Ensembles

2019년 8월 6일 화요일 오후 2:10

1. 집단지성의 힘

❖ 앙상블의 효과

- 이론적으로는 M개의 개별 모델을 결합한 앙상블의 경우 M개의 개별 모델의 평균 오류의 1/M 수준으로 오류가 감소함 (가정: 각 모델은 서로 독립)

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

- 위 가정은 현실세계에서 지켜지지 않는 경우가 많음
- 현실적으로는 M개의 개별 모델을 결합한 앙상블의 경우 개별 모델의 평균 오류보다는 최소한 같거나 낮은 오류를 나타내는 것을 증명할 수 있음

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

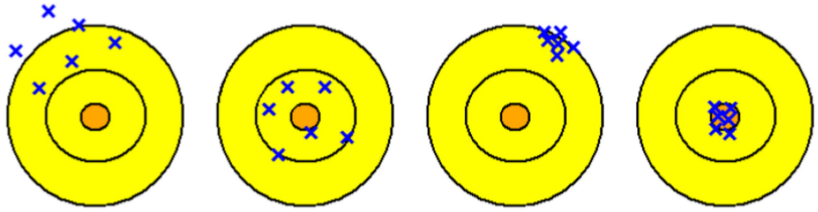
$$E_{Ensemble} \leq E_{Avg}$$

✓ 즉, 1등 모델보다 성능이 우수함을 입증할 수는 없으나 개별 모델들의 평균치보다는 항상 우수하거나 같은 성능을 나타냄

❖ 앙상블의 목적: 다수의 모델을 학습하여 오류의 감소를 추구

- 분산의 감소에 의한 오류 감소: 배깅(Bagging), 랜덤 포레스트(Random Forest)
- 편향의 감소에 의한 오류 감소: 부스팅(Boosting)
- 분산과 편향의 동시 감소: Mixture of Experts

❖ 편향(Bias)과 분산(Variance)의 크기에 따른 모델의 구분



Bias	High	Low	High	Low
Variance	High	High	Low	Low

- 낮은 모델 복잡도: 높은 편향 & 낮은 분산
  - ✓ Logistic regression, LDA, k-NN with large k, etc.
- 높은 모델 복잡도: 낮은 편향 & 높은 분산
  - ✓ DT, ANN, SVM, k-NN with small k, etc.