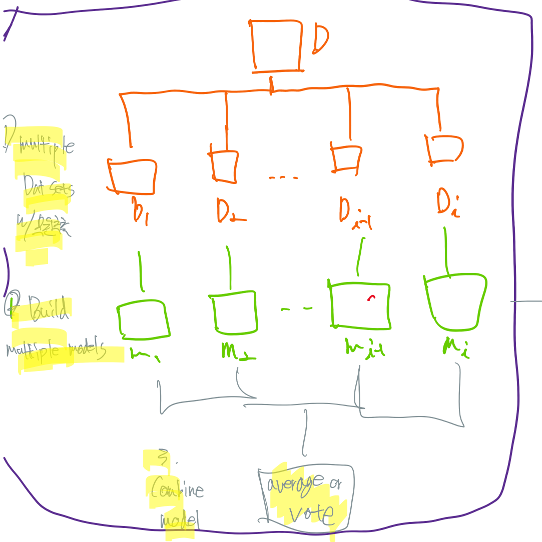
Bagging RF

2019년 2월 21일 목요일 오후 1:39

1. Bagging

- k bootstrap sample ($D_1, D_2, D_3 \cdots D_k$)
- 복원추출
- Train distinct model on each D_i
- Test a new instance by majority vote or average



→ 문제해결을위해
여러전문가에게문의!
→ model diversity

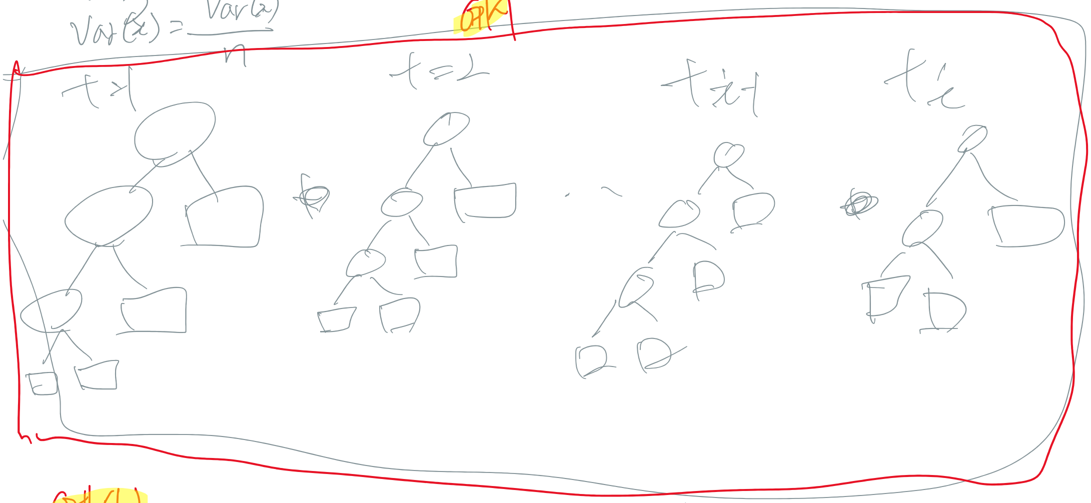
2. Random Forest

- Decision tree 모델을 bagging 하며
full-grown tree 로 low bias 의 여러 DT를 모음

- 다양한 DT가 모이면서 high variance(overfitting)을
줄여줌

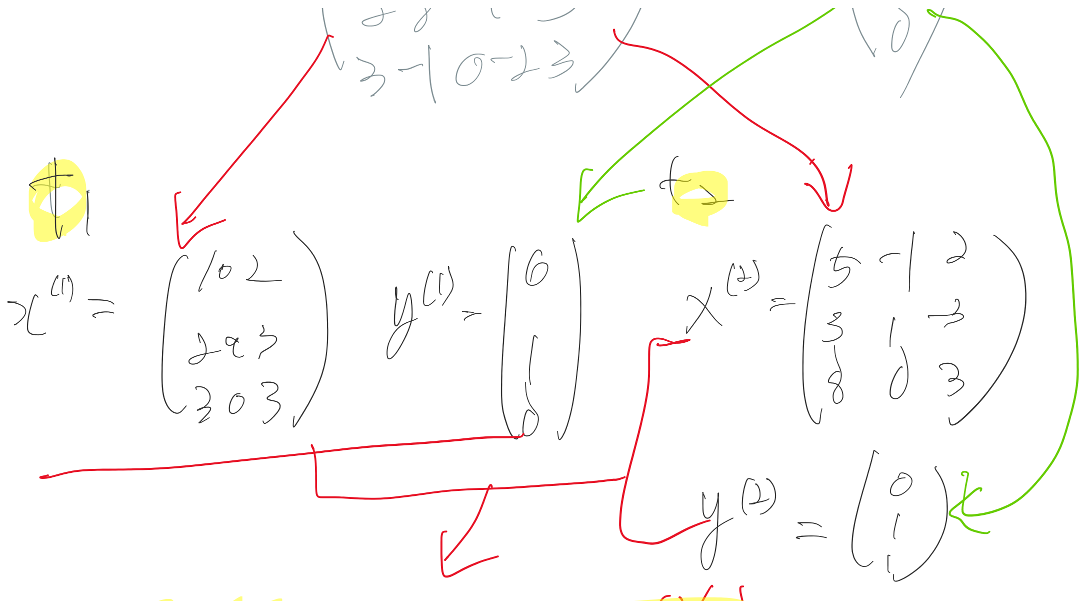
$$\overrightarrow{\mathrm{Var}(x)} = \frac{\mathrm{Var}(x)}{n}$$

OK



예시 -

$$x = \begin{pmatrix} 1 & 5 & 0 & -1 & 2 \\ 0 & 3 & 9 & 1 & -3 \\ 2 & 8 & 9 & 0 & 3 \\ 3 & -1 & 0 & -2 & 3 \end{pmatrix} \qquad y = \begin{pmatrix} 6 \\ 1 \\ 1 \\ 0 \end{pmatrix}$$



$x^{(1)} = \begin{pmatrix} 1 & 0 & 2 \\ 2 & 9 & 3 \\ 3 & 0 & 3 \end{pmatrix}$ $y^{(1)} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ $X^{(2)} = \begin{pmatrix} 5 & -1 & 2 \\ 3 & 1 & -3 \\ 8 & 0 & 3 \end{pmatrix}$

$y^{(2)} = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$

① Bootstrapping trainning set(✓)

+

features (✓)

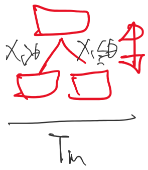
Low bias (full-growed tree) but hig variance → overfitting 가능성 높음

↓

Low variance

2. Feature Importance

$$\text{Importance}(X_i) = \frac{1}{M}\sum_{m=1}^{M} \Delta I(X_i, T_m)$$

- X = feature
- M = 모델(Tree)개수
- T_m = m번째 tree
- ΔI =  ΔI : error cap

⇒ ΔI를 가장 크게하는 feature 순으로 중요도 나레이가 계산됨