Week 5

Sunday, August 10, 2025 5:23 PM

1. Ensemble

: 여러 Base 모델 예측을 통합하여 예측정확성 향상 { 모델독립 / Better than random.

$$e_{ensemble} = \sum_{i=N/2}^{N} \binom{N}{i} e^{i}(1-e)^{N-i}$$: 앙상블 오류율

Decision Tree – Low computation & Non-parametric → Ensemble 적합
데이터 분포 전제 X

2. Random Forest

: Diversity & Randomness

변수 무작위 선택 → Random Subspace.

여러개 Training data 구축 → Bagging

Bagging : Bootstrap Aggregating – 각 Bootstrap 샘플로부터 생성된 모델 합치기.

복원추출을 통해 원래 데이터 크기만큼 샘플링.

◦ Classification

- Majority Voting : $Ensemble(\hat{y}) = \arg\max_i \left(\sum_{j=1} I(\hat{y}_j = i) \right),\ i \in \{0,1\}$

- Weighted Voting : $Ensemble(\hat{y}) = \arg\max_i \left(\frac{\sum (TrainAcc_j) \cdot I(\hat{y}_j = i)}{\sum (TrainAcc_j)} \right),\ i \in \{0,1\}$

or

$$= \arg\max_i \left(\frac{1}{n} \sum P(y = i),\ i \in \{0,1\} \right)$$

Random Subspace

① 원래 변수 중 소수의 입력변수 무작위 선택

② 선택된 입력변수 중 분할될 변수 선택 → 분기점 탐색시, 원래보다 적은 변수 임의 선택하여 그 변수들만 고려

③ full-grown tree 될때까지 반복.

∴ Generalization Error $\le \frac{\bar{\rho}(1-s^2)}{s^2}$: upper bound ($\bar{\rho}$: tree 사이의 평균 상관관계 / s : 올바로 예측한 tree와 잘못 예측한 tree수 차이 평균)

Random forest는 Tree↑ ~ 큰수의 법칙으로 error가 수렴

① 개별 tree 정확도 $\propto s$

② Bagging + Random Subspace → 모델간 독립성, 무작위성 최대화하여 ρ 감소

② Bagging + Random Subspace → 모델간 독립성, 랜덤성 극대화하여 ↑

3. Feature Importance.

원래 데이터에 대해 OOB Error 계산 → 특정변수 섞어도 섞은 데이터의 OOB Error 계산
r_i Y_i e_i

$d_i = e_i - r_i$ $\begin{cases} \bar{d} = \frac{1}{t}\sum d_i \\ S_d^2 = \frac{1}{t-1}\sum (d-\bar{d})^2 \end{cases}$ ⇒ X_i Importance $v_i = \frac{\bar{d}}{S_d}$ (분모: 표준편차)

4. Hyperparams.

- Decision Tree 개수 (충분한 ≥2000)
- 노드별로 무작위 선택되는 변수의 수 $\begin{cases} \text{Classification} : \sqrt{\text{변수의 수}} \\ \text{Regression} : \text{변수의 수} / 3 \end{cases}$