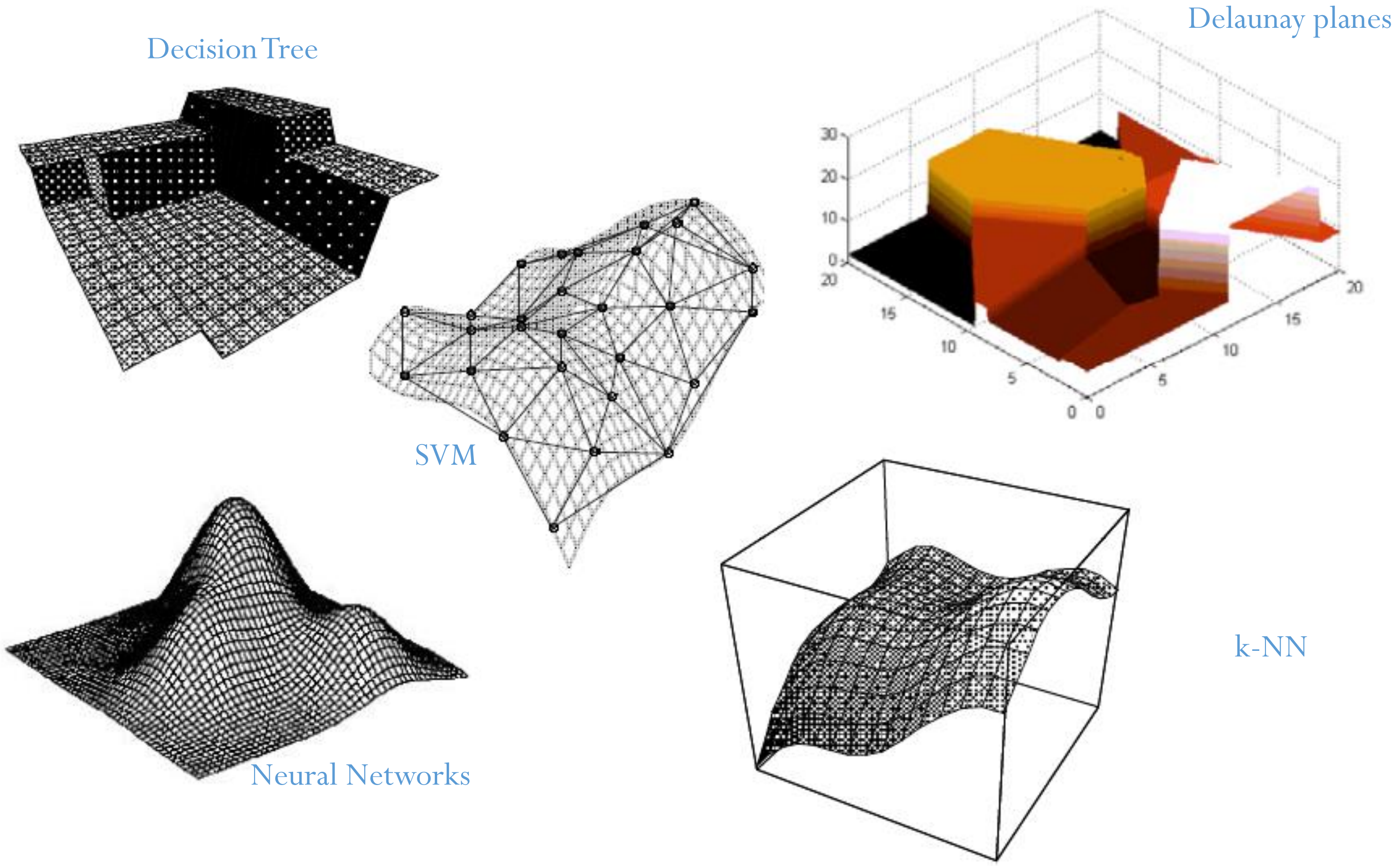
Ensemble Learning:
Bias-Variance Decomposition

Pilsung Kang
School of Industrial Management Engineering
Korea University

Theoretical Backgrounds: Model Space

- Different model produce different class boundaries or fitted functions



Theoretical Backgrounds: Bias-Variance Decomposition

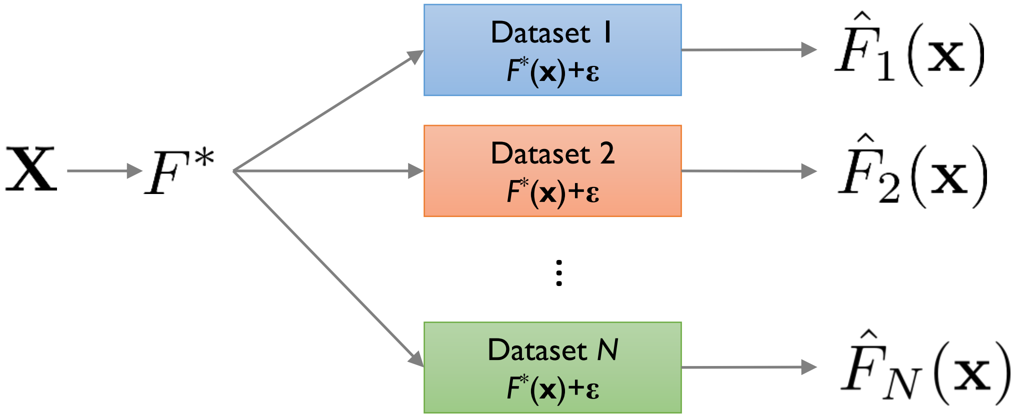
- Suppose the data comes from the "additive error" model

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

✓ $F^*(\mathbf{x})$ 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



✓ 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

$$Err(\mathbf{x}_0) = E\Big[y - \hat{F}(\mathbf{x})|\mathbf{x} = \mathbf{x}_0\Big]^2 \qquad (y = F^*(\mathbf{x}) + \epsilon)$$

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

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

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

Theoretical Backgrounds: Bias-Variance Decomposition

- The MSE for a particular data point

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

✓ By the properties of the expectation operator

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

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

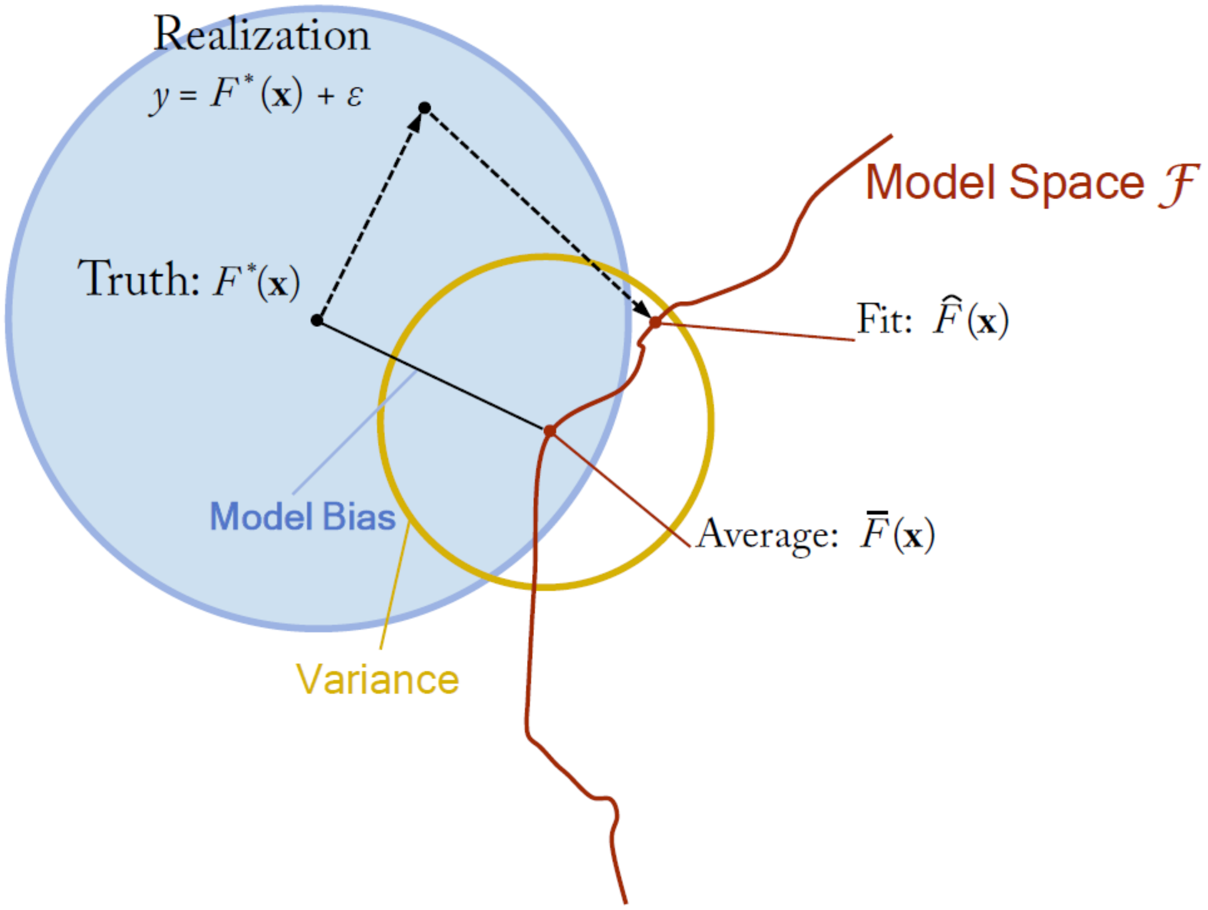
$$= Bias^2\big(\hat{F}(\mathbf{x}_0)\big) + Var\big(\hat{F}(\mathbf{x}_0)\big) + \sigma^2$$

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

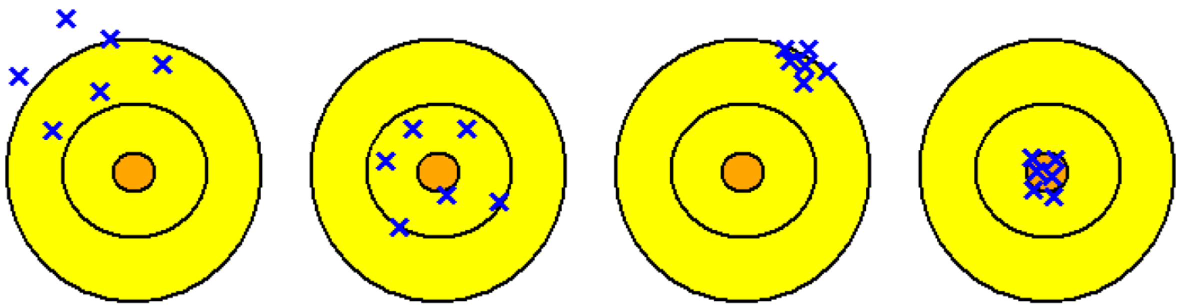
Theoretical Backgrounds: Bias-Variance Decomposition

- Graphical representation of Bias-Variance decomposition



Theoretical Backgrounds: Bias-Variance Decomposition

- Graphical representation of Bias-Variance decomposition



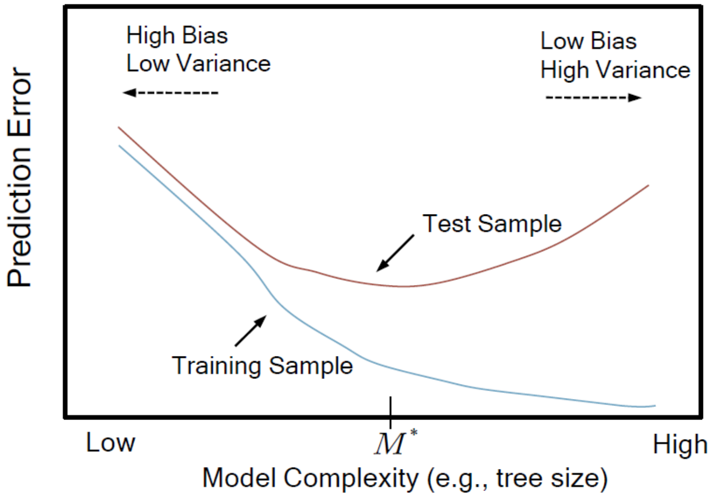
Bias	High	Low	High	Low
Variance	High	High	Low	Low

✓ 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.

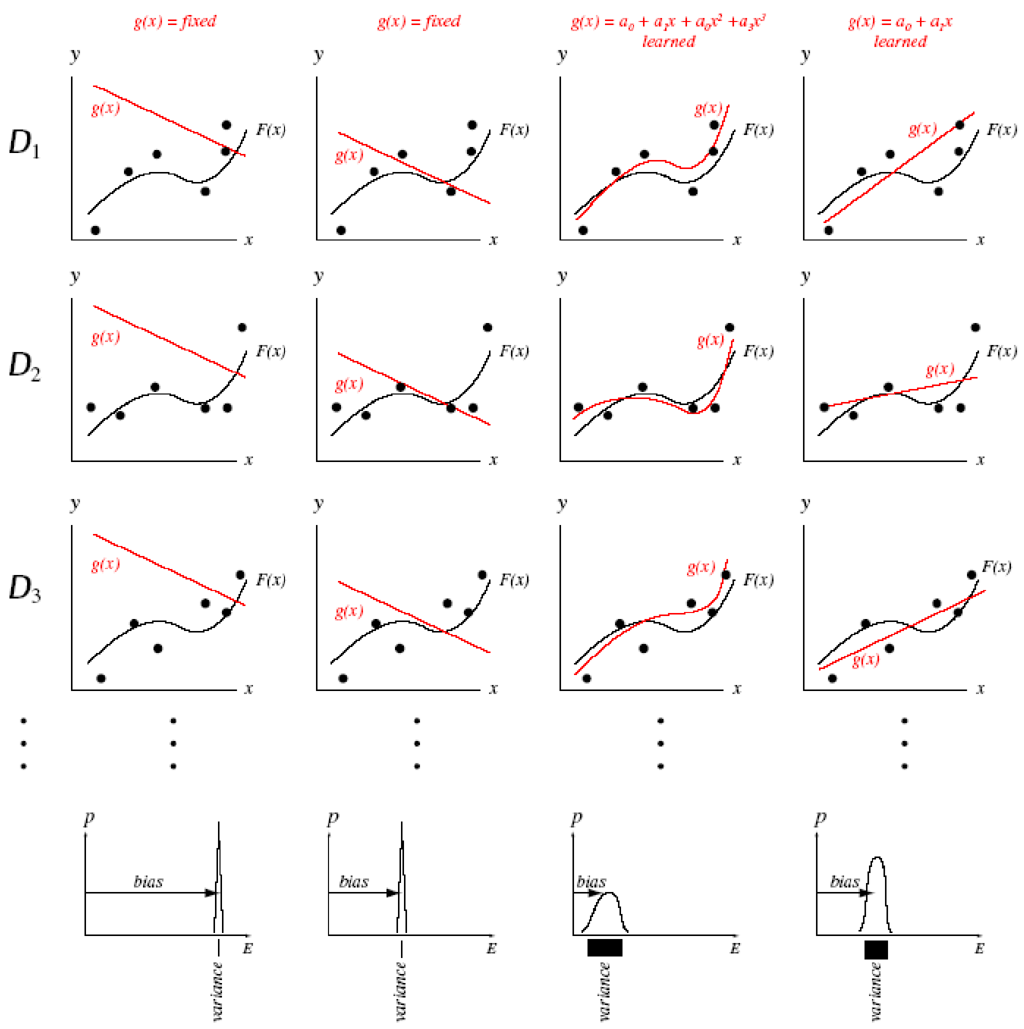
Theoretical Backgrounds: Bias-Variance Decomposition

- Bias-Variance example

Each column is a different model.

Each row is a different dataset of 6 points.

Histograms of mean-squared error of the fit.



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

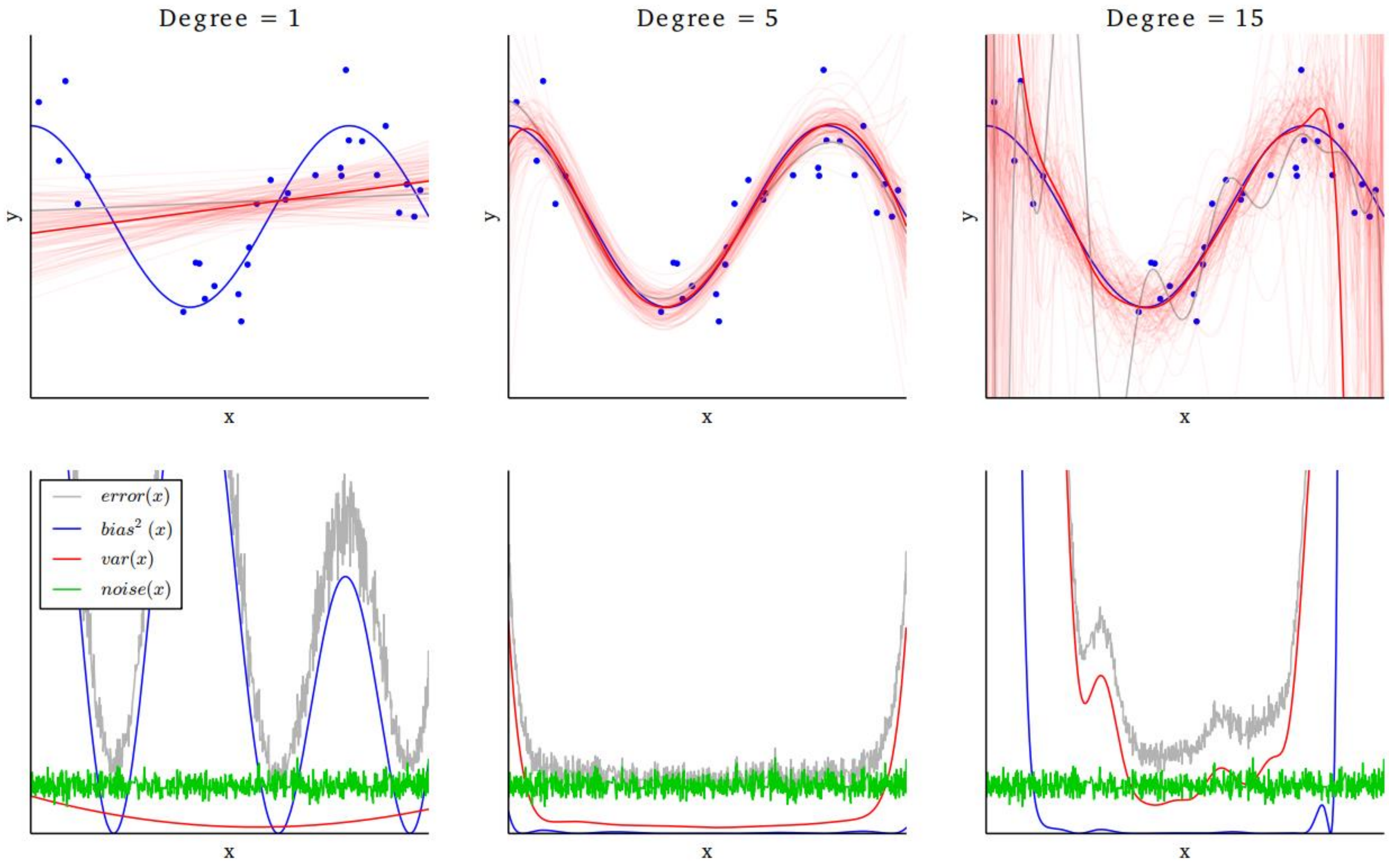
Col 2:
Slightly better fixed linear model;
Lower (but high) bias, zero variance.

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

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

Theoretical Backgrounds: Bias-Variance Decomposition

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

Ensemble Diversity

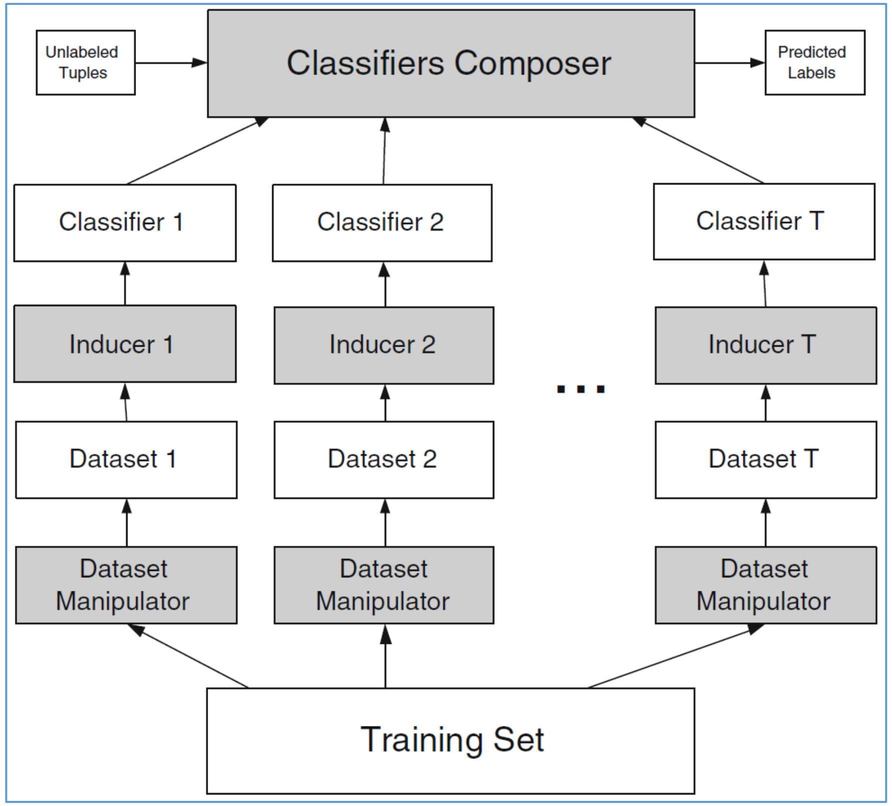
- Ensemble will have no gain from combining a set of identical models
  - ✓ Need base learners whose fitted functions are adequately different from those of others
  - ✓ Wish models to exhibit a certain element of diversity in their group behavior, though still retaining good performance individually.

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

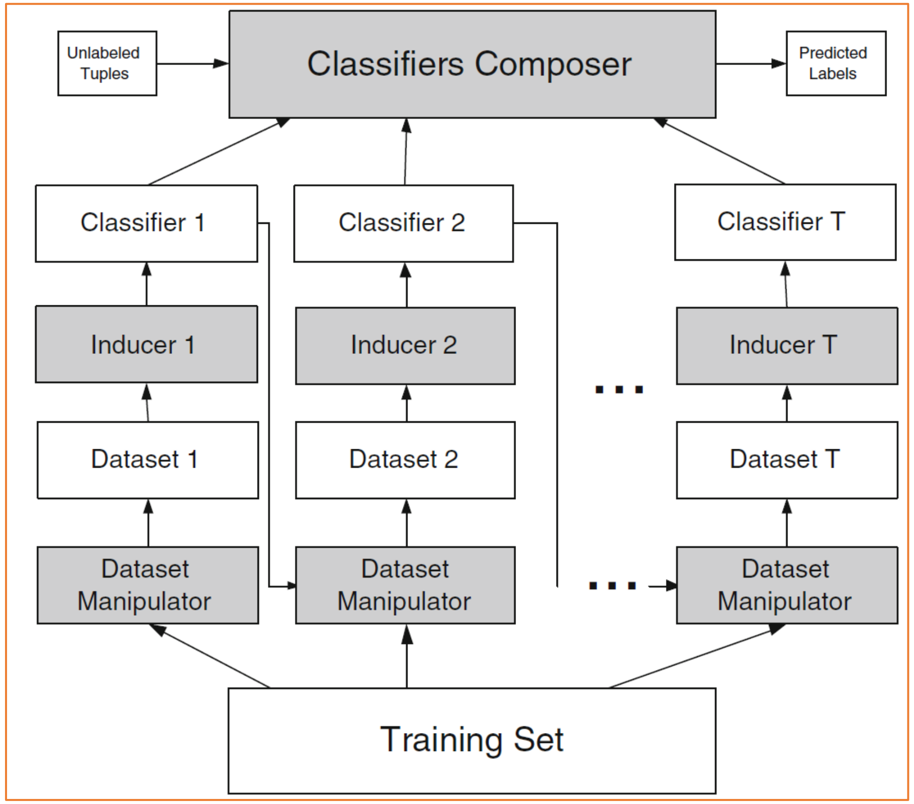
Ensemble Diversity

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

Independent instance selection



Model guided instance selection



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

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

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

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

Why Ensemble?

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

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

  - ✓ 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

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

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


Q & A