I. Machine learning & Type of the algorithms

*Machine learning

: A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E

⇨ 경험을 통해 성능을 향상시키는 것!

*Supervised / unsupervised study

- 지도 학습 / 비지도 학습 => 데이터의 종류에 따라 달라짐. 정답이 주어진 경우는 지도 학습이지만, 정답에 대해서 알지 못하는 상태에서 특징만 알고 있는 데이터의 경우에는 비지도 학습 (머신러닝에서 차지하는 비율이 높음)
- 지도 학습: linear regression (real target values 연속된 값을 예측) / classification (discrete target values)
- 비지도 학습: clustering algorithm (비슷한 것들을 묶어놓는 것 – no target values)

  ex) 구글 주요 뉴스 ; 비슷한 기사들 끼리 묶어주는 것!

  소셜 네트워크 분석 (span account), 시장 분할 (데이터의 숨겨져 있는 패턴이 무엇인지 모르고 정답이 주어지지 않았을 때 이를 사용하는 것!)

*Characteristics of algorithms

- supervised classification: 보지 않은 (unseen) 데이터에 대한 정확도 (accuracy)

  ex) 주택 가격에 대한 데이터 10 만건 – 8 만건 가지고 학습을 한 다음 나머지 2 만 건에 대해서 학습하지 않은 데이터를 가지고 얼마나 예측을 잘 했느냐를 보고 정확도를 판단
- unsupervised classification: 군집화 된 데이터에 대한 검증
  - internal evaluation: 같은 클러스터에 있는 데이터의 유사도는 높고, 다른 클러스터에 있는 데이터의 유사도는 낮은지에 대한 검증 – 내부적으로 검증하는 방법. 이미 클러스터를 한 뒤에 다시 이 클러스터에 있는 데이터끼리 유사한 지에 대해 이야기 하는 것이기 때문에 circular way
  - external evaluation: 골드 스탠다드(정답)이라고 부를 수 있는 외부의 라벨을 이용하여 클러스터를 검증 (like cheating) – But 사람들의 bias 및 사람들의 주관이 들어갈 수 밖에 없음.

  ⇨ so, 결국 여러가지를 섞어서 사용.

*Type of data

- Training data: 모델 학습에 사용
- Validation data: 학습을 할 때 튜닝이 필요한 parameter 들을 최적화하기 위해 사용 (예를 들어, 학습 모델에서 K 라는 클러스터 갯수를 10 개로 할 것인가, 20 개로 할 것인가를 결정하는 등의 문제!)
- Test data: 학습된 모델을 검증하기 위해 사용 (결국 학습된 모델을 test 하는 게 가장 중요한 문제!)

*Overfitting

- 예측할 때 sensor 가 조금씩 오류가 있던지, 데이터 자체가 형성되는 과정에서 오류가 발생했던지, 데이터가 우리 모델에 비해 복잡도가 높을 때 발생!

*표기법

m = 훈련용 데이터 세트의 개수

x = 입력 변수 / 특징

y = 출력 변수 / 특징

(x,y) : 한 훈련 데이터

(x^{i}, y^{i}) : i 번째 훈련 데이터

*가설을 어떻게 묘사할까

- 단일 변수 선형 회귀 1: $h_\theta(x) = \theta_0 + \theta_1 x$
- 이 데이터를 가장 잘 표현하는 hypothesis 는 무엇일까를 이야기하게 됨!
- 실제 y 와 h 의 차이의 제곱이 최소화되는 값을 찾으려는 것!

II. current issues

- Unsupervised study 에 대부분 초점이 맞춰져 있는 듯 함. (pdf 참조)

III. goal of the study

- Bayesian methodology, Clustering 방법에 대해서 배울 거!