

# 데이터마이닝 (**Data Mining**)

Chapter 1.2 랜덤 포레스트(Random Forest)

## 랜덤 포레스트

---

- 배깅과 부스팅보다 더 많은 무작위성을 주어 약한 학습기들을 생성한 후 이를 선형결합하여 최종 학습기를 생성
- 이론적 설명이나 최종 결과에 대한 해석은 어렵지만 예측력은 매우 높음
- 특히 입력변수의 개수가 많을 때에는 배깅이나 부스팅과 비슷하거나 더 좋은 예측력을 보이는 경우가 많고, 조율모수가 없어서 실제 자료분석에 쉽게 사용됨
- 무작위성을 극대화하기 위해 붓스트랩과 더불어 입력변수들에 대한 무작위 추출을 결합

- 
- $\mathcal{L} = \{(x_i, y_i)\}_{i=1}^n, x_i \in \mathbb{R}^p$  : 훈련자료
  - 알고리즘
    1.  $n$  개의 자료를 이용한 붓스트랩 표본  $\mathcal{L}^* = \{(x_i^*, y_i^*)\}_{i=1}^n$  을 생성.
    2.  $\mathcal{L}^*$ 에서 입력변수들 중  $k(k \ll p)$ 개만 무작위로 뽑아 의사결정나무를 생성.
    3. 생성된 의사결정나무들을 선형결합하여 최종학습기를 생성
  - 붓스트랩 표본의 개수는 적절히 커야 함
  - 선형결합 방식 : 회귀의 경우 평균, 분류의 경우 투표 적용