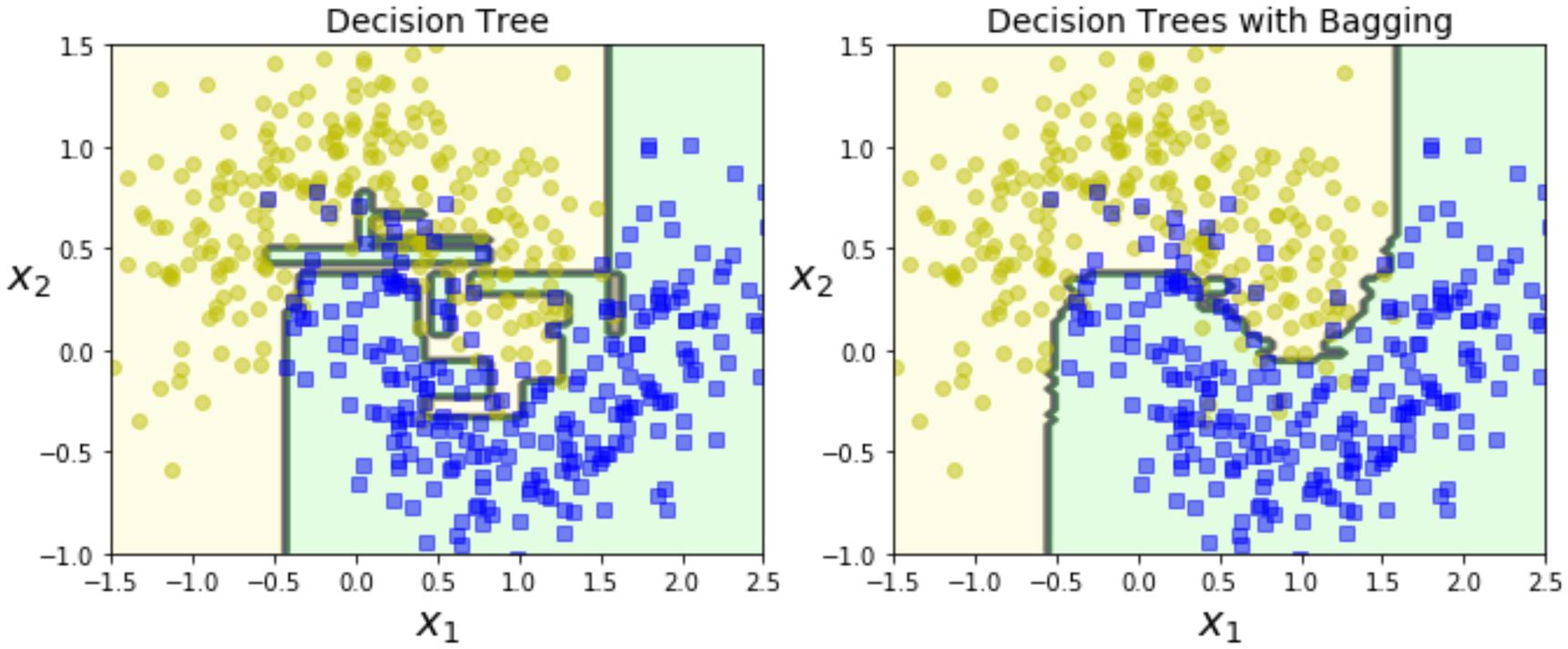
앙상블 학습과 랜덤 포레스트

A/S

둘러보기

1. p251 [그림 7-5] : 어떤 걸 보고 "결정 트리, 배깅을 사용한 결정" 중에서 더 낫다고 말하는 것인가?
2. p255 : 엑스트라 트리가 랜덤 포레스트보다 빠르다고 했는데 왜 그런가?
3. p255 : 엑스트라 트리의 무작위 분할의 장점을 설명했는데 단점은?
4. 배깅, 부스팅, 스태킹 비교



왼쪽은 단일 결정 트리의 결정 경계, 오른쪽은 500개의 트리를 사용한 배깅 앙상블의 결정 경계이다. 왼쪽보다 오른쪽(배깅 앙상블)이 더 일반화가 잘 되어 있어 더 낫다고 말할 수 있다. 근거는 앙상블은 비슷한 편향에서 더 작은 분산을 만들어서 결정 경계가 왼쪽에 비해 덜 불규칙하다는 점에서 알 수 있다.

4. 랜덤 포레스트

엑스트라 트리가 일반적인 랜덤 포레스트보다 작업 속도가 빠른 이유

- 부트스트랩핑의 유무

  랜덤 포레스트는 배깅의 기법 중 하나로, 부트스트랩핑을 기반으로 Weak Tree를 생성한다. 그렇기 때문에 각 Weak Tree 가 다른 분포의 데이터를 학습하고 그것이 집합되었을 때 앙상블 효과를 보인다.

  반면에 엑스트라 트리는 부트스트래핑을 하지 않고, 모든 원래 데이터를 그대로 가져다 쓴다.

  랜덤 포레스트 -〉 복원추출, 엑스트라 트리 -〉 비복원추출

- Split 선택 기준

  랜덤 포레스트는 주어진 샘플에 대해서 모든 변수에 대한 정보 이득을 계산하고, 그 중에서 설명력이 가장 높은 변수를 Split Node 로 선택한다. 이 과정에서 모든 변수에 대한 정보 이득을 계산하므로 시간 복잡도가 높아진다. 반면 엑스트라 트리는 무작위로 변수를 선정한다. 이것이 계산 복잡도를 낮출 수 있다.

4. 랜덤 포레스트

엑스트라 트리 단점

시간 복잡도를 최소화하기 위해 무작위성을 극단적으로 높이면서 몇몇 특성의 중요도를 배제할 수 있다는 단점이 있다.

배깅, 부스팅, 스태킹

	부트스트랩	학습 방식	다른 분류기에 영향	Overfitting 문제
배깅	O	병렬 복원추출	X	X
부스팅	O	순차 복원추출	O	X
스태킹	X	크로스 벨리데이션	X	O (기본적인 방법의 경우)