CHAPTER 1. 한눈에 보는 머신러닝

- 머신러닝이 뭐야?
- 머신러닝이 왜 필요해?
- 머신러닝에 뭐가 있는데?
- 머신러닝에서 해결할 과제는?

1.1 머신러닝이 뭐야?

- 머신러닝(Machine Learning, ML)이란 명시적인 규칙을 코딩하지 않고 기계가 데이터로부터 학습해서 어떤 작업을 더 잘 하도록 만드는것을 의미합니다.
- 다른 의미로는 데이터로부터 학습하도록 컴퓨터를 프로그래밍하는 과학 또는 예술을 의미합니다.

1.2 머신러닝이 왜필요해?

- 기존보다 프로그램이 훨씬 짧아지고 유지보수하기 쉬우며 대부분 정확도가 높기 때문입니다.
- 머신러닝을 통해 겉으로 보이지 않는 패턴 등을 확인하며 배울 수 있습니다.
- 머신러닝을 활용하면 좋은 문제 :
  - 기존 솔루션으로는 많은 수동 조정과 규칙이 필요한 문제
  - 전통적인 방식으로는 전혀 해결방법이 없는 복잡한 문제
  - 유동적인 환경
  - 복잡한 문제와 대량의 데이터에서 통찰 얻기

1.3 머신러닝에 뭐가 있는데?

1.3.1 지도학습 / 비지도학습 / 강화학습

: Lable(레이블)의 여부에 따라 나눌 수 있습니다. 지도학습은 Lable이 있고 비지도학습은 Lable이 없습니다.

1) 지도학습 (Supervised Learning)

- 지도학습은 Train데이터에 Lable이라는 y값이 포함됩니다.
- 회귀문제의 경우는 Lable이 수치형, 분류문제의 경우는 Label이 카테고리형입니다.
- 회귀문제 : 키에 대한 몸무게를 알아보는 모델을 만든다고 가정했을 때 키(x)값에 따른 몸무게(y)를 작성합니다.
- 분류문제 : 스팸메일을 찾는 모델을 만든다고 가정했을 때 이 메일이 스팸인지(1) 아닌지(0)에 관한 y값을 작성합니다.

- 지도학습의 대표적인 알고리즘 :

  - K-최근접(K-Nearest Neighbors)
  - 선형회귀(Linear Regression)
  - 로지스틱 회귀(Logistic Regression)
  - 서포트 벡터 머신(Support Vector Machine, SVM)
  - 결정 트리(Decision Tree)
  - 랜덤 포레스트(Random Forest)
  - 신경망 (Neural Networks)

2) 비지도학습 (Unsupervised Learning)

- 비지도학습은 Lable이 없습니다.
- 비지도학습의 대표적인 알고리즘 :

  - 군집(Clustering)
  - 시각화(Visualization) & 차원축소(Dimensionality reduction)
  - 이상치 탐색(anomaly detection)
  - 연관규칙학습(Association rule learning)

3) 강화학습 (Rainforcement Learning)

- 강화학습에서 학습하는 시스템을 에이전트라 부르며 환경을 관찰해서 행동을 실행합니다.
- 그 결과로 보상 또는 별점을 받고 가장 큰 보상을 얻기 위해 정책(Policy)이라 부르는 최상의 전략을 스스로 학습합니다.
- 정책이 주어진 상황에서 에이전트가 어떤 행동을 선택해야 할지 정의합니다.

1.3.2 배치학습 / 온라인학습

: 실시간으로 점진적인 학습을 하는지 아닌지에 따라 나뉩니다. 점진적인 학습을 안하면 배치학습, 하면 온라인학습입니다.

1) 배치학습 (Batch Learning)

- 배치학습은 가용한 데이터를 모두 사용해 훈련합니다.
- 따라서 데이터의 양이 많은 경우엔 시간이 오래걸려 보통 오프라인에서 수행합니다.

2) 온라인학습 (Online Learning)

- 온라인학습은 데이터를 순차적으로 한개씩 또는 미니배치라고 부르는 작은 묶음 단위로 주입하여 훈련합니다.
- 바로 변화에 스스로 적응해야 할 때 유리합니다. ex) 주식가격
- 온라인 학습에선 Learning Rate가 중요합니다.
- L.R이 크면 시스템이 데이터에 빠르게 적응할 수 있지만 데이터를 금방 잊습니다. 반대로 L.R이 낮으면 시스템이 관성이 커져서 느리게 학습하게 됩니다.

1.3.3 사례기반학습 / 모델기반학습

: 머신러닝 시스템을 어떻게 일반화 시키는지에 따라 나뉩니다.

1) 사례기반학습 (Instance-based Learning)

- 스펨메일을 예로 들면 스펨메일과 공통으로 가지고 있는 단어가 많으면 스팸으로 분류하는 것을 의미합니다.
- 공통으로 가지고 있는 단어 개수를 파악해 스펨인지 아닌지 확인합니다.

2) 모델기반학습 (Model-based Learning)

- 샘플들의 모델을 만들어 예측에 사용하는 것을 말합니다.

1.4 머신러닝에서 해결할 과제는?

1.4.1 과대적합 (Overfitting)

- Overfitting은 모델이 Train데이터에는 너무 잘맞지만 일반성이 떨어지는 것을 의미합니다.
- Overfitting 해결 방법 :
  - 파라미터 수가 적은 모델을 선택합니다.
  - Train 데이터에 있는 특성 수를 줄입니다.
  - 모델에 제약을 가하여 단순화시킵니다.
  - Train 데이터를 더 많이 모읍니다.
  - Train 데이터의 잡음을 줄입니다.(예를 들면 오류 데이터 수정과 이상치 제거)

1.4.2 과소적합 (Underfitting)

- Underfitting은 모델이 너무 단순해서 데이터의 내재된 구조를 학습하지 못하는 경우입니다.
- Underfitting 해결 방법 :
  - 파라미터가 더 많은 강력한 모델을 선택합니다.
  - 학습 알고리즘에 더 좋은 특성을 제공합니다.(특성 엔지니어링)
  - 모델의 제약을 줄입니다.(예를 들면 규제 하이퍼파라미터를 감소시킵니다.)