2022_02_25 김경준

Supervised / Unsupervised / Semisupervised / Reinforcement

Supervised Learning

- K-Nearest Neighbors (K-최근접 이웃)
- Linear Regression (선형 회귀)
- Logistic Regression (로지스틱 회귀)
- Support Vector Machine (서포트 벡터 머신)
- Decision Tree, Random Forests (결정 트리, 랜덤 포레스트)
- Neural networks (신경망)

Unsupervised Learning

- Clustering (군집)
  - k-Means (k-평균)
  - Hierarchical Cluster Analysis (계층 군집 분석)
  - Expectation Maximization (기댓값 최대화)
- Visualization, Dimensionality reduction (시각화, 차원 축소)
  - Principal Component Analysis (주성분 분석)
  - Kernel PCA (커널 PCA)
  - Locally-Linear Embedding (지역적 선형 임베딩)
  - t-distributed Stochastic Neighbor Embedding (t-SNE)
- Association rule learning (연관 규칙 학습)
  - Apriori (어프라이어리)
  - Eclat (이클렛)
- Anomaly detection (이상치 탐지)

Semisupervised Learning

- Restricted Boltzmann machine (unsupervised) -> deep belief network (DBN) (supervised)

Reinforcement Learning

- Agent: scan environment -> action -> reward/penalty -> makes policy

Batch Learning / Online Learning

Batch Learning (Offline Learning)

- 점진적으로 학습 x, 가용 데이터 모두 사용하여 훈련.
- 새로운 데이터 학습? 이전 시스템 중지, 새 시스템으로 다시 훈련.

Online Learning (mini-batch / incremental learning)

- 작은 묶음 단위로 훈련. -> 매 학습 단계가 빠르고 비용 적게 듦.
- Learning rate (학습률)을 높게 하면 데이터에 빠르게 적응, 예전 데이터 금방 잊음.

학습률을 낮게 하면 느리게 학습, 하지만 새로운 데이터에 있는 잡음이나 대표성 없는 데이터 포인트에 덜 민감해짐.

---

1. 머신러닝을 어떻게 정의할 수 있나요?
   A. 데이터 세트의 기계적 학습을 통한 인사이트 도출 작업
   B. **머신러닝**은 데이터로부터 학습할 수 있는 시스템을 만드는 것. **학습**이란 어떤 작업에서 주어진 성능 지표가 더 나아지는 것.
2. 머신러닝이 도움을 줄 수 있는 문제 유형 네가지를 말해보세요.
   A. 주가 예측, 레이블링을 통한 분류 작업, 새로운 상관관계 도출할 수 있는 군집 도출, 이상치 탐지
   B. 명확한 해결책이 없는 복잡한 문제, 수작업으로 만든 긴 규칙 리스트를 대체하는 경우, 변화하는 환경에 적응하는 시스템을 만드는 경우, 사람에게 통찰을 제공해야 하는 경우
3. 레이블된 훈련 세트란 무엇인가요?
   A. 지도 학습 중 하나로, 이미 레이블된 군집에 데이터를 적용하여 분류하는 작업
   B. 레이블된 훈련 세트는 각 샘플에 대해 원하는 정답(레이블)을 담고 있는 훈련 세트
4. 가장 널리 사용되는 지도 학습 작업 두 가지는 무엇인가요?
   A. Linear Regression, Logistic Regression
   B. 회귀(regression)와 분류(classification)
5. 보편적인 비지도 학습 작업 네가지는 무엇인가요?
   A. Clustering, Visualization, Dimensionality reduction, Association rule learning
   B. Clustering, Visualization, Dimensionality reduction, Association rule learning
6. 사전 정보가 없는 여러 지형에서 로봇을 걸어가게 하려면 어떤 종류의 머신러닝 알고리즘을 사용할 수 있나요?
   A. Reinforcement Learning
   B. Reinforcement Learning
7. 고객을 여러 그룹으로 분할하려면 어떤 알고리즘을 사용해야 하나요?
   A. k-means
   B. 그룹을 어떻게 정의할 지 모른다면 비지도 학습(군집 알고리즘), 안다면 지도학습(분류)
8. 스팸 감지의 문제는 지도 학습과 비지도 학습 중 어떤 문제로 볼 수 있나요?
   A. 지도 학습
   B. 지도 학습

9. 온라인 학습 시스템이 무엇인가요?

   A. 데이터 세트를 작은 묶음 단위로 적용하여 학습시키는 방법, 빠르고 비용(메모리, CPU)이 적게 듦, 빠른 변화에 적응해야 하는 시스템에 적합. (주가 예측)

   B. 온라인 학습 시스템은 배치 학습 시스템과 달리 점진적으로 학습할 수 있음. 변화하는 데이터와 자율 시스템에 빠르게 적응하고 매우 많은 양의 데이터를 훈련시킬 수 있음.

10. 외부 메모리 학습이 무엇인가요?

    A. –

    B. 컴퓨터의 주메모리에 들어갈 수 없는 대용량의 데이터를 다룰 수 있음. 데이터를 미니배치로 나누고 온라인 학습 기법을 사용해 학습함.

11. 예측을 하기 위해 유사도 측정에 의존하는 학습 알고리즘은 무엇인가요?

    A. Instance-based learning

    B. Instance-based learning, 새로운 샘플이 주어지면 유사도 측정을 사용해 학습된 샘플 중에서 가장 비슷한 것을 찾아 예측으로 사용함.

12. 모델 파라미터와 학습 알고리즘의 하이퍼파라미터 사이에는 어떤 차이가 있나요?

    A. –

    B. **모델**: 하나 이상의 파라미터(ex. 선형 모델의 기울기)를 사용해 새로운 샘플이 주어지면 무엇을 예측할지 결정. **학습 알고리즘**: 모델이 새로운 샘플에 잘 일반화되도록 이런 파라미터들의 최적 값을 찾음. **하이퍼파라미터**: 모델이 아니라 이런 학습 알고리즘 자체의 파라미터(ex. 적용할 규제의 정도)

13. 모델 기반 알고리즘이 찾는 것은 무엇인가요? 성공을 위해 이 알고리즘이 사용하는 가장 일반적인 전략은 무엇인가요? 예측은 어떻게 만드나요?

    A. 어떤 데이터에서 다중 데이터를 수집하여 그 데이터에 부합하는 Regression을 형성함. 자유도(df)에 따라 평균값을 가진 기울기를 구하고, 그 선에 부합하는 데이터에 대한 예측을 함.

    B. 새로운 샘플에 잘 일반화되기 위한 모델 파라미터의 최적 값을 찾음. 훈련 데이터에서 시스템의 예측이 얼마나 나쁜지 측정하고 모델에 규제가 있다면 모델 복잡도에 대한 패널티를 더한 비용 함수를 최소화함으로써 시스템을 훈련. 예측을 만들려면 학습 알고리즘이 찾은 파라미터를 사용하는 모델의 예측 함수에 새로운 샘플의 특성을 주입.

14. 머신러닝의 주요 도전 과제는 무엇인가요?

    A. 과대적합/과소적합이 없는 정확한 학습으로 인한 결과값 도출

    B. 부족한 데이터, 낮은 데이터 품질, 대표성 없는 데이터, 무의미한 특성, 훈련 데이터에 과소적합된 과도하게 간단한 모델, 훈련 데이터에 과대적합된 과도하게 복잡한 모델

15. 모델이 훈련 데이터에서의 성능은 좋지만 새로운 샘플에서의 일반화 성능이 나쁘다면 어떤 문제가 있는 건가요? 가능한 해결책 세 가지는 무엇인가요?

    A. –

B. 과대적합. 과대적합에 대한 해결책은 더 많은 데이터를 모으거나, 모델을 단순화하거나(간단한 알고리즘 선택, 특성이나 파라미터의 수 줄이기, 모델에 규제 추가하기), 훈련 데이터에 있는 잡음을 감소시키는 것.

16. 테스트 세트가 무엇이고 왜 사용해야 하나요?

A. 같은 데이터 세트에서 20%를 추출해 내어 훈련세트에 적용된 모델이 정상적으로 작동하는지 검증하기 위해 테스트 세트를 만들어야 함

B. 실전에 배치되기 전에 모델이 새로운 샘플에 대해 만들 일반화 오차를 추정하기 위해 사용.

17. 검증 세트의 목적은 무엇인가요?

A. –

B. 모델을 비교하는 데 사용. 이를 사용해 가장 좋은 모델을 고르고 하이퍼파라미터를 튜닝.

18. 테스트 세트를 사용해 하이퍼파라미터를 튜닝하면 어떤 문제가 생기나요?

A. –

B. 테스트 세트에 과대적합될 위험, 일반화 오차는 매우 낙관적으로 측정.

19. 교차 검증이 무엇이고, 왜 하나의 검증 세트보다 선호하나요?

A. –

B. **교차 검증:** 검증 세트를 별도로 분리하지 않고(모델 선책과 하이퍼파라미터 튜닝을 위해) 모델을 비교할 수 있는 기술. 훈련 데이터를 최대한 활용하도록 도와줌.