chapter1

Chpater 1

머신러닝 스팸필터 예시

- 머신러닝 사용 시 전통적인 방법보다 유지보수 쉬움
- 새로운 패턴/단어가 생겼을 때 자동으로 인식하고 별도의 작업 필요 없음
- 알고리즘이 학습한 것을 확인할 수 있음 : 연관관계 및 새로운 추세 등 → 보이지 않던 패턴 발견 (= **데이터마이닝**)

머신러닝 시스템의 종류

- 지도 / 비지도
  - 지도 : 훈련 데이터에 레이블(=class)라고 하는 답이 포함됨. (x, y 모두 제공)
    - 분류 : ex. 스팸필터
    - 회귀 : 특성(= 예측 변수) 를 통해 타깃 수치를 예측

    ** 분류 ↔ 회귀 ⇒ 로지스틱 회귀

    - K-최근접 이웃 / 선형회귀 / 로지스틱 회귀 / SVM / 결정트리&랜덤포레스트 / 신경망
  - 비지도 : 훈련데이터에 레이블(=class) 정보가 없음
    - 군집 : k-means, DBSCAN, 계층 군집 분석, 이상치탐지&특이치 탐지, 원-클래스, 아이솔레이션 포레스트
    - 시각화와 차원 축소 : PCA, 커널 PCA, 지역 선형 임베딩, t-SNE
    - 연관 규칙 학습 : 어프라이어리, 이클렛
  - 준지도 : 훈련 데이터 중 일부만 레이블이 있음 / 일반적으로 지도와 비지도의 조합.
  - 강화 학습 : 관찰과 실행의 결과로 보상 혹은 벌점을 받음 → 이러한 과정에서 가장 큰 보상을 받도록 학습되는 것이 강화학습.

- 온라인 / 배치
  - 온라인
    - 빠른 학습 단계 / 저비용
    - 연속적으로 데이터 받고 빠른 변화에 적응할 수 있음.
    - 큰 데이터 → 일부 데이터를 읽어들이며 학습 가능
    - 학습률 중요
    - 면밀히 모니터링 하는 것이 중요함.
  - 배치
    - 점진적 학습 불가
    - 가용 데이터 모두 사용해야 함.
    - 새로운 데이터에 대해 학습하려면 이전에 사용한 전체 데이터도 사용해서 새롭게 학습.
- 사례기반 / 모델기반
  - 사례 기반
    - 유사도 측정을 통해 학습 데이터와 새로운 데이터를 비교함
  - 모델 기반
    - 샘플들의 모델을 만들어 예측에 사용함.

나쁜 데이터의 사례

- 충분하지 않은 양의 훈련 데이터
- 대표성이 없는 훈련 데이터
- 낮은 품질의 데이터 (에러, 이상치, 결측 등)
- 관련 없는 특성
  - 훈련 시 좋은 특성을 찾는 방법 ⇒ 특성 공학 (특성 선택 / 특성 추출)

나쁜 알고리즘의 사례

- 훈련 데이터 과대적합

    - 과대적합 → 모델이 너무 복잡할 때!
    - 해결 방법 : 파라미터 수가 적은 모델 선택, 특성 수 줄이기, 모델에 제약(규제), 훈련 데이터 모으기, 훈련데이터 정제(잡음 줄이기)
    - 해결 방법 중 규제의 양은 → 하이퍼파라미터가 결정하는데, 이는 학습 알고리즘의 파라미터이다.

- 훈련 데이터 과소적합
    - 과대적합의 반대 케이스

테스트와 검증

- 데이터를 훈련세트와 테스트 세트로 나눈다.
- 테스트 세트를 사용하여 모델의 성능을 평가할 수 있음. 일반적으로 테스트셋은 20%

하이퍼파라미터 튜닝과 모델 선택

- 테스트 세트에 최적화된 모델을 만들면 새로운 데이터에 대해 일반화가 어려울 수 있음
- 이를 위해 hold out 검증 → 훈련셋중 일부를 떼어내고 여러 후보 모델을 평가하여 가장 좋은 하나 선택
- 훈련세트에서 다양한 하이퍼파라미터 값을 가진 여러 모델 훈련 → validation set에서 가장 성능이 높은 것 택 → train+val set에서 다시 훈련하여 최종모델 만들기 → testset을 활용하여 최종모델 평가.