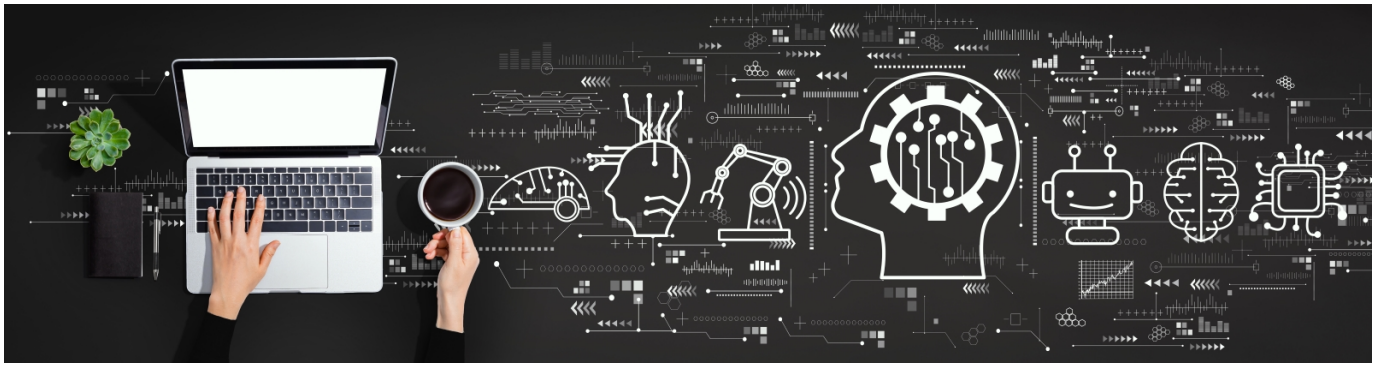


머신러닝 (Machine Learning)

- 명시적인 프로그래밍 없이 컴퓨터가 학습하는 능력을 갖추게 하는 연구 분야
- 머신러닝은 데이터를 통해 다양한 패턴을 감지하고, 스스로 학습할 수 있는 모델 개발에 초점

+ 코드 + 텍스트

머신러닝 분류

지도 학습(Supervised Learning)

- 지도 학습은 주어진 입력으로 부터 출력 값을 예측하고자 할 때 사용
- 입력과 정답 데이터를 사용해 모델을 학습 시킨 후 새로운 입력 데이터에 대해 정확한 출력을 예측하도록 하는 것이 목표
- 지도 학습 알고리즘의 학습 데이터를 만드는 것은 많은 사람들의 노력과 자원이 필요하지만 높은 성능을 기대할 수 있음

분류와 회귀

- 지도 학습 알고리즘은 크게 **분류(classification)**와 **회귀(regression)**로 구분
- 분류는 입력 데이터를 미리 정의된 여러개의 클래스 중 하나로 예측하는 것
- 분류는 클래스의 개수가 2개인 이진 분류(binary classification)와 3개 이상인 다중 분류(multi-class classification)로 나눌 수 있음
- 회귀는 연속적인 숫자를 예측하는 것으로 어떤 사람의 나이, 농작물의 수확량, 주식 가격 등 출력 값이 연속성을 갖는 다면 회귀 문제라고 할 수 있음

지도 학습 알고리즘

- 선형 회귀(Linear Regression)
- 로지스틱 회귀(Logistic Regression)
- 서포트 벡터 머신(Support Vector Machine)
- k-최근접 이웃(k-Nearest Neighbors)

- 결정 트리(Decision Tree)
- 앙상블(Ensemble)
- 신경망(Neural Networks)

비지도 학습(Unsupervised Learning)

- 비지도 학습은 원하는 출력 없이 입력 데이터를 사용
- 입력 데이터의 구조나 패턴을 찾는 것이 목표
- 미리 정해진 결과가 없고, 방대한 양의 데이터에서 유용한 통찰력을 얻을 수 있음

클러스터링, 차원 축소, 연관규칙

- 비지도 학습 알고리즘은 크게 **클러스터링(Clustering)**, **차원 축소(Dimensionality Reduction)**, **연관 규칙(Association Rules)**으로 구분
- 클러스터링은 공간상에서 서로 가깝고 유사한 데이터를 클러스터로 그룹화
- 차원 축소는 고차원의 데이터에 대해서 너무 많은 정보를 잃지 않으면서 데이터를 축소시키는 방법
- 연관 규칙은 데이터에서 특성 간의 연관성이 있는 흥미로운 규칙을 찾는 방법

비지도 학습 알고리즘

- 클러스터링(Clustering)
  - k-Means
  - DBSCAN
  - 계층 군집 분석(Hierarchical Cluster Analysis)
  - 이상치 탐지(Outlier Detection), 특이값 탐지(Novelty Detection)
- 차원축소(Dimensionality Reduction)
  - 주성분 분석(Principal Component Analysis)
  - 커널 PCA(Kernel PCA)
  - t-SNE(t-Distributed Stochastic Neighbor Embedding)
- 연관 규칙(Association Rule Learning)
  - Apriori
  - Eclat

준지도 학습(Semi-supervised Learning)

- 레이블이 있는 것과 없는 것이 혼합된 경우 사용
- 일반적으로는 일부 데이터에만 레이블이 있음
- 준지도 학습 알고리즘은 대부분 지도 학습 알고리즘과 비지도 학습 알고리즘의 조합으로 구성

강화 학습(Reinforcement Learning)

- 동적 환경과 함께 상호 작용하는 피드백 기반 학습 방법
- 에이전트(Agent)가 환경을 관찰하고, 행동을 실행하고, 보상(reward) 또는 벌점(penality)를 받음
- 에이전트는 이러한 피드백을 통해 자동으로 학습하고 성능을 향상시킴
- 어떤 지도가 없이 일정한 목표를 수행

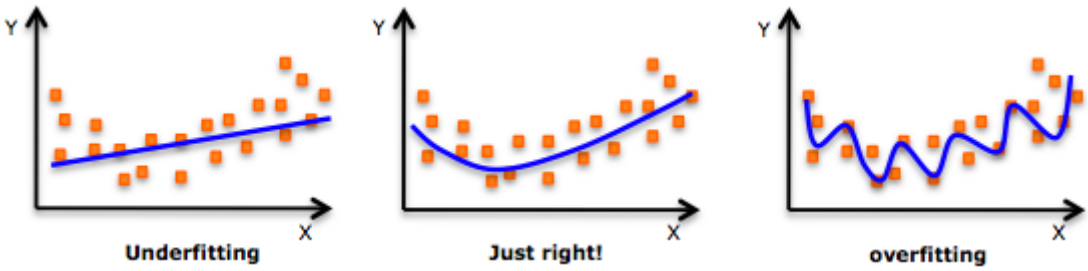
온라인 vs. 배치

- 온라인 학습(Online Learning)
    - 적은 데이터를 사용해 미니배치(mini-batch) 단위로 점진적으로 학습
    - 실시간 시스템이나 메모리 부족의 경우 사용
- 배치 학습(Batch Learning)
    - 전체 데이터를 모두 사용해 오프라인에서 학습
    - 컴퓨팅 자원이 풍부한 경우 사용

사례 기반 vs. 모델 기반

- 사례 기반 학습(Instance-based Learning)
    - 훈련 데이터를 학습을 통해 기억
    - 예측을 위해 데이터 사이의 유사도 측정
    - 새로운 데이터와 학습된 데이터를 비교
- 모델 기반 학습(Model-based Learning)
    - 훈련 데이터를 사용해 모델을 훈련
    - 훈련된 모델을 사용해 새로운 데이터를 예측

일반화, 과대적합, 과소적합



일반화(generalization)

- 일반적으로 지도 학습 모델은 학습 데이터로 훈련 시킨 뒤 평가 데이터에서도 정확하게 예측하기를 기대함
- 훈련된 모델이 처음보는 데이터에 대해 정확하게 예측한다면, 이러한 상태를 모델이 **일반화(generalization)** 되었다고 함
- 모델이 항상 일반화 되는 것은 아님

과대적합(overfitting)

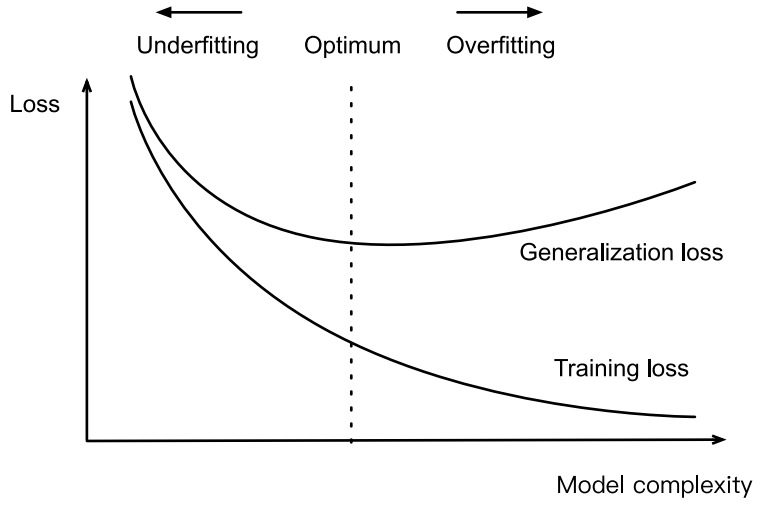
- 주어진 훈련 데이터에 비해 복잡한 모델을 사용한다면, 모델은 훈련 데이터에서만 정확한 성능을 보이고, 평가 데이터에서는 낮은 성능을 보임
- 즉, 모델이 주어진 훈련 데이터는 잘 예측하지만 일반적인 특징을 학습하지 못해 평가 데이터에서는 낮은 성능을 보이는 상태를 **과대적합(overfitting)**이라고 함

과소적합(underfitting)

- 과대적합과 반대로 주어진 훈련 데이터에 비해 너무 간단한 모델을 사용하면, 모델이 데이터에 존재하는 다양한 정보들을 제대로 학습하지 못함
- 이러한 경우 모델은 훈련 데이터에서도 나쁜 성능을 보이고 평가 데이터에서도 낮은 성능을 보이는 **과소적합(underfitting)**되었다고 함

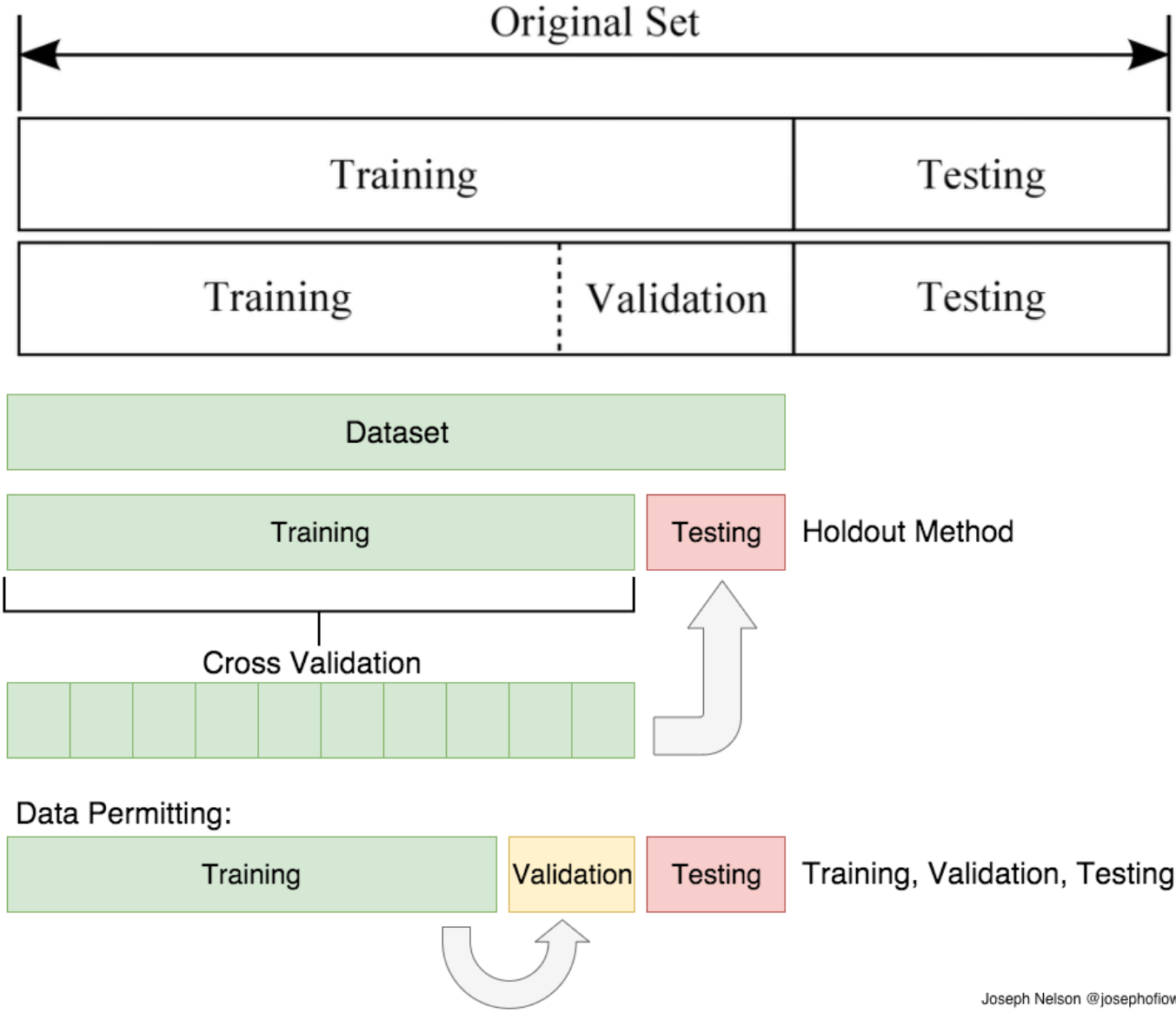
모델 복잡도와 데이터셋 크기의 관계

- 데이터의 다양성이 클수록 더 복잡한 모델을 사용하면 좋은 성능을 얻을 수 있음
- 일반적으로 더 큰 데이터셋(데이터 수, 특징 수)일수록 다양성이 높기 때문에 더 복잡한 모델을 사용할 수 있음
- 하지만, 같은 데이터를 중복하거나 비슷한 데이터를 모으는 것은 다양성 증가에 도움이 되지 않음
- 데이터를 더 많이 수집하고 적절한 모델을 만들어 사용하면 지도 학습을 사용해 놀라운 결과를 얻을 수 있음



훈련 세트 vs. 테스트 세트 vs. 검증 세트

- 머신러닝 모델의 일반화 성능을 측정하기 위해 훈련 세트, 테스트 세트로 구분
- 훈련 세트로 모델을 학습하고 테스트 세트로 모델의 일반화 성능 측정
- 하이퍼파라미터는 알고리즘을 조절하기 위해 사전에 정의하는 파라미터
- 테스트 세트를 이용해 여러 모델을 평가하면 테스트 세트에 과대적합됨
- 모델 선택을 위해 훈련 세트, 테스트 세트, 검증 세트로 구분



참고문헌

- scikit-learn 사이트: https://scikit-learn.org/
- Jake VanderPlas, "Python Data Science Handbook", O'Reilly
- Aurelien Geron, "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems", O'Reilly