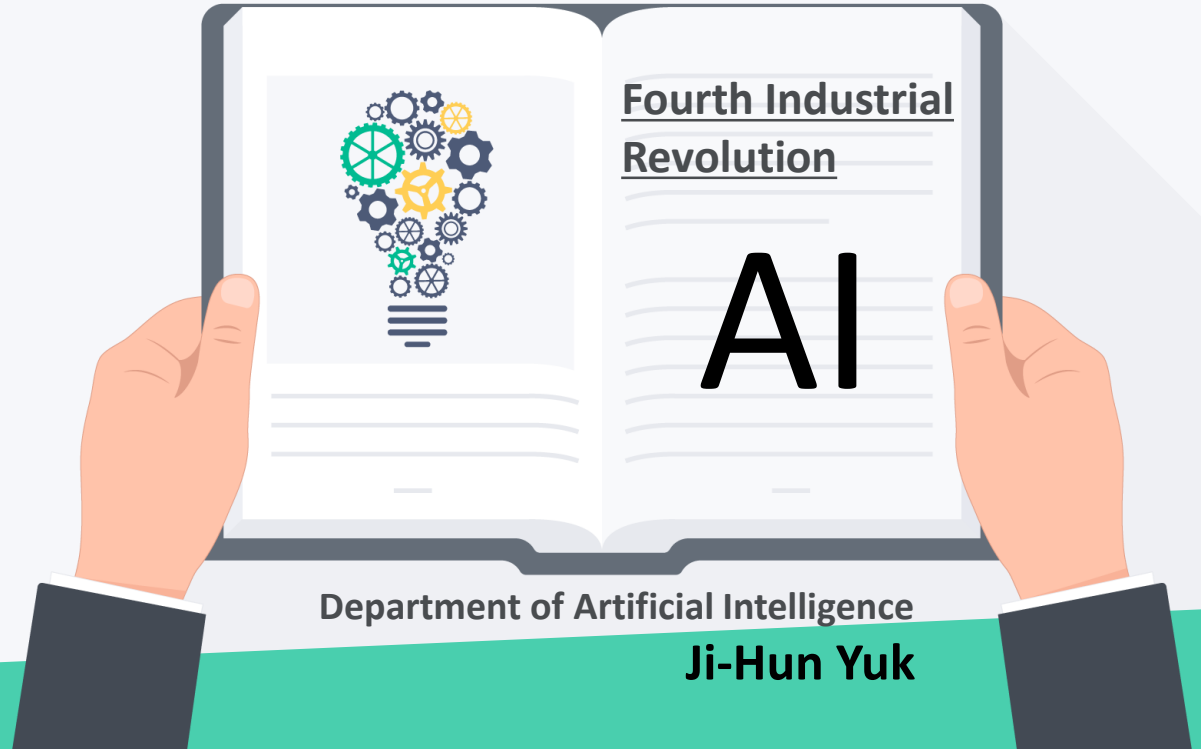
What is Artificial Intelligence(AI)

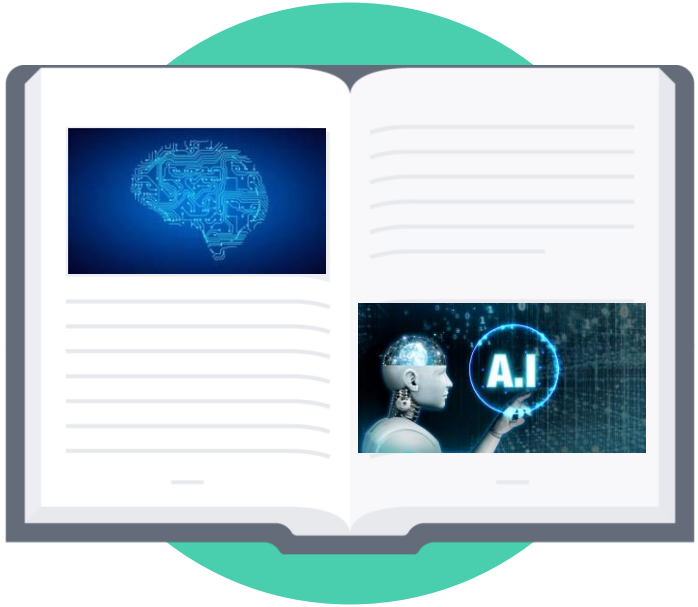
affective.AI Lab



Department of Artificial Intelligence
Ji-Hun Yuk

20 July 2023

Table of contents



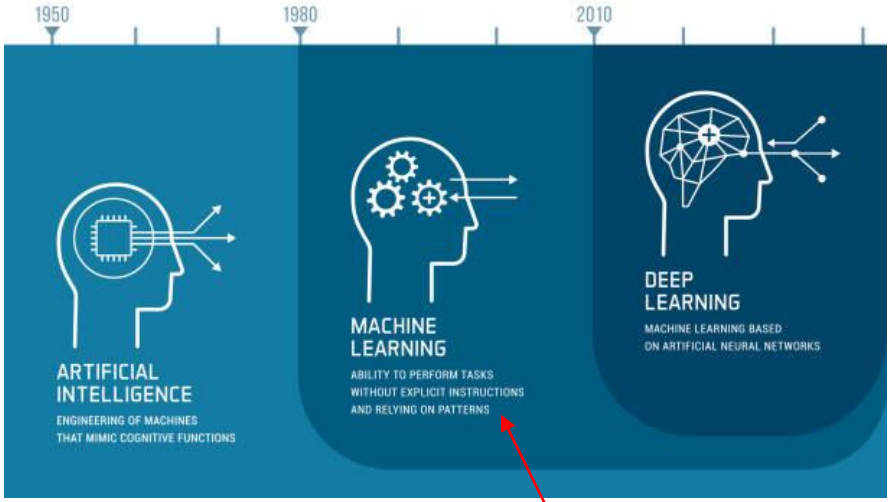
목차

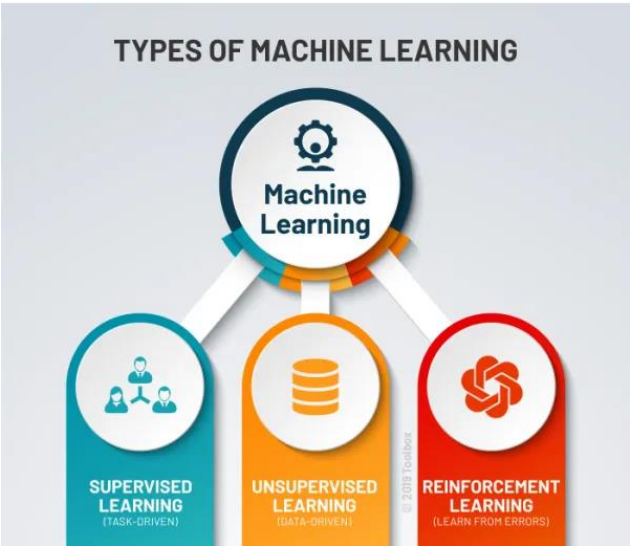
1. What is AI?

Definition:

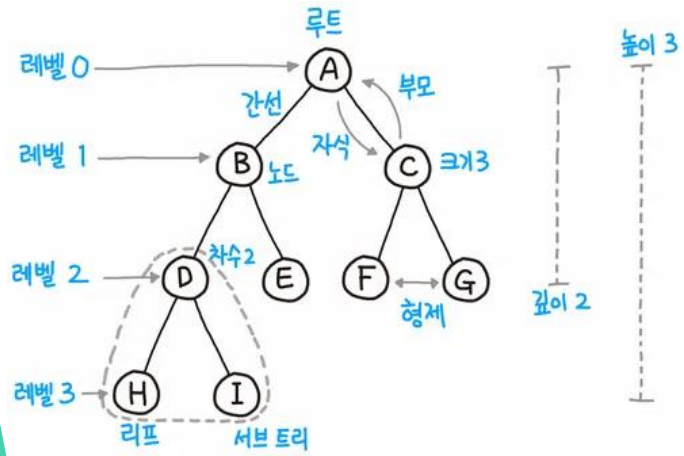
Artificial intelligence (AI) is intelligence—perceiving, synthesizing, and inferring information—demonstrated by machines, as opposed to intelligence displayed by humans or by other animals.

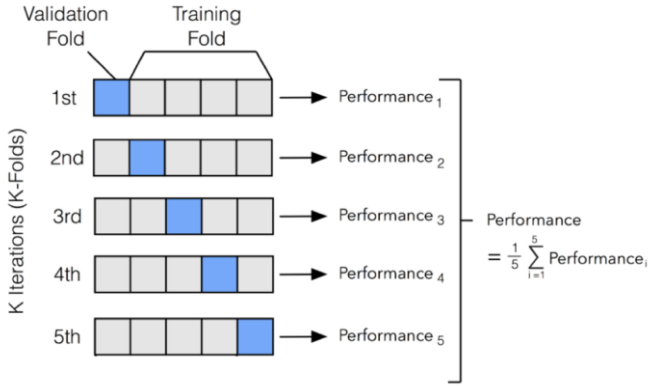
2. What is Machine Learning?

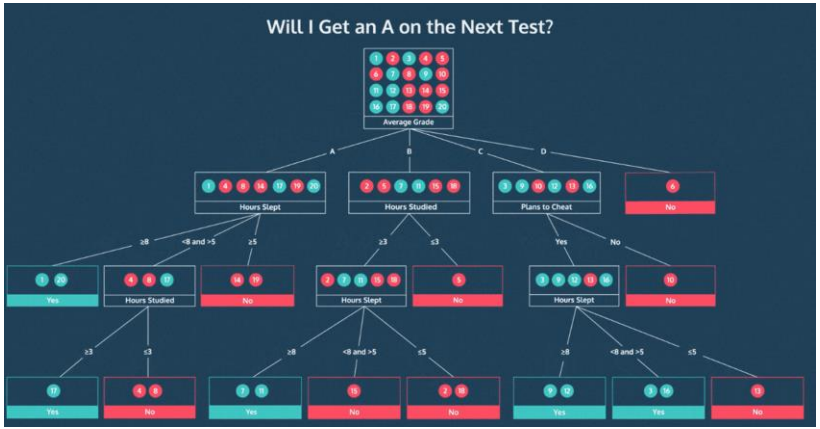




3-0. Weekly study
Ch5. 트리 알고리즘





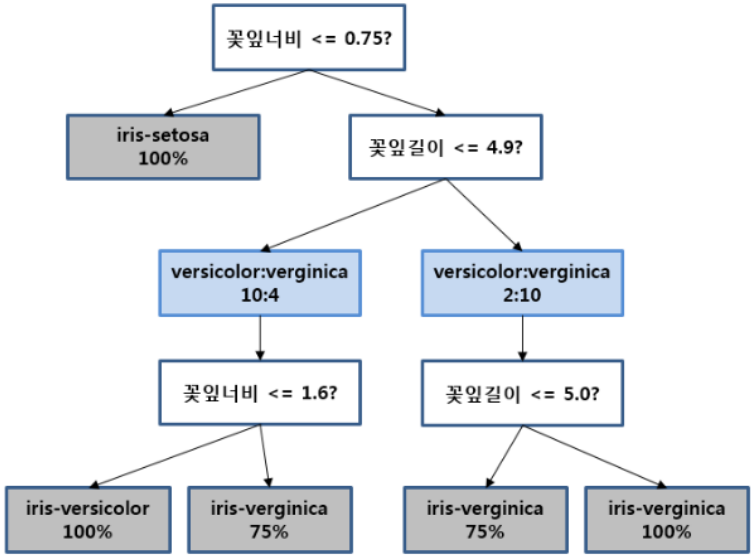


20 July 2023

3-1. 결정 트리

- 예 / 아니오에 대한 질문을 이어나가면서 정답을 찾아 학습하는 알고리즘.
- 비교적 예측 과정을 이해하기 쉽고 성능도 뛰어남.
- 특정 기준에 따라 데이터를 구분함.

- Leaf 노드가 순도 100%의 한 가지 카테고리만 가지게 되는 최대 트리를 형성하게 되면 새로운 데이터에 적용할 때 과대적합 문제가 발생하여 일반화 성능이 떨어진다. -> 가지치기



3-1. 결정 트리

$$I(A) = 1 - \sum_{k=1}^{m} p_k^2 \qquad E = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

- 불순도: 결정 트리가 최적의 질문을 찾기 위한 기준.
  (e.g. 지니 불순도, 엔트로피 불순도)
- 정보 이득: 부모 노드와 자식 노드의 불순도 차이.
  결정 트리 알고리즘은 정보 이득이 최대화되도록 학습.

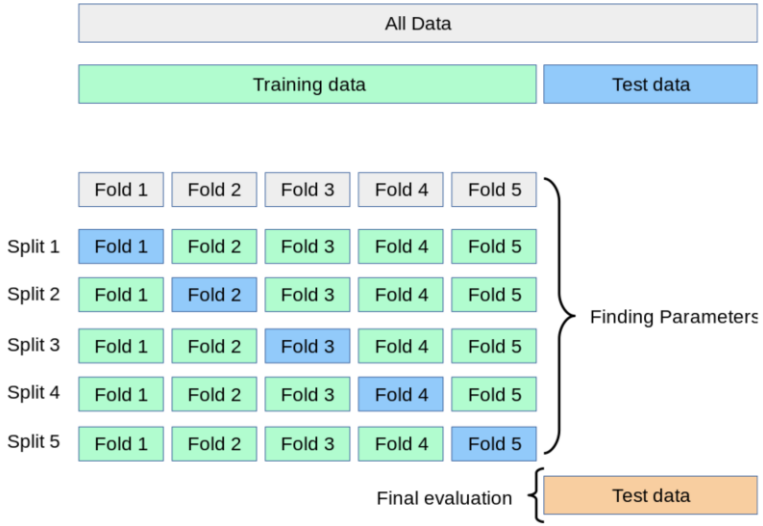
지니 불순도 = 1 − (음성 클래스 비율² + 양성 클래스 비율²)

엔트로피 불순도 = −음성 클래스 비율 × $\log_2$(음성 클래스 비율) − 양성 클래스 비율 × $\log_2$(양성 클래스 비율)

부모의 불순도 − (왼쪽 노드 샘플 수 / 부모의 샘플 수) × 왼쪽 노드 불순도 −
(오른쪽 노드 샘플 수 / 부모의 샘플 수) × 오른쪽 노드 불순도 = 정보 이득

3-2. 교차 검증과 그리드 서치

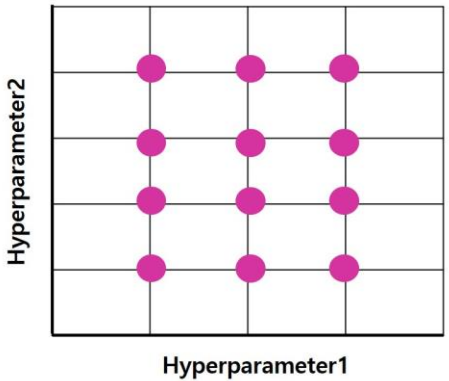
- 검증 세트: 하이퍼파라미터 튜닝을 위해 모델을 평가할 때, 테스트 세트를 사용하지 않기 위해 훈련 세트에서 다시 떼어 낸 데이터 세트.
- 교차 검증: 훈련 세트를 여러 폴드로 나눈 다음 한 폴드가 검증 세트의 역할을 하고 나머지 폴드에서는 모델을 훈련. 교차 검증은 모든 폴드에 대해 검증 점수를 얻어 평균을 냄.



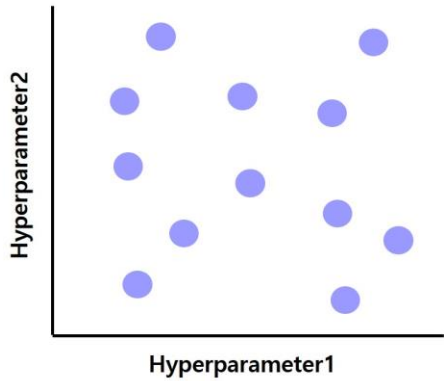
3-2. 교차 검증과 그리드 서치

- 그리드 서치: 하이퍼파라미터 탐색을 자동화해 주는 도구. 탐색할 매개변수를 나열하면 교차 검증을 수행하여 가장 좋은 검증 점수의 매개변수 조합을 선택.
- 랜덤 서치: 연속된 매개변수 값을 탐색할 때 유용. 탐색할 값을 직접 나열하는 것이 아니고 탐색 값을 샘플링할 수 있는 확률 분포 객체를 전달. 지정된 횟수만큼 샘플링하여 교차 검증을 수행하기 때문에 시스템 자원이 허락하는 만큼 탐색량 조절 가능.

<Grid Search>



<Random Search>

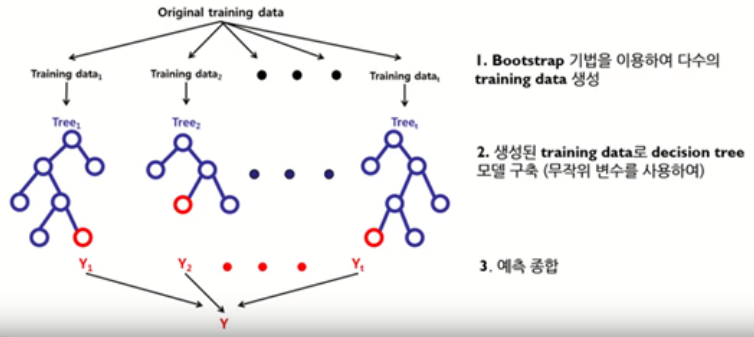


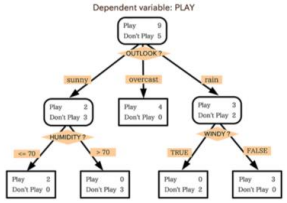
3-3. 트리의 앙상블

- 앙상블 학습: 더 좋은 예측 결과를 만들기 위해 여러 개의 모델을 훈련하는 알고리즘.
- 랜덤 포레스트: 대표적인 결정 트리 기반의 앙상블 학습 방법. 부트스트랩 샘플을 사용하고 랜덤하게 일부 특성을 선택하여 트리를 만듦.

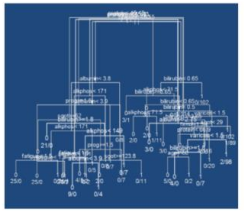
랜덤 포레스트 개요

- 다수의 의사결정나무모델에 의한 예측을 종합하는 앙상블 방법
- 일반적으로 하나의 의사결정나무모델 보다 높은 예측 정확성을 보여줌
- 관측치 수에 비해 변수의 수가 많은 고차원 데이터에서 중요 변수 선택 기법으로 널리 활용됨

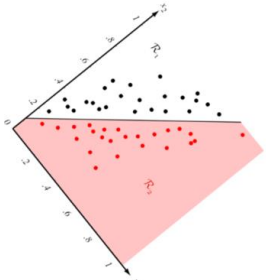




Tree



Random Forest



Rotation Forest

3-3. 트리의 앙상블

- 엑스트라 트리: 랜덤 포레스트와 비슷하게 결정 트리를 사용하여 앙상블 모델을 만들지만 부트스트랩 샘플을 사용하지 않는다. 대신 랜덤하게 노드를 분할해 과대적합을 감소시킨다.
- 그레디언트 부스팅: 랜덤 포레스트나 엑스트라 트리와 달리 결정 트리를 연속적으로 추가하여 손실 함수를 최소화하는 앙상블 방법.

  - **Random Forest vs Extra Trees**
  - 부트스트랩 샘플(중복된 훈련 샘플) 사용 유무 차이. 엑스트라 트리는 결정 트리를 만들어 낼 때 훈련 세트 전체를 사용하기 때문에 Bagging이라고는 할 수 없다.

  - 랜덤포레스트는 주어진 모든 feature에 대한 정보이득을 계산하고 가장 높은 정보 이득을 가지는 feature를 Split Node로 선택하고 그것들은 전부 비교해서 가장 최선의 feature를 선정한다. 이 과정을 통해 성능이 좋은 결정트리를 만들 수 있지만 연산량이 많이 든다는 단점이 있다.

  - 반면에 엑스트라 트리는 Split을 할 때 무작위로 feature를 선정합니다. feature중에 아무거나 고른 다음 그 feature에 대해서 최적의 Node를 분할한다. 성능이 낮아지지만 생각보다 준수한 성능을 보이고 과대적합을 막고 검증 세트의 점수를 높이는 효과가 있다. 그리고 속도가 빠르다는 장점 또한 존재한다.

4. Summary

5. Q&A

Thanks for listening