Ensemble Learning: Bagging

Pilsung Kang

School of Industrial Management Engineering

Korea University

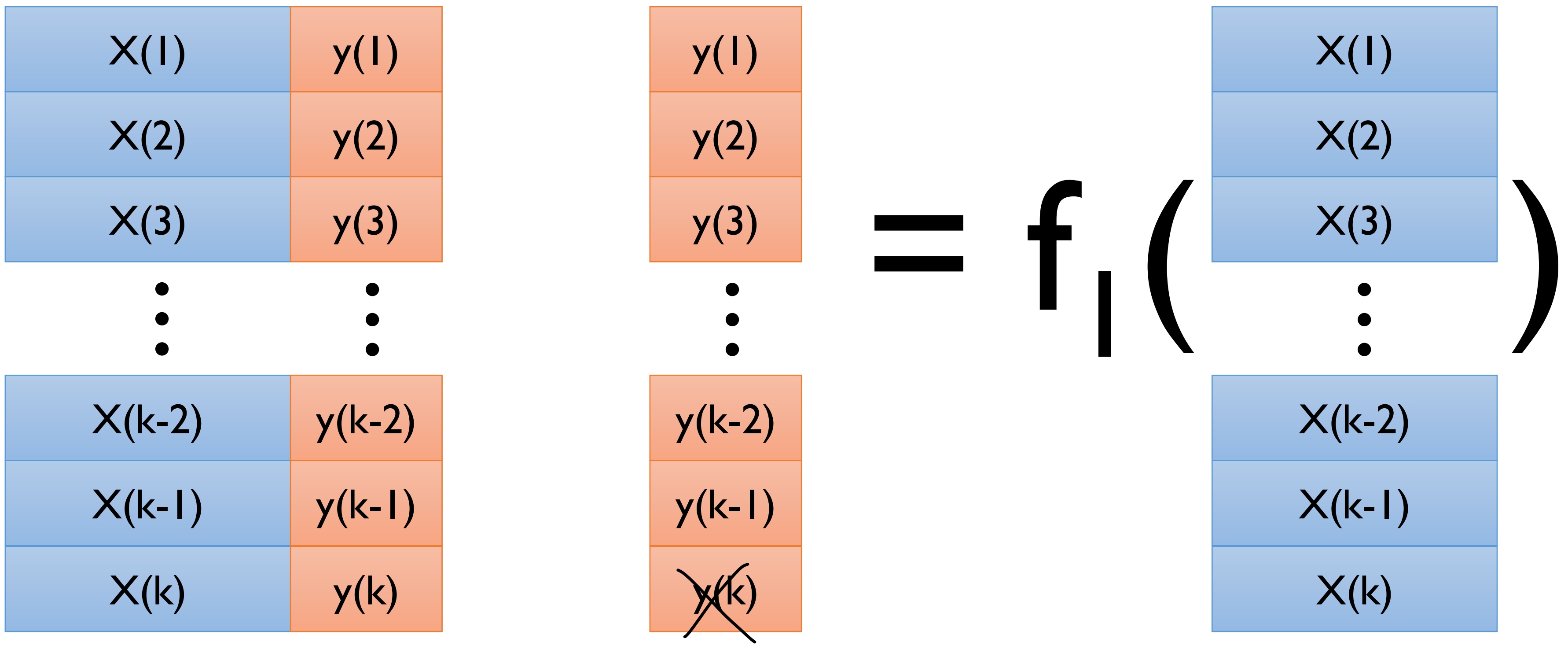
Diversity

Data → Bagging

Modeling

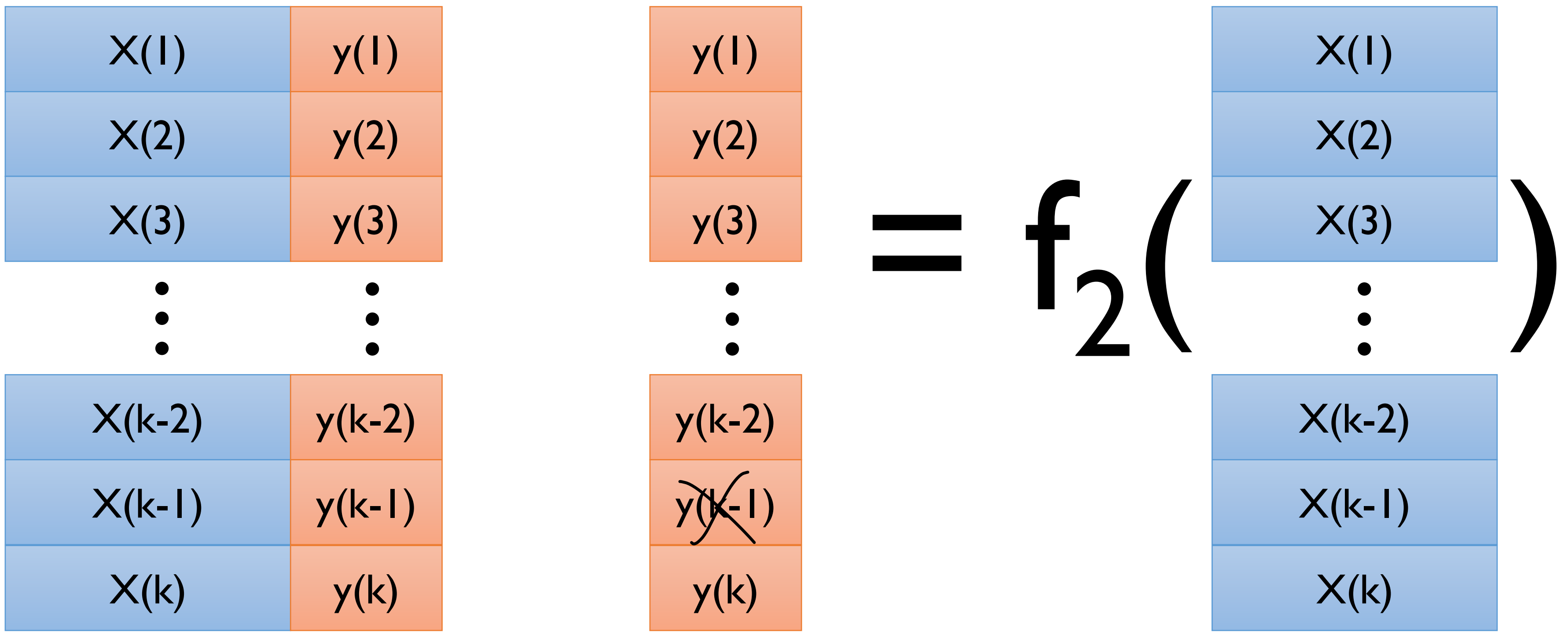
Sampling without Replacement

- K-fold data split
  - ✓ Entire data is split into k blocks; each classifier is trained only on different subset of (k-1) blocks



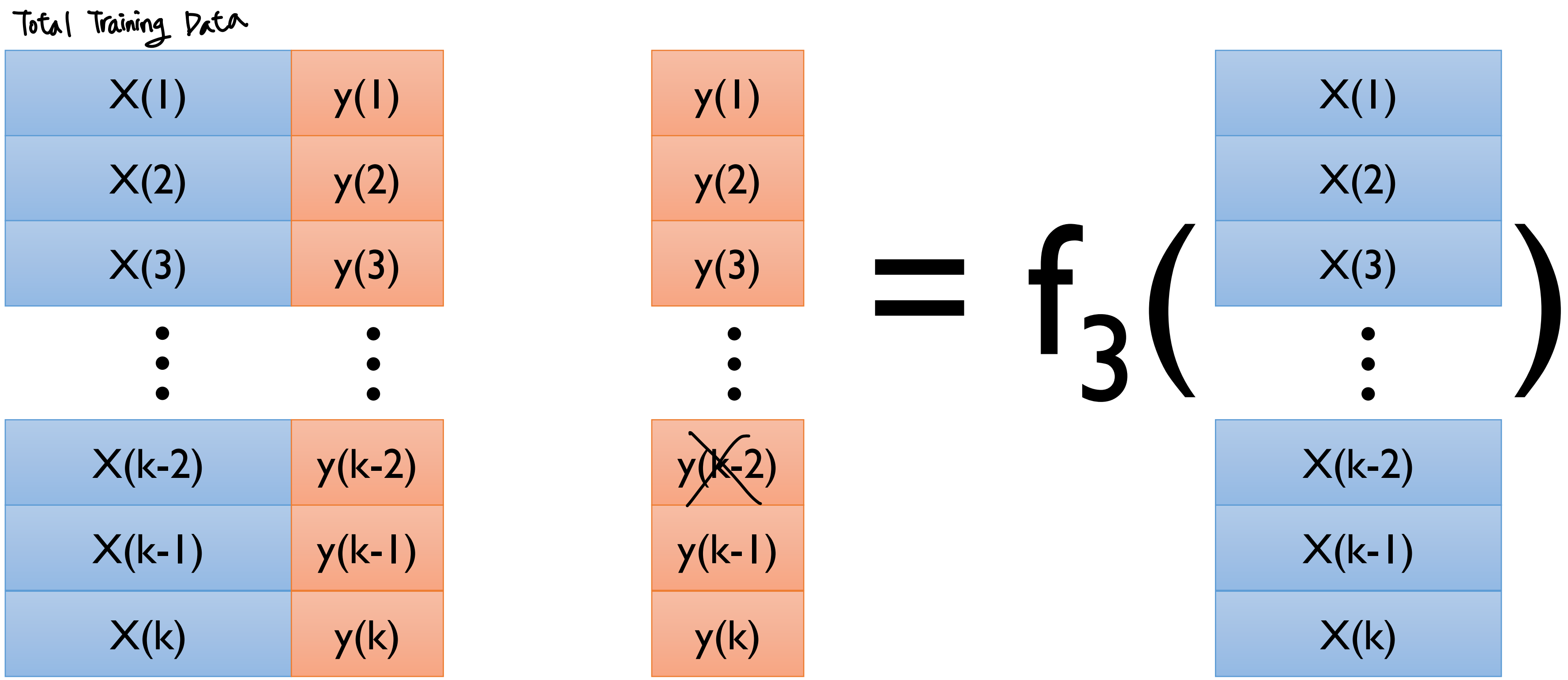
Sampling without Replacement

- K-fold data split
  - ✓ Entire data is split into k blocks; each classifier is trained only on different subset of (k-1) blocks



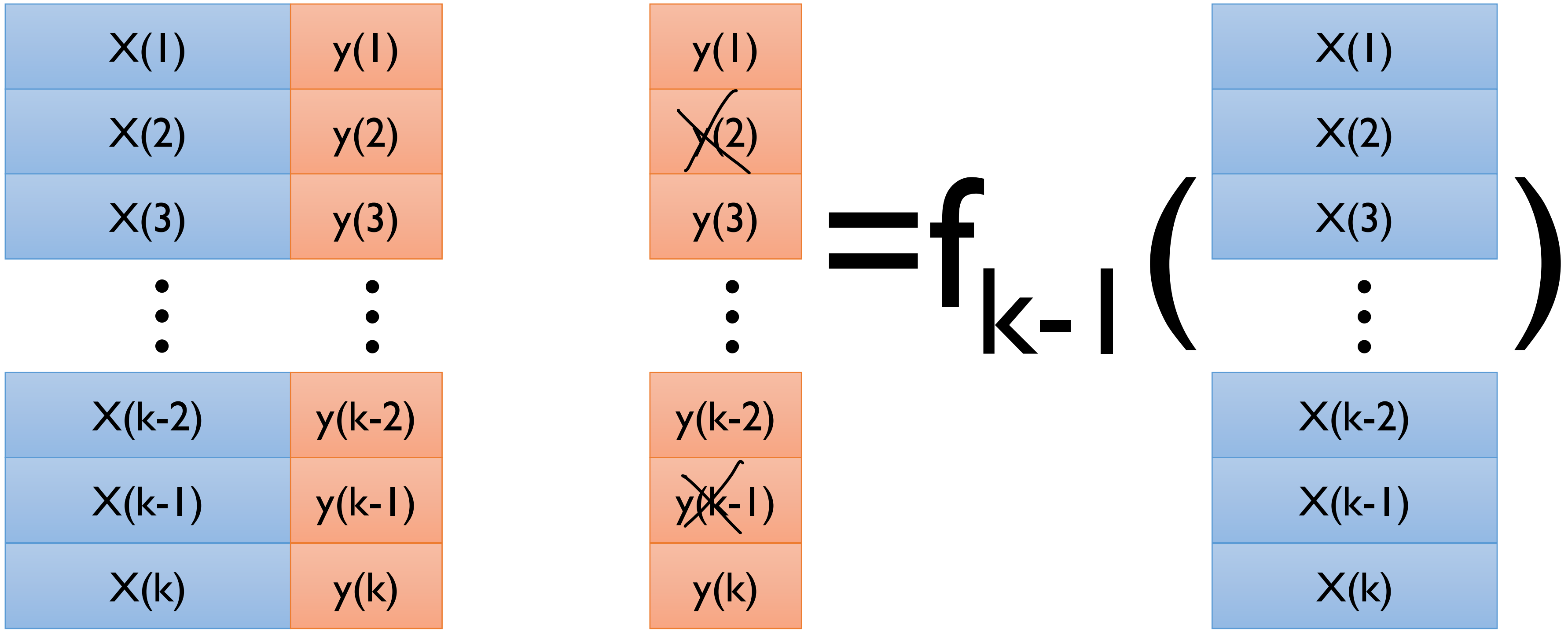
Sampling without Replacement

- K-fold data split
  - ✓ Entire data is split into k blocks; each classifier is trained only on different subset of (k-1) blocks



Sampling without Replacement

- K-fold data split
  - ✓ Entire data is split into k blocks; each classifier is trained only on different subset of (k-1) blocks



Sampling without Replacement

f_i와 f_j는
(k-2) folds는 공통

- K-fold data split
  - ✓ Entire data is split into k blocks; each classifier is trained only on different subset of (k-1) blocks

X(1)	y(1)
X(2)	y(2)
X(3)	y(3)
⋮	⋮
X(k-2)	y(k-2)
X(k-1)	y(k-1)
X(k)	y(k)

$$\begin{bmatrix} \cancel{y(1)} \\ y(2) \\ y(3) \\ \vdots \\ \cancel{y(k-2)} \\ y(k-1) \\ y(k) \end{bmatrix} = f_k\left(\begin{bmatrix} X(1) \\ X(2) \\ X(3) \\ \vdots \\ X(k-2) \\ X(k-1) \\ X(k) \end{bmatrix}\right)$$

Sampling without Replacement

- K-fold data split
  - ✓ Entire data is split into k blocks; each classifier is trained only on different subset of (k-1) blocks
- Final output $\hat{y} = \delta\Big(f_1(\mathbf{x}), f_2(\mathbf{x}), \cdots, f_{k-1}(\mathbf{x}), f_k(\mathbf{x})\Big)$
  - ✓ $\delta(\cdot)$: An aggregation function of individual outputs (ex: simple average)

고려대학교 KOREA UNIVERSITY

DSBA Data Science & Business Analytics

Bootstrap Aggregating: Bagging

Breiman (1996)

- Main Idea
  - ✓ Each member of the ensemble is constructed from a different training dataset
  - ✓ Each dataset is generated by sampling from the total N data examples, choosing N items uniformly at random with replacement 복원추출
  - ✓ Each dataset sample is known as a bootstrap → 원하는 수만큼 데이터셋 구성 가능 ⊕ 분포를 긍정적인 방향으로 왜곡 (기존은 ε에 종속된 모델이 생성됨)

$$y = f(x) + \varepsilon$$

Original Dataset		Bootstrap 1		Bootstrap 2		…	Bootstrap B	
x^1	y^1	x^3	y^3	x^7	y^7		x^9	y^9
x^2	y^2	x^6	y^6	x^1	y^1		x^5	y^5
x^3	y^3	x^2	y^2	x^{10}	y^{10}		x^2	y^2
x^4	y^4	x^{10}	y^{10}	x^1	y^1		x^4	y^4
x^5	y^5	x^8	y^8	x^8	y^8		x^7	y^7
x^6	y^6	x^7	y^7	x^6	y^6		x^2	y^2
x^7	y^7	x^7	y^7	x^2	y^2		x^5	y^5
x^8	y^8	x^3	y^3	x^6	y^6		x^{10}	y^{10}
x^9	y^9	x^2	y^2	x^4	y^4		x^8	y^8
x^{10}	y^{10}	x^7	y^7	x^9	y^9		x^2	y^2

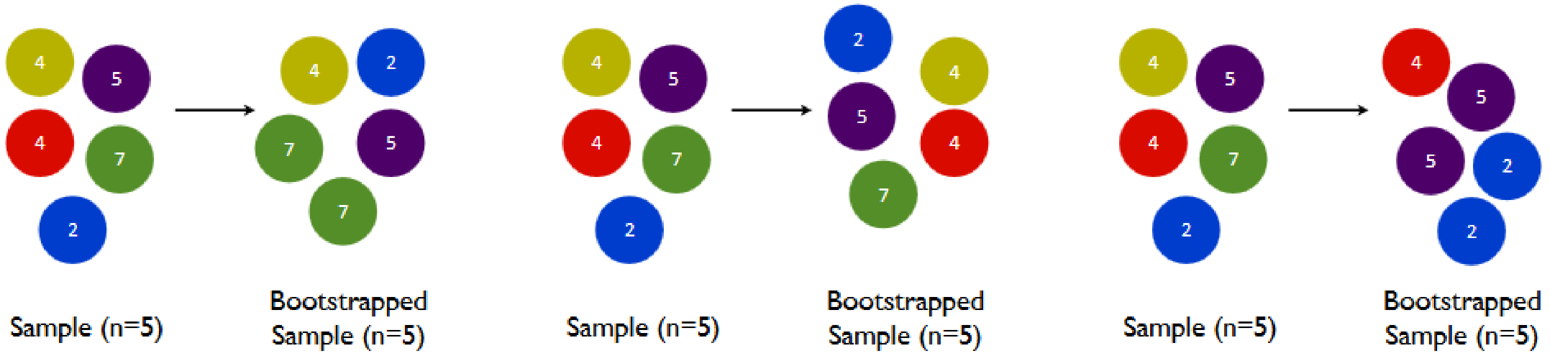
Bootstrap Aggregating: Bagging

- Bagging: Bootstrapp Aggregating
  - ✓ Probability that an instance is not included in a bootstrap

$$p=\left(1-\frac{1}{N}\right)^{N} \rightarrow \lim_{N\rightarrow\infty}\left(1-\frac{1}{N}\right)^{N}=e^{-1}=0.368$$

선택되지 않을 확률

N이 일정 수준으로 크면
2/3는 Bootstrap에 1회 이상 sampling
1/3 〃 0회 sampling → OOB Dataset
⇓
Bagging의 검증관점 Stochastic



Tra 70% → Model
Val 30%
→ Train과 Val의 구성에 따라 결과가 달라짐

  - ✓ Fits well with the models with low bias and high variance

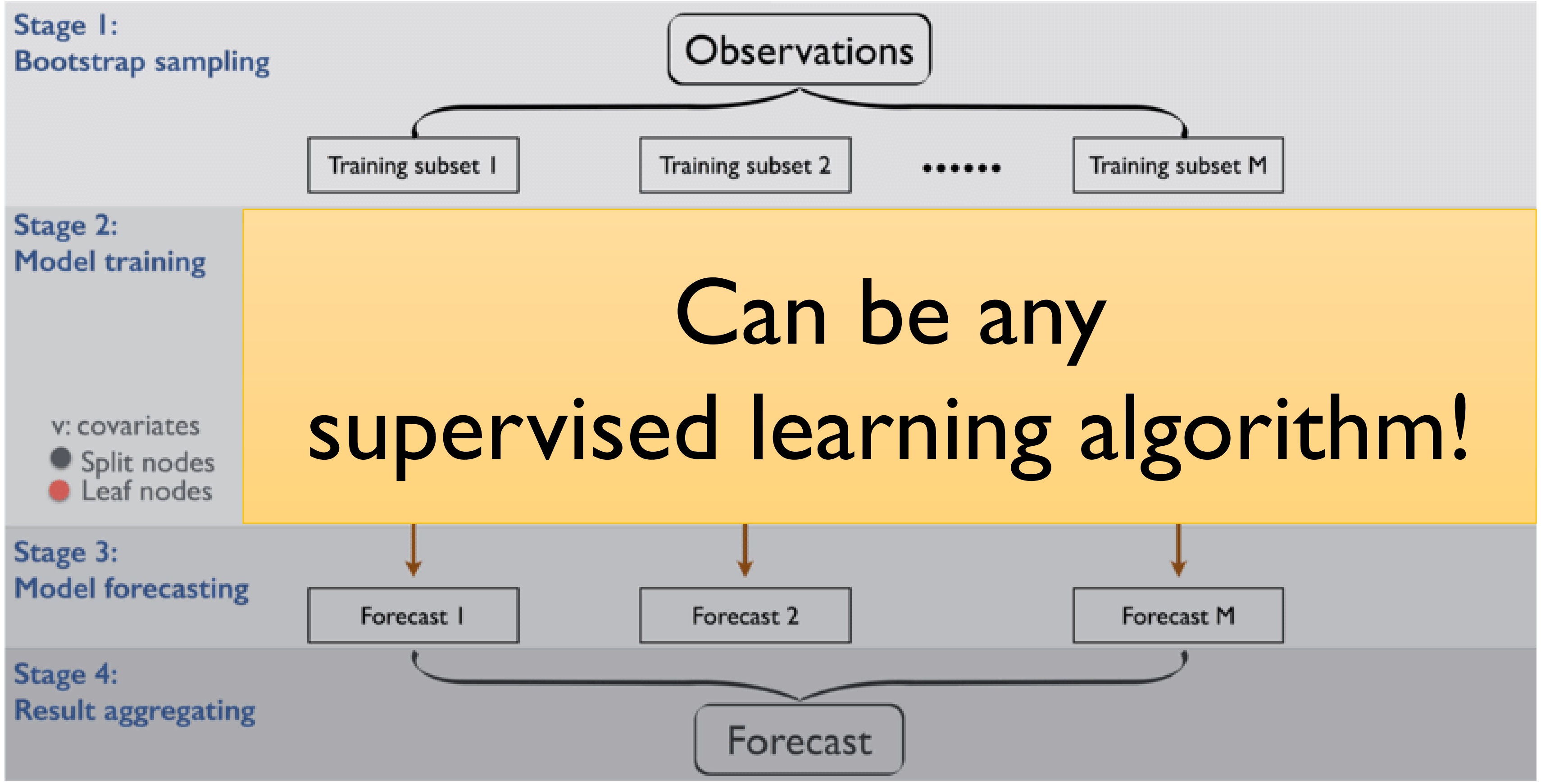
모델의 복잡도가 높은 것들과 조합이 맞음

B1 B2 B3 ...
Train
M1 M2 M3
Random ← Val OOB1 OOB OOB

고려대학교 KOREA UNIVERSITY
DSBA Data Science & Business Analytics

Bootstrap Aggregating: Bagging

- Bagging with Decision Tree ≠ Random Forest → About Data Sampling



ANN
SVM
DT
⋮
↕
LDA
↳ High variance
Low Bias
↓
조합이 어울리지 않음

Bootstrap Aggregating: Bagging

* Classification → Majority voting
* Regression → simple Average

- Result Aggregating
  - ✓ For classification problem
    - Majority voting

$$\hat{y}_{Ensemble} = arg\max_i \left(\sum_{j=1}^{n} \delta(\hat{y}_j = i), \quad i \in \{0, 1\} \right)$$

Xnew

OOB ~~Training~~ Accuracy	Ensemble population	P(y=1) for a test instance	Predicted class label
0.80	Model 1	0.90	1
0.75	Model 2	0.92	1
0.88	Model 3	0.87	1
0.91	Model 4	0.34	0
0.77	Model 5	0.41	0
0.65	Model 6	0.84	1
0.95	Model 7	0.14	0
0.82	Model 8	0.32	0
0.78	Model 9	0.98	1
0.83	Model 10	0.57	1

$$\sum_{j=1}^{n} \delta(\hat{y}_j = 0) = 4$$

$$\sum_{j=1}^{n} \delta(\hat{y}_j = 1) = 6$$

Majority voting

$$\hat{y}_{Ensemble} = 1$$

고려대학교 KOREA UNIVERSITY

DSBA Data Science & Business Analytics

Bootstrap Aggregating: Bagging

- Result Aggregating
  - ✓ For classification problem
    - Weighted voting (weight = training accuracy of individual models)

$$P(y=1 \mid X_{new}) = \frac{0.80+0.75+0.88+0.65+0.78+0.83}{A}$$

$$P(y=0 \mid X_{new}) = \frac{0.91+0.77+0.95+0.82}{A}$$

$$\hat{y}_{Ensemble} = arg\max_i \left(\frac{\sum_{j=1}^{n} (TrnAcc_j) \cdot \delta(\hat{y}_j = i)}{\sum_{j=1}^{n} (TrnAcc_j)}, \quad i \in \{0, 1\} \right)$$

$A = \Sigma$

OOB ~~Training~~ Accuracy	Ensemble population	P(y=1) for a test instance	Predicted class label
0.80	Model 1	0.90	1
0.75	Model 2	0.92	1
0.88	Model 3	0.87	1
0.91	Model 4	0.34	0
0.77	Model 5	0.41	0
0.65	Model 6	0.84	1
0.95	Model 7	0.14	0
0.82	Model 8	0.32	0
0.78	Model 9	0.98	1
0.83	Model 10	0.57	1

$$\frac{\sum_{j=1}^{n} (TrnAcc_j) \cdot \delta(\hat{y}_j = 0)}{\sum_{j=1}^{n} (TrnAcc_j)} = 0.424$$

$$\frac{\sum_{j=1}^{n} (TrnAcc_j) \cdot \delta(\hat{y}_j = 1)}{\sum_{j=1}^{n} (TrnAcc_j)} = 0.576$$

$$\hat{y}_{Ensemble} = 1$$

고려대학교 KOREA UNIVERSITY

DSBA Data Science & Business Analytics

Bootstrap Aggregating: Bagging

- Result Aggregating
  - ✓ For classification problem
    - ▪ Weighted voting (weight = predicted probability for each class)

30가지 이상의 조합방법이 존재

$$\hat{y}_{Ensemble} = arg \max_i \left(\frac{1}{n} \sum_{j=1}^{n} P(y = i), \quad i \in \{0, 1\} \right)$$

Training Accuracy	Ensemble population	P(y=1) for a test instance	Predicted class label
0.80	Model 1	0.90	1
0.75	Model 2	0.92	1
0.88	Model 3	0.87	1
0.91	Model 4	0.34	0
0.77	Model 5	0.41	0
0.65	Model 6	0.84	1
0.95	Model 7	0.14	0
0.82	Model 8	0.32	0
0.78	Model 9	0.98	1
0.83	Model 10	0.57	1

$$\frac{1}{n} \sum_{j=1}^{n} P(y = 0) = 0.375$$

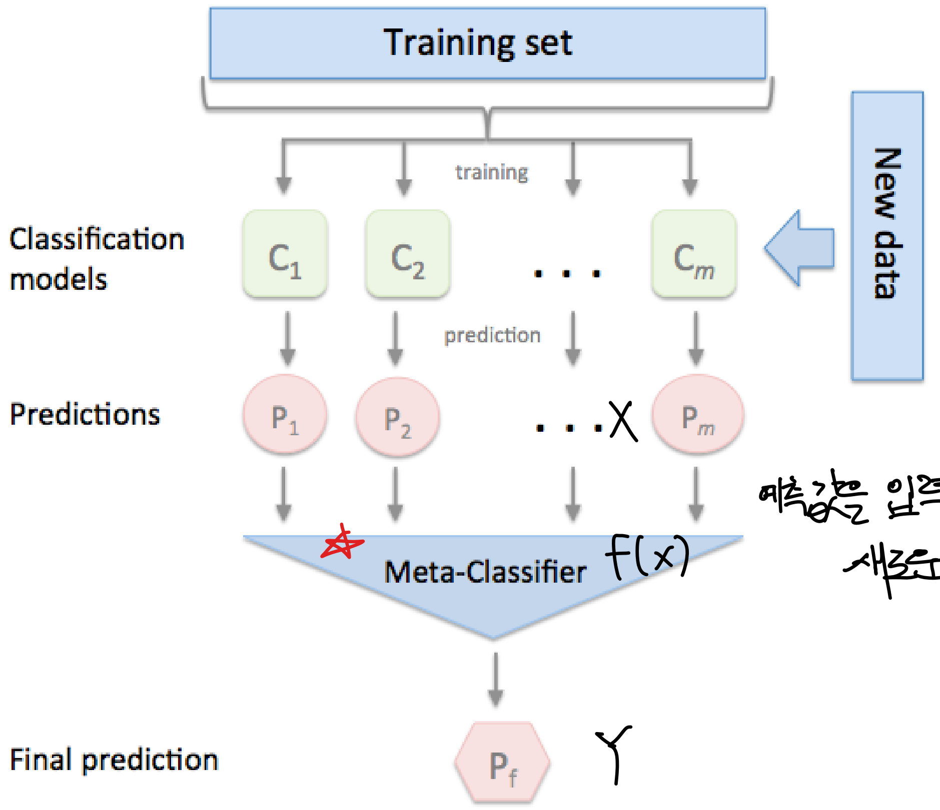
$$\frac{1}{n} \sum_{j=1}^{n} P(y = 1) = 0.625$$

→ 0.98 → 영향력↑ $\hat{y}_{Ensemble} = 1$

→ 0.57

Bootstrap Aggregating: Bagging

- Result Aggregating: Stacking
  - ✓ Use another prediction model to aggregate the results
    - Input: Predictions made by ensemble members
    - Target: Actual true label



예측값을 입력값으로 받아
새로운 모델 생성

$$x \rightarrow \left.\begin{matrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_m(x) \end{matrix}\right\} g(f_1(x), f_2(x), \ldots, f_m(x)) = y$$

Bootstrap Aggregating: Bagging

- Result Aggregating: Stacking
  - ✓ The winner of KDD-cup 2015
    - MOOC dropout prediction

•Jeong-Yoon Lee, Winning Data Science Competitions

Bootstrap Aggregating: Bagging

- Bagging: Algorithm

Algorithm 1 Bagging

Input: Required ensemble size T
Input: Training set $S = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$
for $t = 1$ to T **do**
  Build a dataset S_t, by sampling N items, randomly *with replacement* from S.
  Train a model h_t using S_t, and add it to the ensemble.
end for
For a new testing point (x', y'),
If model outputs are continuous, combine them by averaging.
If model outputs are class labels, combine them by voting.

고려대학교 KOREA UNIVERSITY

DSBA Data Science & Business Analytics

Bootstrap Aggregating: Bagging

- Bagging: Illustration

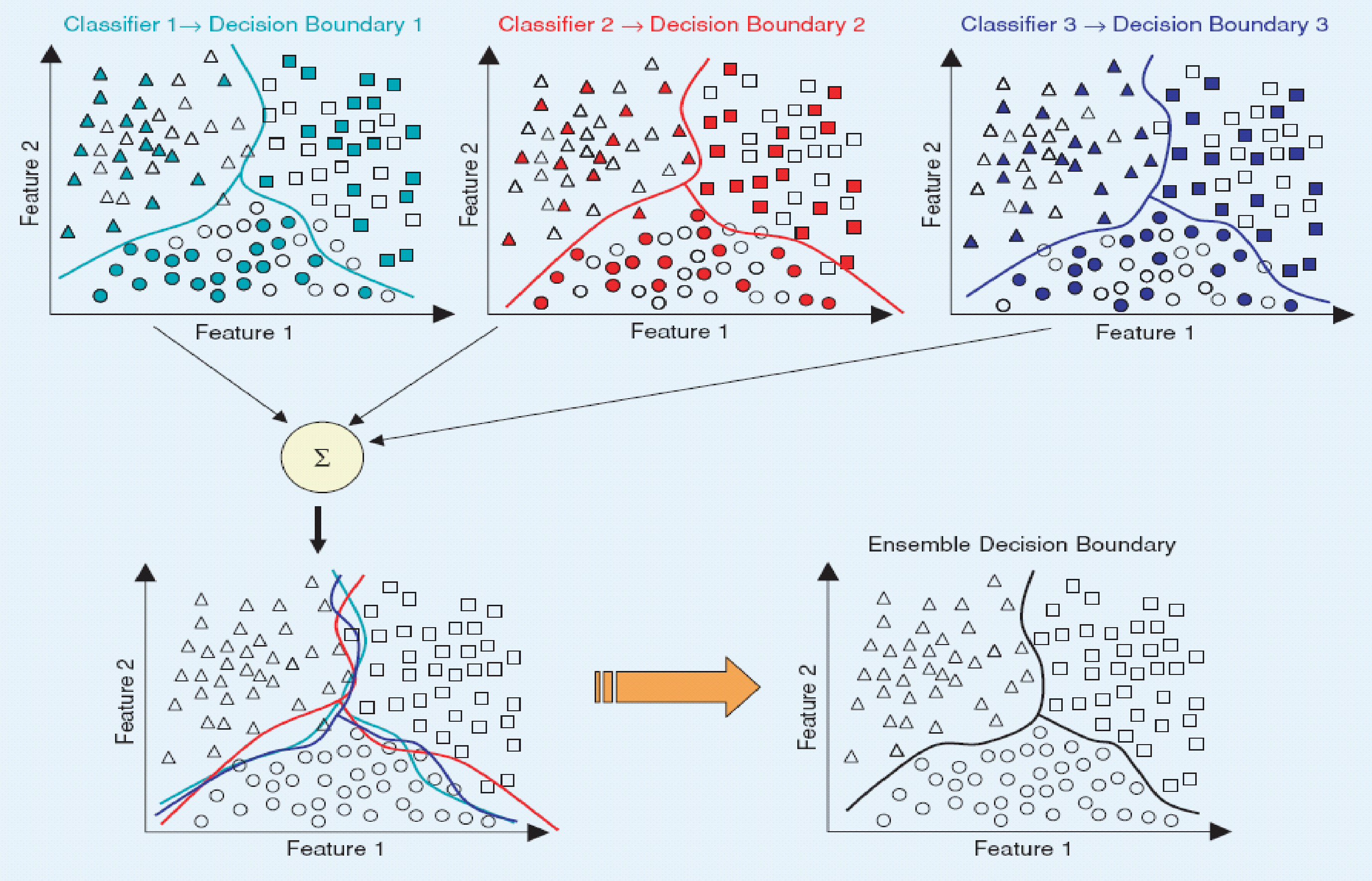
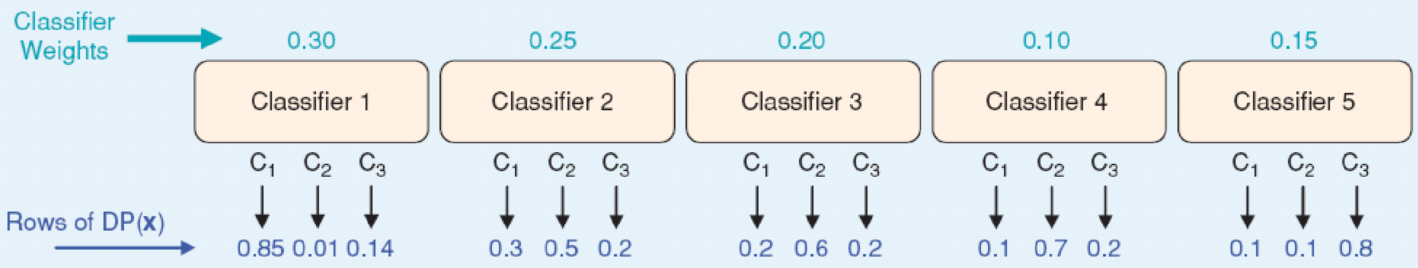


Figure 3. Combining classifiers that are trained on different subsets of the training data.

→ 색깔X → OOB

Bootstrap Aggregating: Bagging

- Aggregation examples

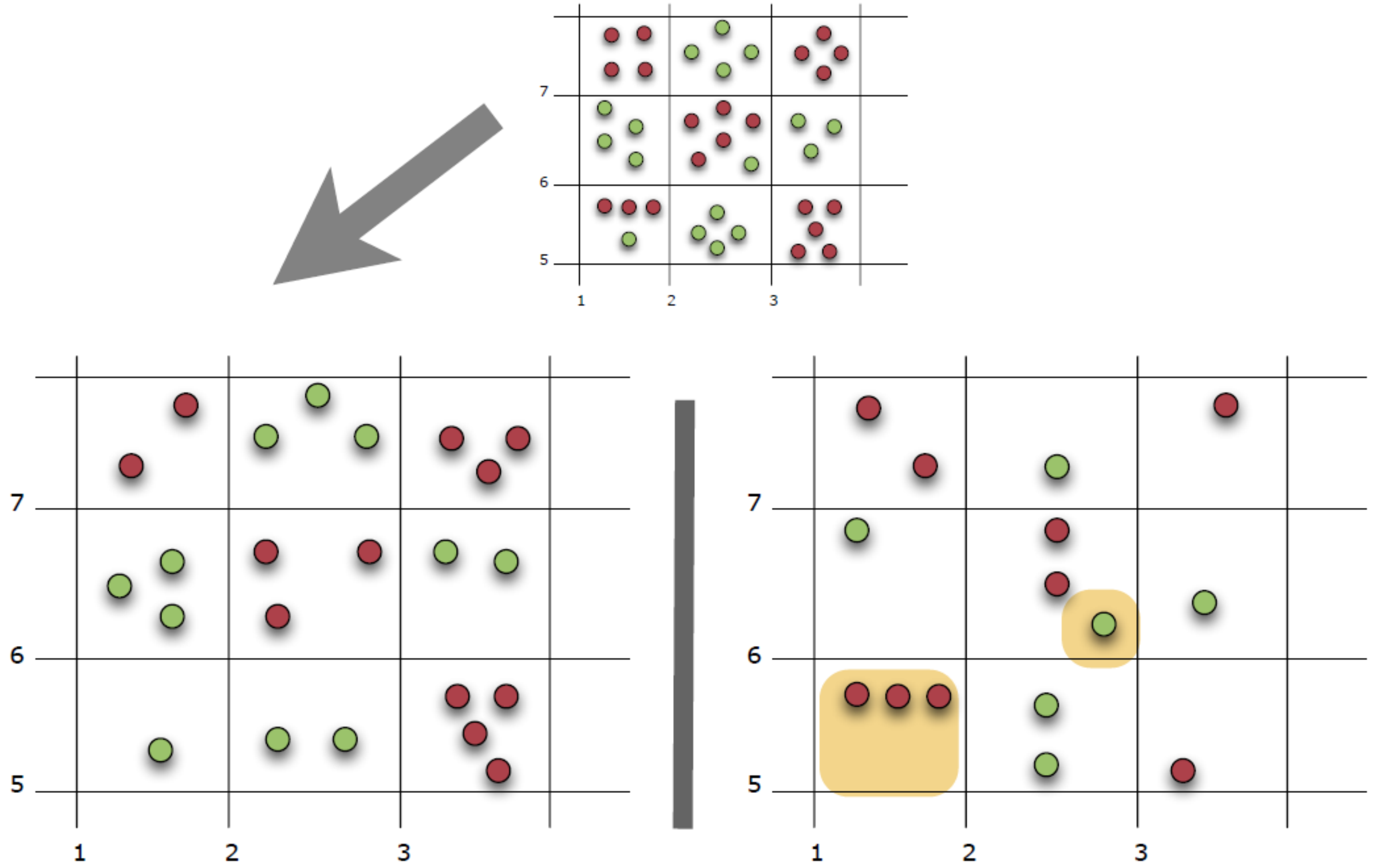


Product Rule:	**C_1: 0.0179,**	C_2: 0.00021,	C_3: 0.0009
Sum Rule:	C_1: 1.55,	**C_2: 1.91,**	C_3: 1.54
Mean Rule:	C_1: 0.310,	**C_2: 0.382,**	C_3: 0.308
Max Rule:	**C_1: 0.85,**	C_2: 0.7,	C_3: 0.8
Median Rule:	C_1: 0.2,	**C_2: 0.5,**	C_3: 0.2
Minimum Rule:	C_1: 0.1,	C_2: 0.01,	**C_3: 0.14**
Weighted Average	**C_1: 0.395,**	C_2: 0.333,	C_3: 0.272
Majority Voting:	C_1: 1,	**C_2: 3,**	C_3: 1 Vote
Weighted Majority Voting:	C_1: 0.3,	**C_2: 0.55,**	C_3: 0.15 Votes
Borda Count:	C_1: 5	**C_2: 6,**	C_3: 4 Votes

Figure 15. Example on various combination rules.

Bootstrap Aggregating: Bagging

- Out of bag error (OOB Error)
  - ✓ Use the training instances that are not sampled for validation



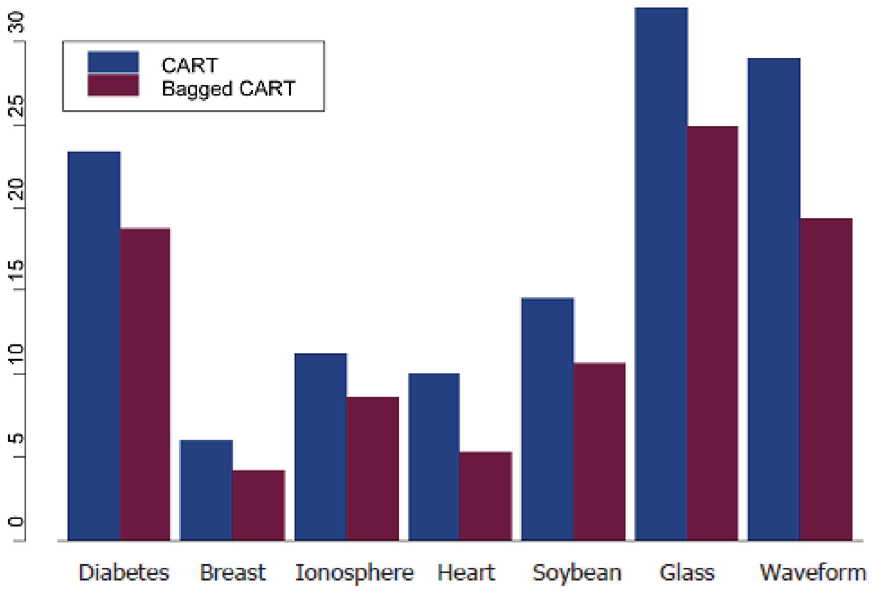
http://www.slideshare.net/oxdata/jan-vitek-distributedrandomforest522013

Bootstrap Aggregating: Bagging

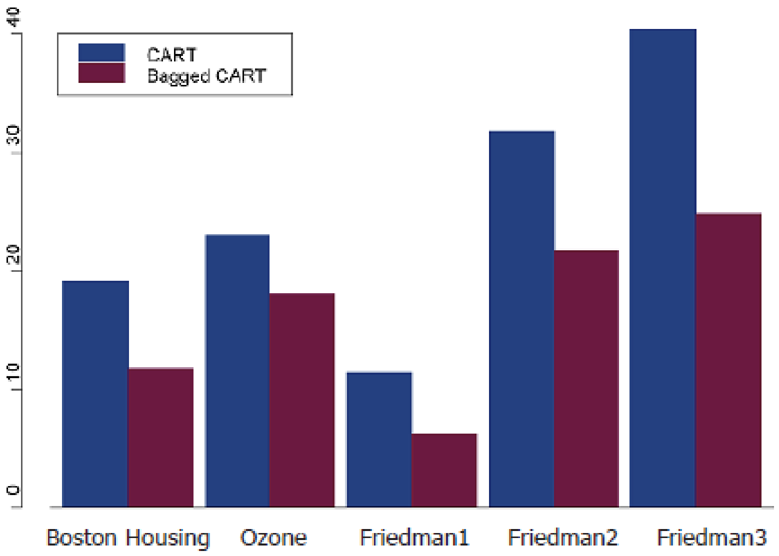
- Bagged Trees vs. Single Tree

Error Rate

Classification



Regression




Q & A