Hands-On Machine Learning

Chapter 1. 한눈에 보는 머신러닝

* 머신러닝 이란?

머신러닝은 해결하고자 하는 **문제**에 대하여, 컴퓨터에게 훈련 데이터를 **학습**하여, 문제에 대한 해결 및 정확도(Accuracy)를 상승하는 것을 의미한다. 예를 들면, 무수히 강아지의 사진을 훈련 데이터를 주어 학습을 진행하여 새로운 사진을 보고 강아지인지 아닌지 컴퓨터가 판단할 수 있게 하는 것이다.

머신러닝의 장점과 사용 이유는 무엇일까? 사람이 하지 못하는 일에 대해서 생각하는 것이 간단하다.

1. 오늘 날 데이터의 수는 폭발적으로 증가하였으며, 이에 따라 인간이 처리할 수 있는 시간과 능력이 불가능하다. 즉 컴퓨터로 훨씬 가격이 **저렴하고 신속**하게 진행할 수 있다.

2. 데이터는 가변적이다. 스팸 메일의 종류와 형태가 바뀌듯, 규칙이 변화하고 일일이 실시간으로 파악할 수 없다. 이를 **자동적**으로 해줄 수 있다.

3. 인간보다 **정확**하다. 항상 올바른 명제는 아니다. 인간에게 쉬운 동물 사진 맞히기를 컴퓨터는 잘 해내지 못할 수 있다. 하지만 알파고가 그렇듯 컴퓨터는 인간을 능가하였으며, 점점 발전하여 더욱 더 특정 일이나 문제에 정확성을 가진다.

이는 어디까지나 어떤 문제를 다루냐에 따라 달라진다. 하지만 점점 더 많은 문제와 상황에서 앞서 말한 부분들이 증가하고 있는 것은 사실이다.

* 머신러닝 시스템의 종류

머신러닝의 종류는 지도 비지도(준지도) 강화학습이 존재한다. 각 파트마다 내용이 매우 방대하고 쓰임의 상황과 목적이 다르다. 이에 대해서 간단히 파악하려 한다.

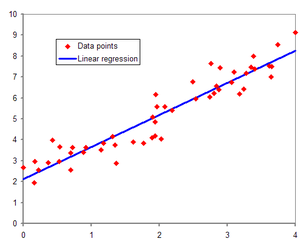
- 지도 학습(supervised learning) : 우리는 타이타닉 사건에 대한 데이터를 가지고 있다. 탑승객들의 나이와 객실 등급, 성별과 같은 **특징(feature)**과 이 승객의 생존여부(목표, **label**)이 있는 데이터이다. 특징들을 통해서 label을 얼마나 잘 맞치게 학습하는 것, 즉 데이터와 학습의 목표로 하는 값이 존재하는 것을 지도학습이라고 한다.

- 비지도 학습(unsupervised learning) : 비지도 학습은 앞서 타이타닉 문제에서 승객들의 생존여부가 없는 데이터이다. 이에 대해서 특징들을 통해서, 비슷한 군집을 찾는 것을 목표로 할 수 있다. 즉 비지도 학습의 가장 큰 특징은 목표값(label)이 존재하지 않는 것이라고 할 수 있다.

- 강화학습(reinforce learning) : 강화학습의 가장 유명한 예시는 물론 알파고이다. 강화학습은 학습하는 시스템을 에이전트라고 부르며, 행동에 대한 결과에 보상(reward)과 벌점(penalty)를 부여합니다. 이를 무수히 학습하여 가장 보상이 높아지는 정책policy을 찾는 것입니다.

* 머신러닝관련 문제들

머신러닝은 다양한 종류를 가지고 있지만, 직관적으로 파악하기 위해서는



간단한 예로 위의 그림과 같이 특징에 대한 값을 설명하는 **y = ax + b**와 같은 선을 가진다. 실제 값을 가장 잘 잘 맞추기 위해서는 회귀분석과 비선형 회귀분석과 같이 **알고리즘 선택**을 하여야 한다. 또한 판단하기 위해서는 a의 기울기와 b의 절편, **parameter**를 적절히 찾아야 하며, 이를 위해 **비용함수(cost function)** 혹은 **효용함수(utility function)**를 적절히 설정시킬 필요가 있다.

이를 위해서는

1. 충분하지 않은 양의 훈련 데이터 - 알고리즘이 충분히 훈련하기 위해서는 많은 데이터가 필요하다. 특히 복잡하고 정교한 알고리즘의 경우 데이터가 적을 경우 쉽게 overfitting이 발생함으로, 데이터는 가능한 큰 것이 모집단의 특징을 많이 반영한다.

2. 대표성 없는 훈련 데이터 - 서울의 집값을 예측하는 프로젝트에서 훈련데이터로 강원도에 있는 산이 포함된다면 이 훈련데이터는 서울의 집값을 대표하지 못하며, 이에 따라 유의미한 학습이 불가능하다. 이에 대해서 샘플링 잡음(sampling noise)은 데이터에 오타나 오류가 들어간 것을 의미고 샘플링 편향(sampling bias)은 데이터가 대표성이 없고 한쪽으로 치우쳐진 것으로 이해할 수 있다.

3. 낮은 품질의 데이터 - 데이터에 에러, 이상치(outlier), 잡음, na값이 있으면 학습하기 어렵다. 이에 대해서 소거하거나 값을 채워주는(보간법)이 필요하다.

4. 관련 없는 특성 - 목표에 적절한 특징을 가질 필요가 있다. 나이를 예측하는데 키와 몸무게 등은 유의미한 특징이 될 수 있지만, 가족원의 수와 같이 무의미한 특징이 있을 수 있다. 이에 대해서도 특성 선택(feature selection)은 무수히 많은 특징 중 유용한 특징을 파악 및 선택하는 것이며, 특징 추출(feature extraction)은 특성을 결합하여 새로운 유용한 특성을 만드는 것이다.

5. 훈련 데이터 과대적합 - 우리의 목표는 훈련데이터를 통하여, 실제 테스트 데이터를 유의미한게 맞추는 것을 목표로 한다. 하지만 훈련에 과대적합하여 훈련은 잘 맞추지만, 테스트에서는 정확도가 낮은 경우가 있다. 이에 대해서는 i) 특성의 수 감소 혹은 모델 변경, ii) 규제 및 데이터 정제, iii) 훈련 데이터의 크기를 늘리기 와 같은 방법이 존재한다.

6. 훈련 데이터 과소적합 - 앞서 과대적합과 반대로 훈련이 부족하게 된 경우도 존재한다.