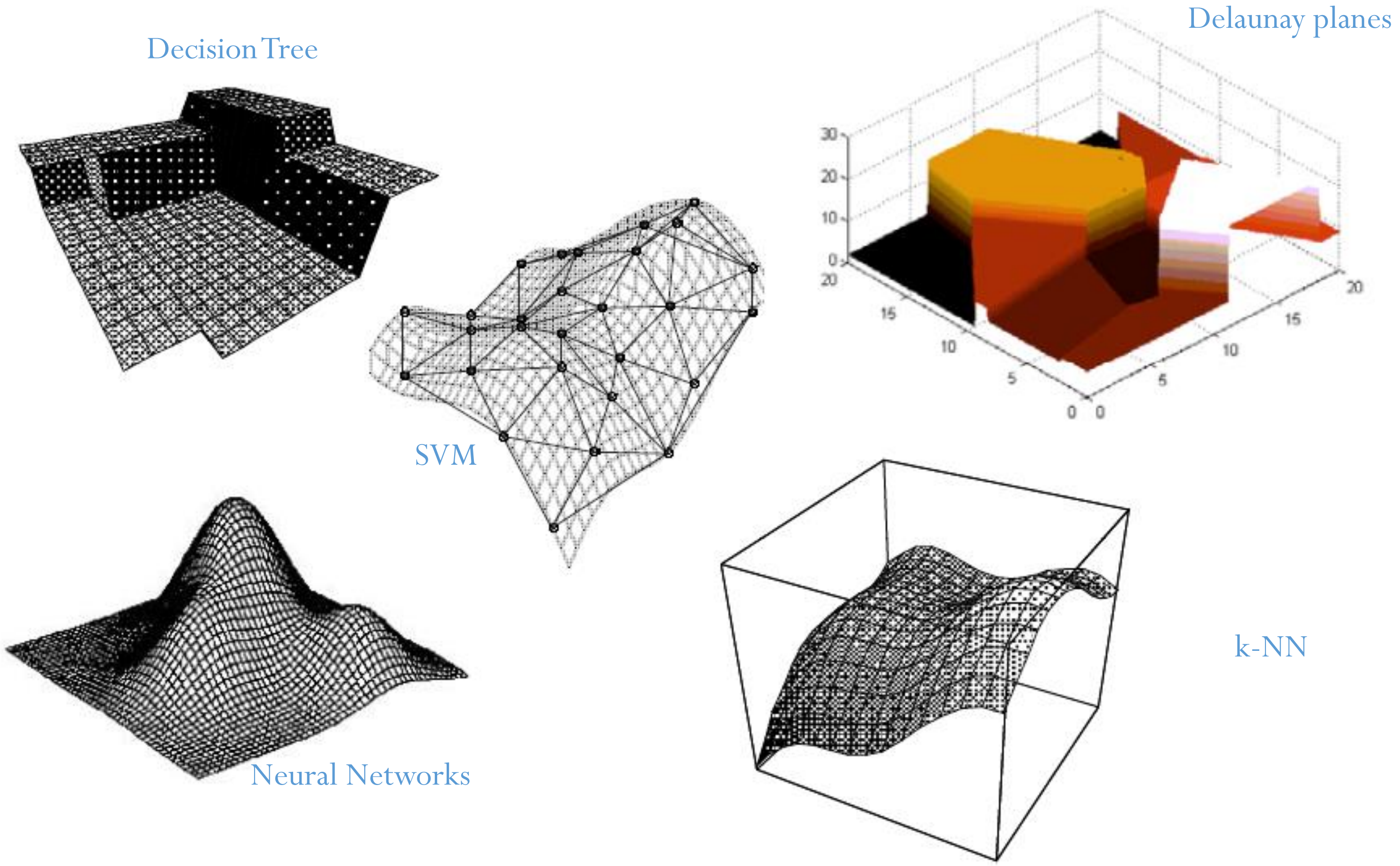
School of Industrial Management Engineering

Korea University

Theoretical Backgrounds: Model Space

집단지성

- Different model produce different class boundaries or fitted functions



결론 도출까지의
Logical thinking 방식이
다르다

Theoretical Backgrounds: Bias-Variance Decomposition

- Suppose the data comes from the "additive error" model

→ 현실 함수 → 노이즈, 에러

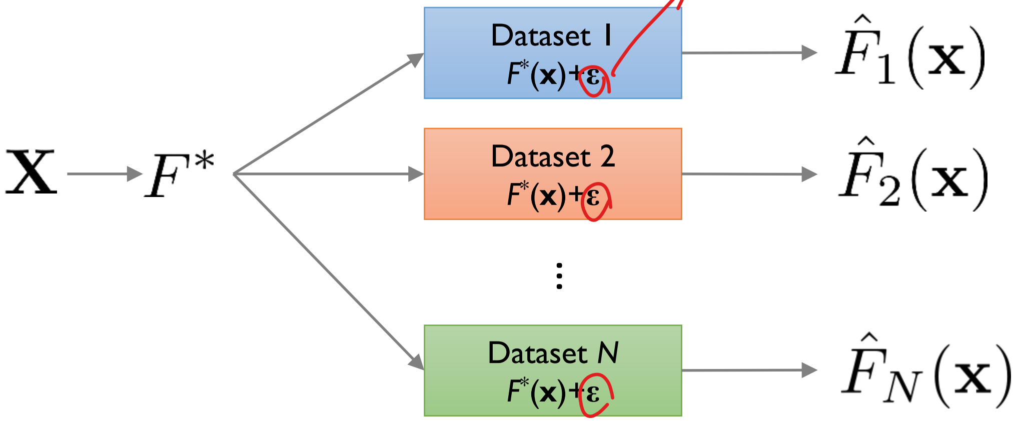
$$y = F^*(\mathbf{x}) + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

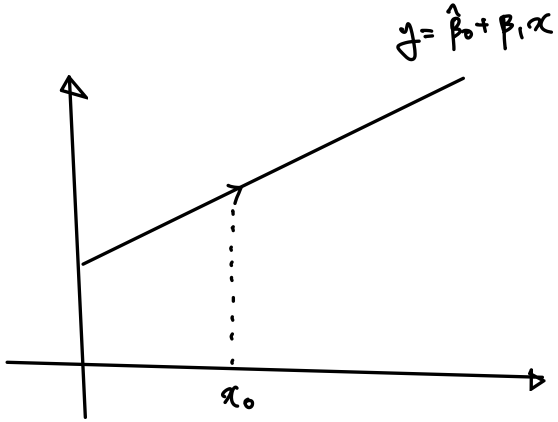
  ✓ $F^*(\mathbf{x})$ is the target function that we are trying to learn, but do not really know

  ✓ The errors are independent and identically distributed

- Consider the estimation process (Linear Regression)

노이즈에 따라 데이터셋이 달라진다. → 역으로 ε을 반영X하고 진행





$E[(\hat{\beta}_0 + \hat{\beta}_1 x_1)]$

$= \beta_0 + \beta_1 x_0$

  ✓ The average fit over all possible datasets:

$$\bar{F}(\mathbf{x}) = E[\hat{F}_D(\mathbf{x})]$$

Theoretical Backgrounds: Bias-Variance Decomposition

- The MSE for a particular data point

$$Err(\mathbf{x}_0) = E\left[\underbrace{y}_{\text{실제값}} - \underbrace{\hat{F}(\mathbf{x})}_{\text{예측값}} \middle| \mathbf{x} = \mathbf{x}_0\right]^2$$

$(y = F^*(\mathbf{x}) + \epsilon)$

$\bar{F}(x) = E[\hat{F}_D(x)]$

$$= E\left[\overset{A}{F^*(\mathbf{x}_0)} + \overset{B}{\epsilon} - \hat{F}(\mathbf{x}_0)\right]^2$$

$\rightarrow E(A+B)^2 = E(A^2+2AB+B^2)$

$= E(A^2) + \underbrace{E(B^2)}_{\sigma^2} + \underbrace{2E(AB)}_{\text{생략 } 0}$ $\rightarrow \epsilon \sim N(0, \sigma^2)$

$$= \underbrace{E\left[F^*(\mathbf{x}_0) - \hat{F}(\mathbf{x}_0)\right]^2}_{E(A^2)} + \underbrace{\sigma^2}_{E(B^2)}$$

$$= E\left[F^*(\mathbf{x}_0) - \bar{F}(\mathbf{x}_0) + \bar{F}(\mathbf{x}_0) - \hat{F}(\mathbf{x}_0)\right]^2 + \sigma^2$$

수식 변형을 위해 추가

Theoretical Backgrounds: Bias-Variance Decomposition

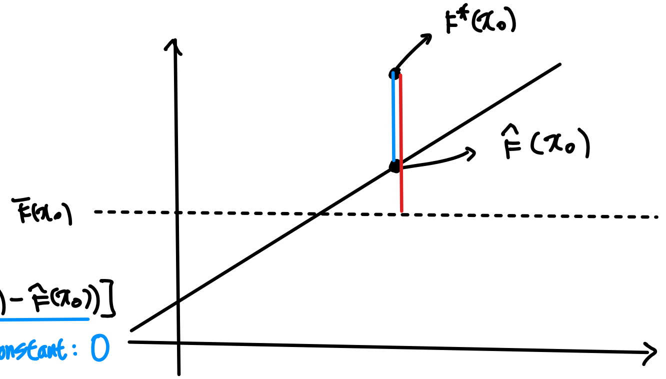
- The MSE for a particular data point

$$= E\Big[F^*(\mathbf{x}_0) - \bar{F}(\mathbf{x}_0) + \bar{F}(\mathbf{x}_0) - \hat{F}(\mathbf{x}_0)\Big]^2 + \sigma^2$$

✓ By the properties of the expectation operator

$2E[(F^*(x_0) - \bar{F}(x_0))(\bar{F}(x_0) - \hat{F}(x_0))]$

↳ constant : 0



$$= E\Big[F^*(\mathbf{x}_0) - \bar{F}(\mathbf{x}_0)\Big]^2 + E\Big[\bar{F}(\mathbf{x}_0) - \hat{F}(\mathbf{x}_0)\Big]^2 + \sigma^2$$

↳ 정답값 ↳ 평균값

Constant

$$= \Big[F^*(\mathbf{x}_0) - \bar{F}(\mathbf{x}_0)\Big]^2 + E\Big[\bar{F}(\mathbf{x}_0) - \hat{F}(\mathbf{x}_0)\Big]^2 + \sigma^2$$

정답값 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_N$ 평균값 예측값

$$= Bias^2(\hat{F}(\mathbf{x}_0)) + Var(\hat{F}(\mathbf{x}_0)) + \sigma^2$$

Bias : 모델의 노이즈를 바꿔갈 때 평균값이 정답값과 얼마나 가까운가

Var : 모델의 노이즈를 바꿔갈 때 평균값이 정답값과 차이가 나는가

고려대학교 KOREA UNIVERSITY

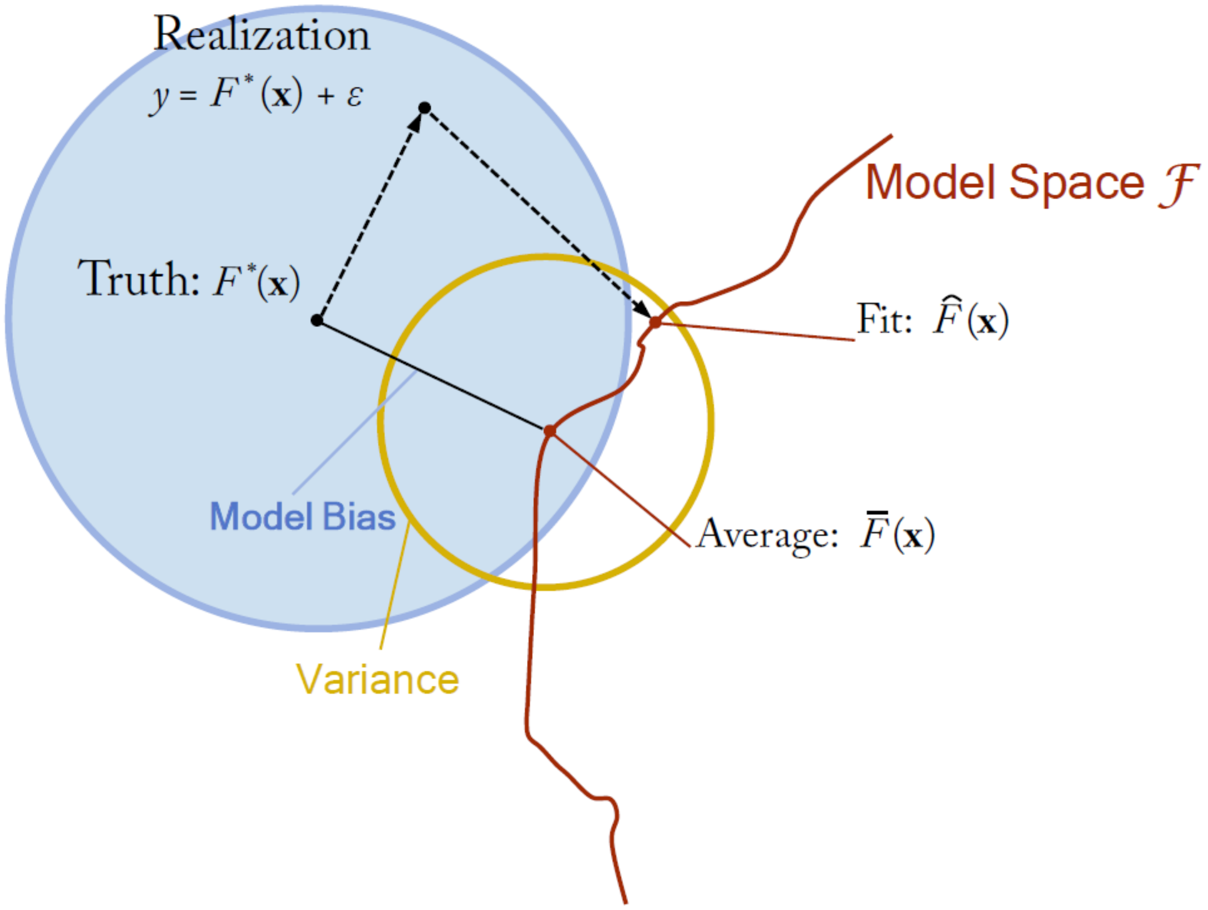
DSBA Data Science & Business Analytics

Theoretical Backgrounds: Bias-Variance Decomposition

- Properties of Bias and Variance
  - ✓ Bias^2: the amount by which the average estimator differs from the truth
    - Low bias: on average, we will accurately estimate the function from the dataset
    - High bias implies a **poor** match
  - ✓ Variance: spread of the individual estimations around their mean
    - Low variance: estimated function does not change much with different datasets
    - High variance implies a **weak** match
  - ✓ Irreducible error: the error that was present in the original data
  - ✓ Bias and variance are not independent of each other

Theoretical Backgrounds: Bias-Variance Decomposition

- Graphical representation of Bias-Variance decomposition



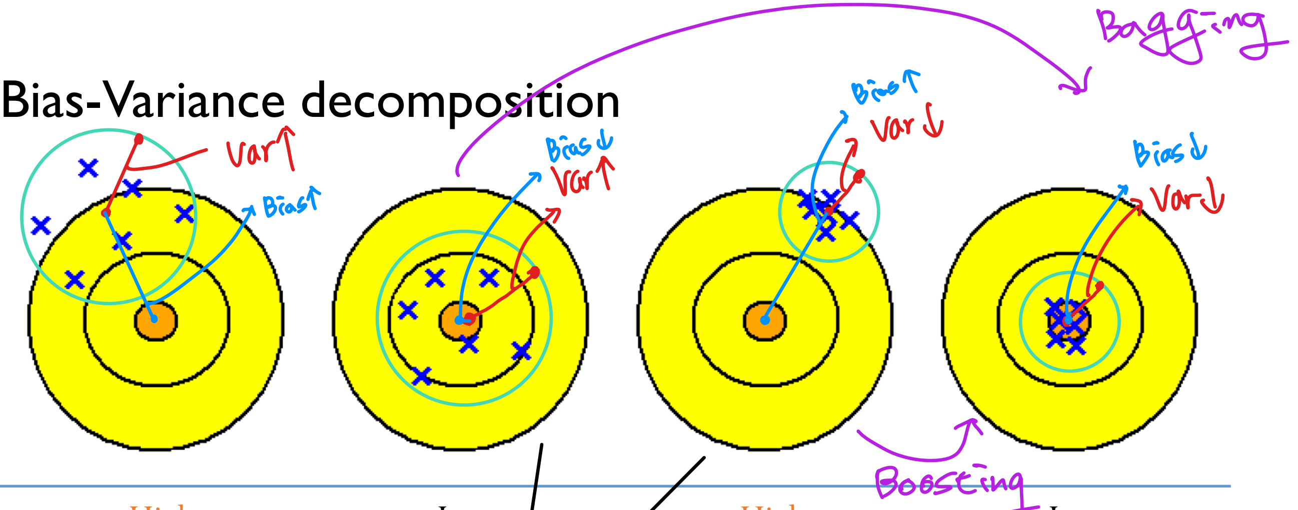
Theoretical Backgrounds: Bias-Variance Decomposition

- Graphical representation of Bias-Variance decomposition

● : 정답 (F*(x)) 분포

X : 예측값 (F̂(x))

● : 예측값 분포



Bias	High	Low	High	Low
Variance	High	High	Low	Low

↳ Best Case!

✓ Lower model complexity: high bias & low variance

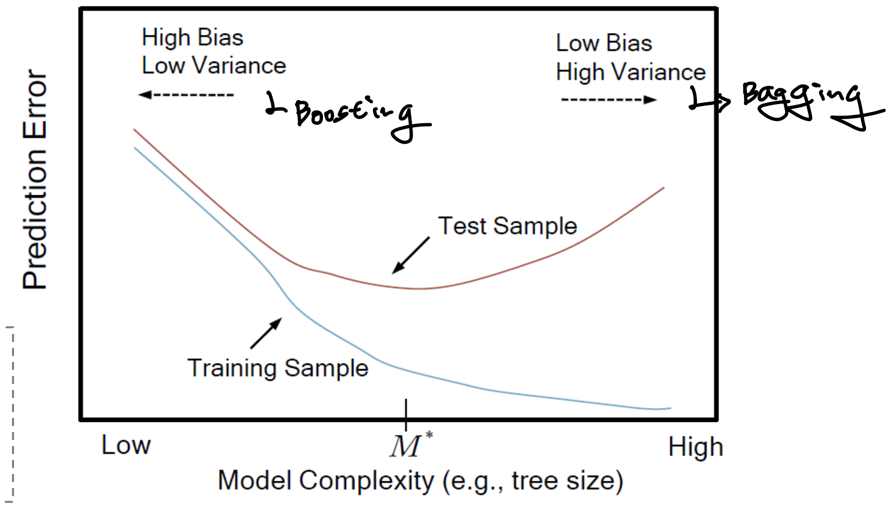
- Logistic regression, LDA, k-NN with large k, etc.

✓ Higher model complexity: low bias & high variance

- DT, ANN, SVM, k-NN with small k, etc.

Bias-Variance Dilemma

The more complex (flexible) we make the model,
the lower the bias but the higher the variance it is subjected to.



Theoretical Backgrounds: Bias-Variance Decomposition

- Bias-Variance example

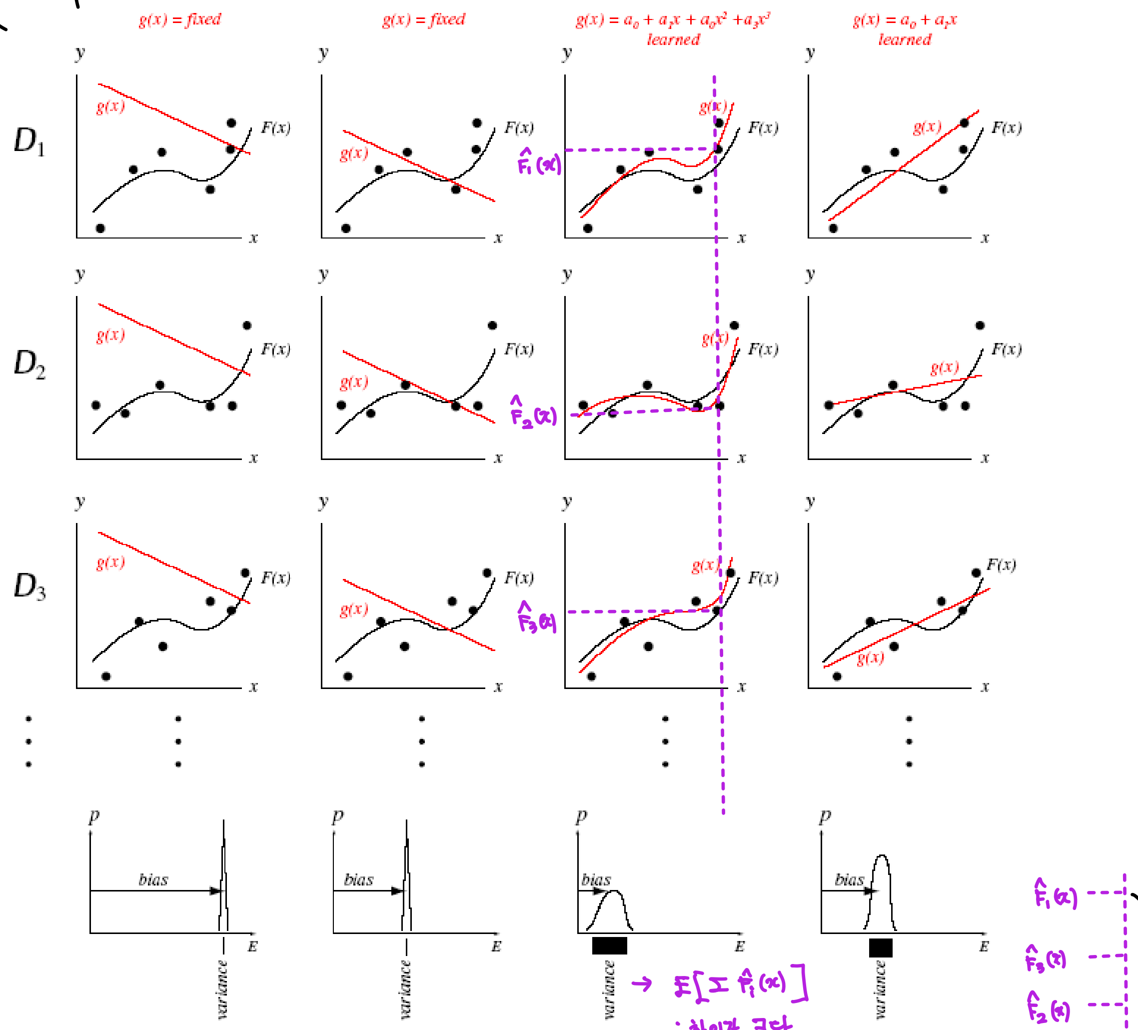
D_i : Same Dataset　　～ : $F^*(x)$　　• : $F^*(x)+\varepsilon$

데이터셋 \ 모델종류

Each column is a different model.

Each row is a different dataset of 6 points.

Histograms of mean-squared error of the fit.



Var

Col 1:
Poor fixed linear model;
High bias, zero variance

Col 2:
Slightly better fixed linear model;
Lower (but high) bias,
zero variance.

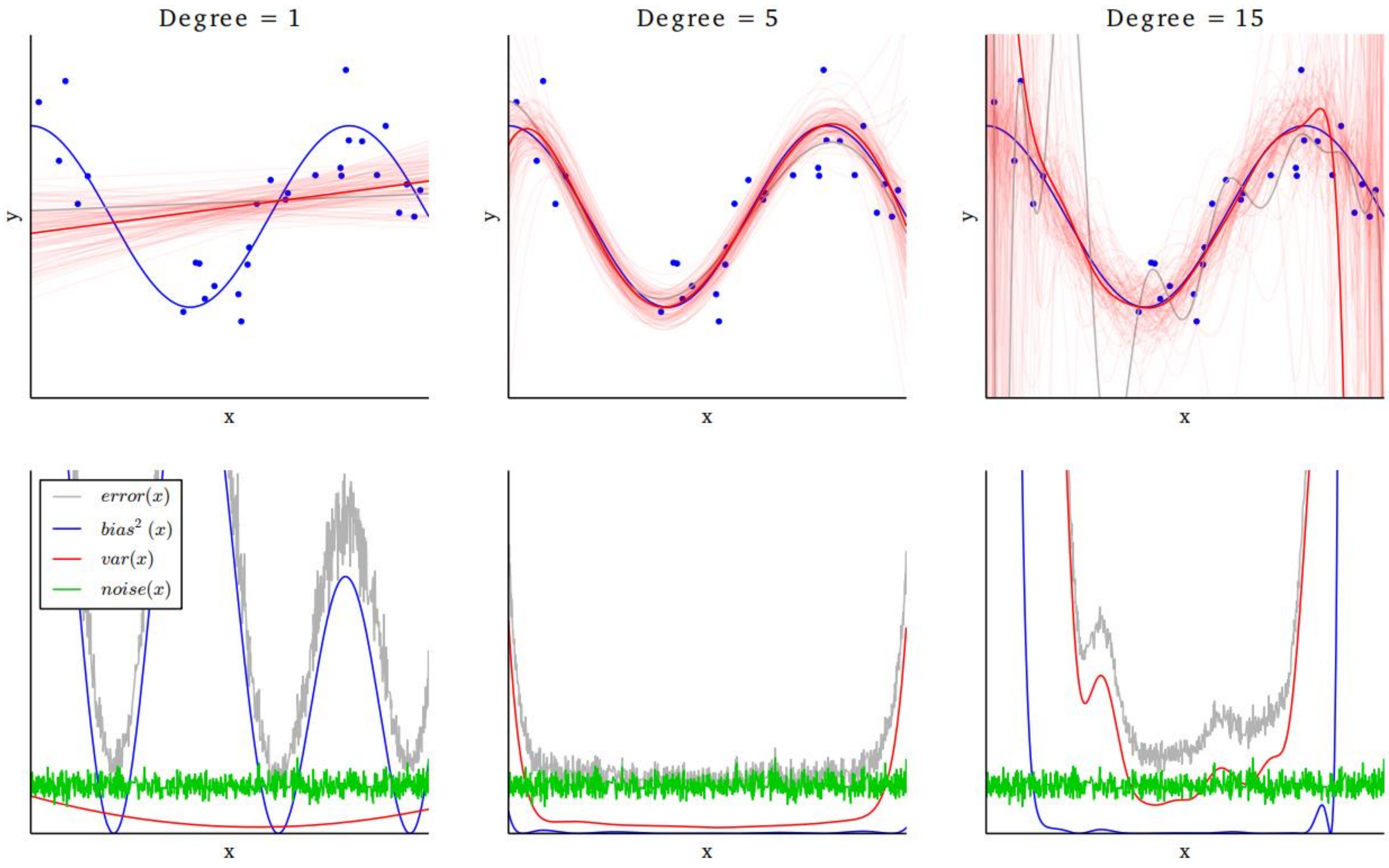
Col 3:
Learned cubic model;
Low bias, moderate variance.

Col 4:
Learned linear model;
Intermediate bias and variance.

Theoretical Backgrounds: Bias-Variance Decomposition

bias↑ / Boosting   var↑ / Bagging

- Bias-Variance example



Purpose of Ensemble

- Goal: Reduce the error through constructing multiple learners to
  - ✓ Reduce the variance: Bagging, Random Forests
  - ✓ Reduce the bias: AdaBoost
  - ✓ Both: Mixture of experts

- Two key questions on the ensemble construction
  - ✓ Q1: How to generate individual components of the ensemble systems (base classifiers) to achieve sufficient degree of diversity?

    → Identical한 모델을 여러개 만드는 것은 의미가 없다

  - ✓ Q2: How to combine the outputs of individual classifiers?

Ensemble Diversity

- Ensemble will have no gain from combining a set of identical models
  - ✓ Need base learners whose fitted functions are adequately different from those of others
  - ✓ Wish models to exhibit a certain element of diversity in their group behavior, though still retaining good performance individually.

Diversity	Implicit	Explicit
Description	Provide different random subset of the training data to each learner	Use some measurement ensuring it is substantially different from the other members
Ensemble Algorithms	Instance: Bagging Variables: Random Subspaces, Rotation Forests Both: Random Forests	Boosting, Negative Correlation Learning ↳ Diversity↑ = Model간의 corr ↓

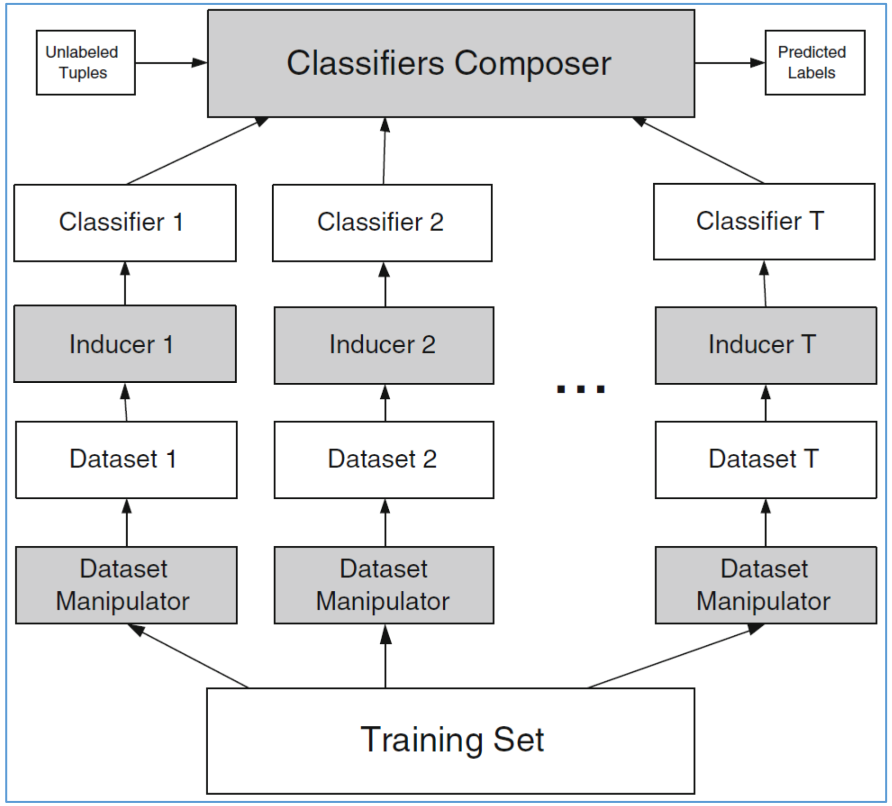
Ensemble Diversity

```python
import xgboost as xgb

# XGBoost의 native API를 사용하는 경우
params = {
    'objective': 'binary:logistic',
    'nthread': 4  # 병렬 처리를 위한 코어 수 설정
}
train_data = xgb.DMatrix(X_train, label=y_train)
bst = xgb.train(params, train_data, num_boost_round=100)

# scikit-learn API를 사용하는 경우
clf = xgb.XGBClassifier(n_jobs=4)  # 병렬 처리를 위한 코어 수 설정
clf.fit(X_train, y_train)
```

- Independent (implicit) vs. Model guided (explicit) instance selection

→ Individual Model에 대한 계산복잡도 ↑ → 시간이 생각보다 오래 걸리는 편

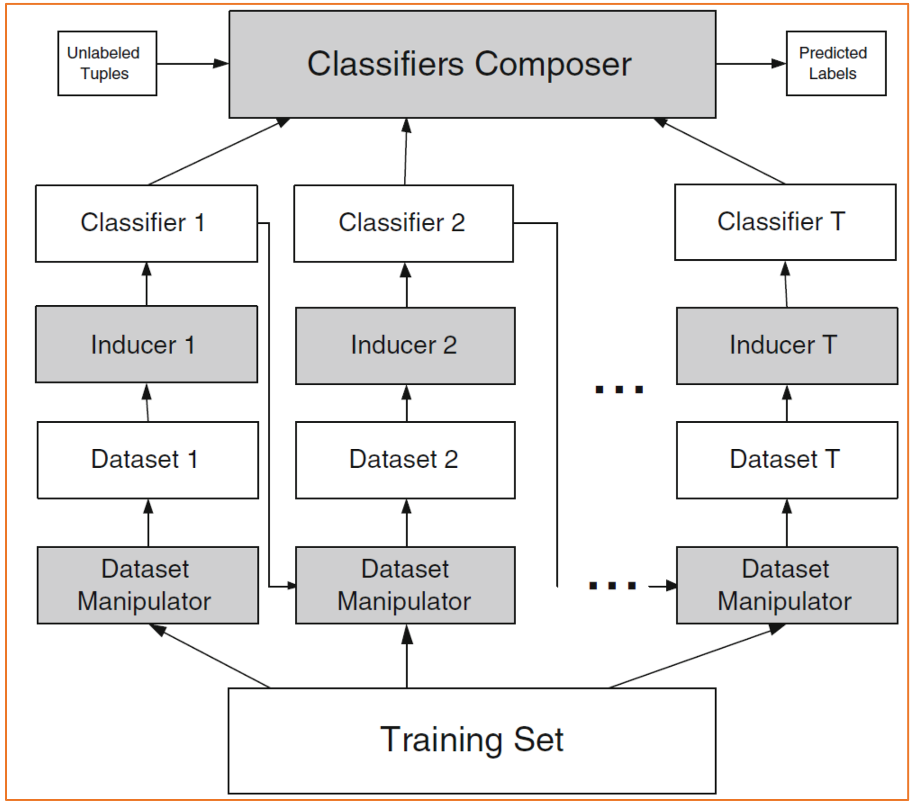
Independent instance selection



Bagging

병렬처리 O (항상 수행시간이 짧은 것 X)

Model guided instance selection



Boosting

병렬처리 X → 하드웨어적으로 처리 가능하긴 함. (e.g. XGBoost)

① Parallel Tree Boosting
② Feature Parallelism
③ Data Parallelism

고려대학교 KOREA UNIVERSITY

DSBA Data Science & Business Analytics

Why Ensemble?

- Why Ensemble works?
  - ✓ True functions, estimations, and the expected error

$$y_m(\mathbf{x}) = f(\mathbf{x}) + \epsilon_m(\mathbf{x}). \quad \mathbb{E}_{\mathbf{x}}\left[\{y_m(\mathbf{x}) - f(\mathbf{x})\}^2\right] = \mathbb{E}_{\mathbf{x}}\left[\epsilon_m(\mathbf{x})^2\right]$$

  - ✓ The average error made by M individual models vs. Expected error of the ensemble

$$E_{Avg} = \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\mathbf{x}}\left[\epsilon_m(\mathbf{x})^2\right]$$

앙상블의 출력을
개별 모형 출력들의 평균으로 정의

$$E_{Ensemble} = \mathbb{E}_{\mathbf{x}}\left[\left\{\frac{1}{M}\sum_{m=1}^{M} y_m(\mathbf{x}) - f(\mathbf{x})\right\}^2\right]$$

$$\frac{1}{M} \cdot M \cdot f(x)$$

$$= \mathbb{E}_{\mathbf{x}}\left[\left\{\frac{1}{M}\sum_{m=1}^{M}\epsilon_m(\mathbf{x})\right\}^2\right]$$

Why Ensemble?

- Why Ensemble works?
  - ✓ Assume that the errors have zero mean and are uncorrelated,

이론적으로
개별 모형의 ε가 독립이라는
가정하에

$$\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})] = 0, \qquad \mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})\epsilon_l(\mathbf{x})] = 0 \ (m \neq l)$$

  - ✓ The average error made by M individual models vs. Expected error of the ensemble

$$E_{Ensemble} = \frac{1}{M} E_{Avg}$$

  - ✓ In reality (errors are correlated), by the Cauchy's inequality

$$\left[\sum_{m=1}^{M} \epsilon_m(\mathbf{x})\right]^2 \leq M \sum_{m=1}^{M} \epsilon_m(\mathbf{x})^2 \Rightarrow \left[\frac{1}{M}\sum_{m=1}^{M} \epsilon_m(\mathbf{x})\right]^2 \leq \frac{1}{M}\sum_{m=1}^{M} \epsilon_m(\mathbf{x})^2$$

$E_{Ensemble}$ E_{Avg}

$(ax+by)^2 \leq (a^2+b^2)(x^2+y^2)$

$$\boxed{E_{Ensemble} \leq E_{Avg}}$$


Q
&
A