Introduction to Machine Learning

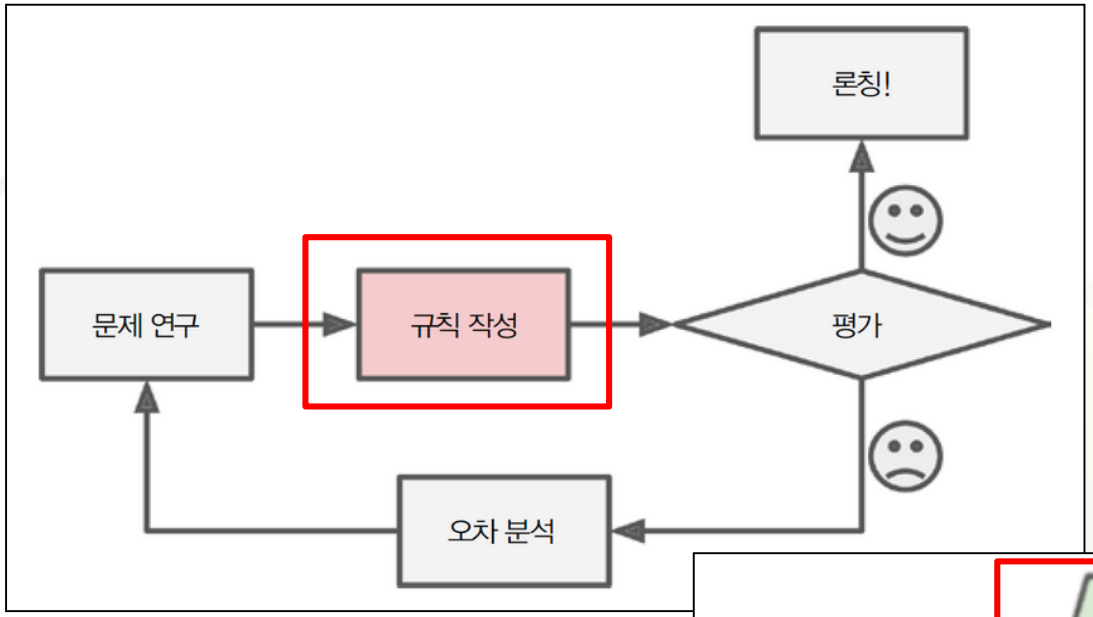
머신 러닝

- 데이터로부터 학습하도록 컴퓨터를 프로그래밍하는 과학 또는 예술

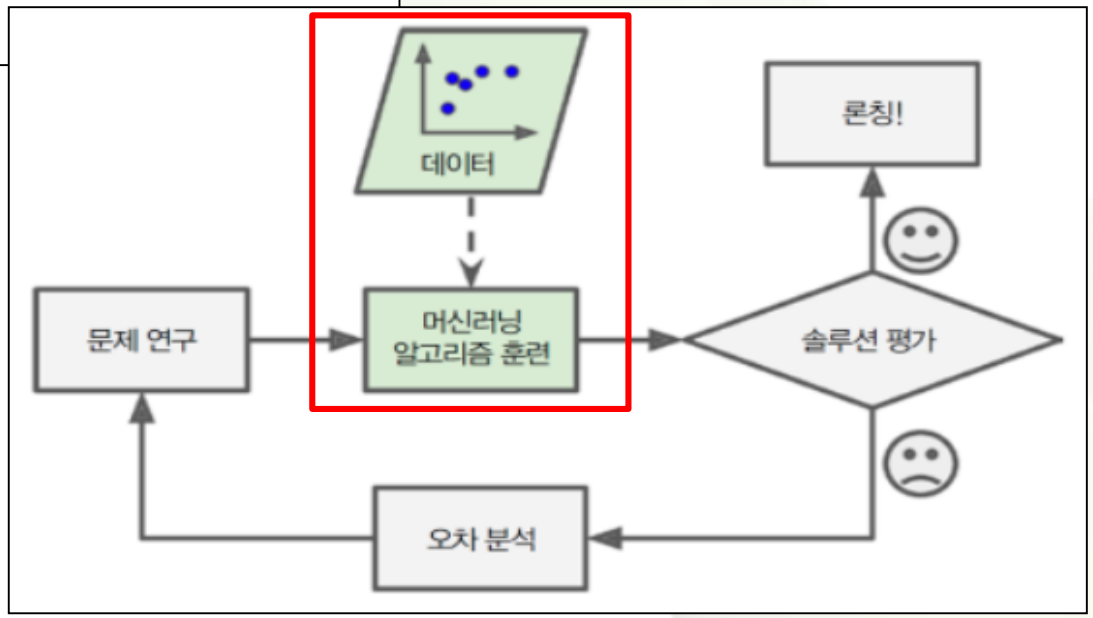
> 명시적인 프로그래밍 없이 컴퓨터가 학습하는 능력을 갖추게 하는 연구 분야.
>
> Arthur Samuel, 1959

> 어떤 작업 T에 대한 컴퓨터 프로그램의 성능을 P로 측정했을 때 경험 E로 인해 성능이 향상되었다면 이 컴퓨터 프로그램은 작업 T와 성능 측정 P에 대해 경험 E로 학습한 것이다.
>
> Tom Mitchell, 1997

일반적인 프로그램과 머신 러닝



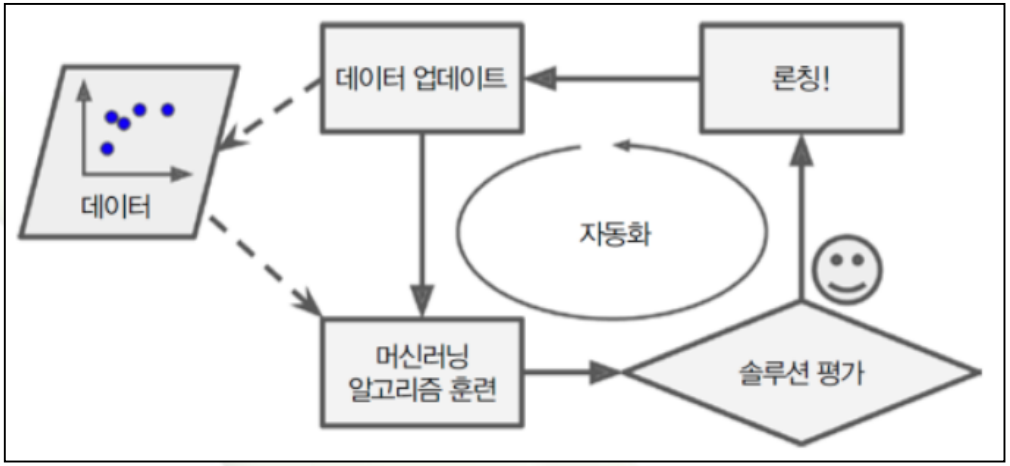
일반적인 프로그램



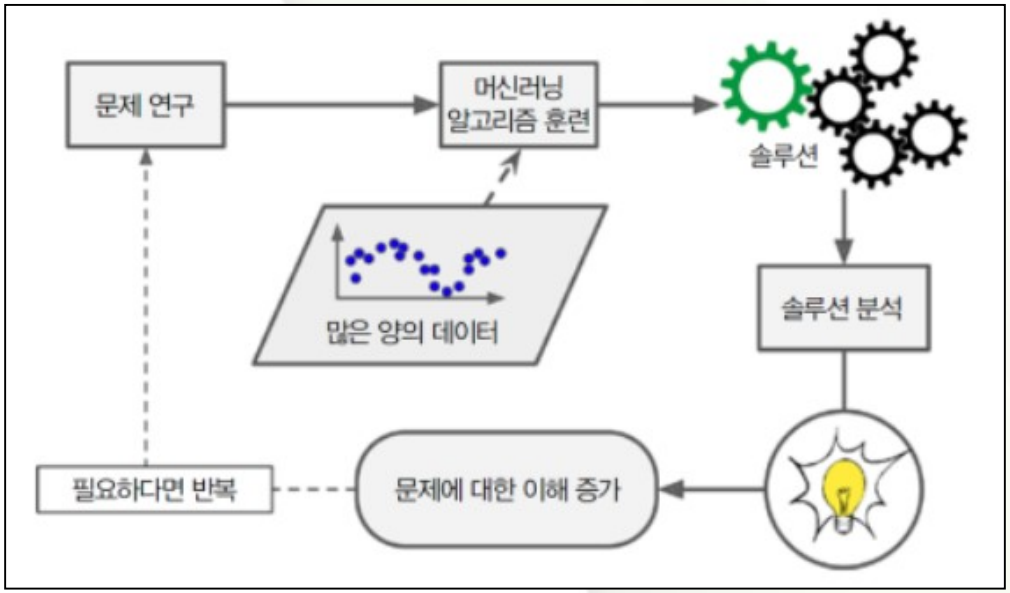
머신러닝 프로그램

머신 러닝을 사용하는 이유

- 규칙에 변경 사항이 발생하는 경우 유지 보수 자동화
- 규칙이 매우 복잡하거나 알려진 알고리즘이 없는 문제도 구현 가능
- 새로운 데이터에 적응 가능



- 머신 러닝을 적용한 분석을 통해 숨겨진 패턴 발견



머신 러닝 시스템 종류

- 학습하는 동안의 감독 형태나 정보량 기준

학습방법	설명
지도학습	▪ 알고리즘에 주입하는 데이터에 레이블이라는 답 포함 • 범주에 대한 분류와 수치를 예측하는 회귀
비지도학습	▪ 알고리즘에 주입하는 데이터에 레이블 없음
강화학습	▪ 환경을 관찰해서 행동을 실행하고 그 결과로 보상 (또는 벌칙) 부여 → 가장 큰 보상을 얻기 위해 [정책]이라는 최상의 전략 학습

머신 러닝 시스템 종류

- 점진적 학습 여부 기준

학습방법	설명
배치학습	▪ 점진적으로 학습할 수 없고 가용한 데이터를 모두 사용 ▪ 시간과 자원 소모량이 많아서 오프라인으로 학습 ▪ 먼저 시스템을 훈련시키고 제품 시스템에 적용하면 더 이상의 학습은 없음 ▪ 새로운 데이터를 적용하려면 전체 데이터로 다시 학습 후 적용
온라인학습	▪ 데이터를 순차적으로 한 개씩 또는 작은 묶음 단위로 주입하여 학습 ▪ 새로운 데이터를 즉시 적용할 수 있음

- 일반화 방법 기준

학습방법	설명
사례 기반 학습	▪ 시스템이 사례를 기억함으로써 학습 → 유사도 측정을 사용해서 새로운 데이터에 일반화
모델 기반 학습	▪ 샘플들의 모델을 만들어 예측에 사용

머신 러닝 알고리즘

- 주요 머신 러닝 알고리즘

학습방법	주요 알고리즘
지도학습	k-Nearest Neighbors, Linear Regression, Logistic Regression, Support Vector Machine (SVM), Decision Tree & Random Forests, Neural Networks
비지도학습	[Clustring] k-Means, Hierachical Cluster Analysis, Expectation Maximization [Visualization & Dimensionality Reduction] Principal Component Analysis (PCA), Kernel PCA, Locally-Linear Embedding, t-Distributed Stochastic Neighbor Embedding (t-SNE) [Association Rule Learning] Apriori, Eclat

머신러닝의 과제

- 적은 양의 훈련 데이터
  - » 대부분의 머신러닝 알고리즘이 잘 동작하려면 충분히 많은 양의 데이터 필요

- 대표성 없는 훈련 데이터
  - » 샘플링 잡음 → 샘플이 작아서 우연히 뽑힌 대표성 없는 데이터 사용
  - » 샘플링 편향 → 잘못된 표본 추출 방법에 의해 대표성 없는 데이터 사용

- 낮은 품질의 데이터
  - » 훈련 데이터에 이상치, 에러가 많이 포함되면 좋은 예측 모델을 만들 수 없음
  - » 많은 데이터 과학자들이 데이터 정제에 상당한 시간 사용

머신러닝의 과제

- 관련 없는 특성
  - » 성공적인 머신 러닝 프로젝트의 핵심 요소는 훈련에 사용할 좋은 특성을 찾는 것 → 이러한 과정이 특성 공학
    - › 특성 선택 : 가지고 있는 특성 중에서 훈련에 유용한 특성 선택
    - › 특성 추출 : 특성을 결합해서 더 유용한 특성 도출

- 훈련 데이터 과대적합 및 과소적합
  - » 과대적합 → 모델이 훈련 데이터에는 잘 맞지만 다른 데이터에는 잘 맞지 않는 경우
  - » 과소적합 → 모델이 지나치게 단순해서 훈련 데이터와 실제 데이터 모두에 잘 맞지 않는 경우