


ML_01. Intro

🕒 생성일	@2022년 6월 6일 오전 9:45
📁 유형	머신러닝/딥러닝
👤 작성자	 동훈 오

해당 내용은 'HandsOn 2판 - chapter 1. 한눈에 보는 머신러닝' 과 '혼자 공부하는 머신러닝 + 딥러닝' 도서를 참고해서 만들었습니다.

머신러닝 시스템 - 지도학습

지도학습(Supervised learning)에는 알고리즘에 주입하는 훈련 데이터에 label이라는 원하는 답이 포함된다.

지도학습의 방식에는 분류(classification)과 회귀(regression) 두 가지가 있다.

- 분류는 스팸 필터와 같이 데이터가 어떤 label을 가질지 예측하는 작업을 의미한다.
- 회귀는 확률과 수치의 의미에서, 데이터가 스팸일 확률 또는, 중고차가 연식 / 주행거리에 따라 가격이 얼마로 책정될 지 등 타겟 수치를 예측한다.
- 일부 회귀 알고리즘은 분류에 사용할 수 있다. 로지스틱 회귀(Logistic Regression)는 클래스에 속할 확률을 출력한다.

지도학습 알고리즘

- k-최근접 이웃 (k-nearest neighbors)
- 선형 회귀 (linear regression)
- 로지스틱 회귀 (logistic regression)
- SVM (support vector machine)
- 결정 트리 (decision tree) & Random Forest
- neural networks (→ DNN; deep neural network)

머신러닝 시스템 - 비지도 학습

비지도 학습(Unsupervised learning)은 훈련 데이터에 label이 없다. 시스템이 아무런 도움 없이 학습해야 한다. 데이터의 분포, 데이터 간 거리 정보 등 여러 특성을 알고리즘이 분석해야 하고 분석 결과를 바탕으로 데이터를 분류하거나 특정 값을 도출할 수 있다.

비지도학습 알고리즘

- clustering
 - k-means
 - DBSCAN
 - 계층 군집 분석 (hierarchical cluster analysis)
 - 이상치 탐지 & 특이치 탐지
 - isolation forest
- 시각화와 차원 축소
 - 주성분 분석 (principal component analysis; PCA)
 - kernel PCA
 - 지역적 선형 임베딩 (locally-linear embedding)

- t-SNE

머신러닝 시스템 - 준지도 학습

어떤 알고리즘은 일부만 레이블이 있는 데이터를 다룰 수 있고 이 때 사용하는 방식이 준지도 학습(semisupervised learning) 이다.

대부분의 준지도 학습은 지도 학습과 비지도 학습의 조합으로 이루어져 있다. 제한된 볼츠만 머신 (restricted Boltzman machine; RBM) 을 여러 겹으로 쌓은 심층 신뢰 신경망 (deep belief network; DBN) 을 주로 사용한다.

머신러닝 시스템 - 강화학습

강화 학습 (reinforcement learning) 은 앞서 언급한 알고리즘과는 성격이 많이 다르다. 학습하는 시스템을 에이전트라고 부르며 환경을 관찰해서 행동을 실행하고 그 결과로 보상 또는 벌점을 받는다. 시간이 지남에 따라 가장 큰 보상을 얻기 위해 '정책' 이라고 부르는 최상의 전략을 스스로 학습한다.

다음의 내용은 데이터가 input으로 적용되는 방식에 따라 머신러닝 시스템이 어떻게 진행되는지에 관한 내용이다.

배치 학습 (batch learning) 과 mini-batch

배치 학습은 전체 데이터셋을 사용해 훈련 하는 방식으로, 새로운 데이터에 대해 학습하려면 새로운 데이터를 포함한 이전 데이터 전체를 사용하여 시스템의 새로운 버전을 처음부터 다시 훈련해야 한다. 컴퓨팅 자원과 훈련 시간에 한계가 있다.

배치 학습에 변화를 주어 점진적으로 데이터를 학습하는 방식은 데이터를 한 개씩 또는 mini-batch 라 부르는 작은 묶음 단위로 데이터를 주입하여 시스템을 훈련시킨다. 점진적 학습은 알고리즘이 데이터 일부를 읽어 들이고 훈련 단계를 수행하고, 전체 데이터가 모두 적용될 때까지 이 과정을 반복한다.

점진적 학습에서 변화하는 데이터에 얼마나 빠르게 적응할 것인지 결정하는 학습률이 중요한 파라미터이다.

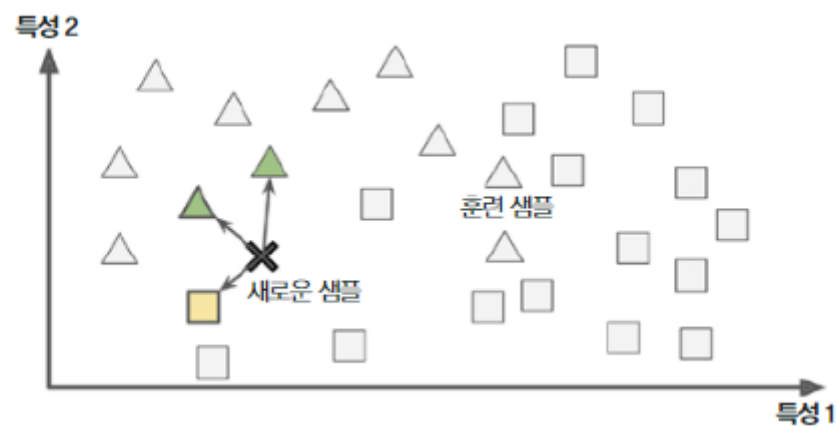
점진적 학습의 치명적 단점은 나쁜 데이터에 약하다는 것이며 이를 해결하기 위해 성능 감소 시 즉각 학습을 중단하는 방법이나, 입력 데이터를 모니터링해서 비정상 데이터를 판별하는 방법을 사용할 수 있다.

머신러닝 시스템의 목적

머신러닝 시스템의 목적은 훈련 데이터에 잘 맞는 모델을 찾아내는 것이 아니다. 궁극적인 목적은 새로운 데이터에서 좋은 예측을 만들어내는 것이며 시스템이 일반화되어야 한다는 것과 같은 의미이다.

사례 기반 학습

사례 기반 학습은 일반화의 첫 번째 접근법이다. 이 방식은 시스템이 훈련 샘플을 기억함으로써 학습한다. 유사도 측정을 통해 새로운 데이터와 학습한 샘플을 비교하는 식으로 일반화 한다.



모델 기반 학습

샘플들의 모델을 만들어 예측에 사용하는 것이 모델 기반 학습이다. 간단한 선형 모델을 가지고 데이터의 분포를 설명하는 방식과 유사하다. 이 때 선형 모델에 사용되는 모델 파라미터는 데이터를 나타내는 어떤 특징이라고 볼 수 있으며 새로운 데이터 예측에 영향을 미친다.

파라미터가 적절한지 확인하기 위해 모델이 전반적인 성능 평가가 필요하며 이 때 사용되는 함수가 비용 함수 또는 손실 함수이다. 손실 함수를 사용해 error 값이 낮아지도록 파라미터를 업데이트 할 수 있다.

요약

- 머신러닝은 명시적인 규칙을 코딩하지 않고 기계가 데이터로부터 학습하여 어떤 작업을 더 잘하도록 만드는 것이다.
- 여러 종류의 머신러닝 시스템이 있다. 학습 방식에 따라 (지도 학습 / 비지도 학습 / 준지도 학습 / 강화 학습), 데이터 사용 전략에 따라 (배치 학습 / 점진적 학습).
- 학습 알고리즘이 모델 기반이면 훈련 세트에 모델을 맞추기 위해 모델 파라미터를 조정하고(즉, 훈련 세트에서 좋은 예측을 만들기 위해), 새로운 데이터에서도 좋은 예측을 만들 거라 기대한다. 알고리즘이 사례 기반이면 샘플의 특성을 기억하는 것이 학습이고 유사도 측정을 통해 새로운 샘플을 비교하는 식으로 일반화 한다.

(+추가)

- 훈련 세트가 너무 작거나, 대표성이 없는 데이터이거나, 잡음이 많고 관련 없는 특성으로 오염되어 있다면 시스템이 잘 작동하지 않는다. 마지막으로, 모델이 너무 단순하거나 데이터가 적거나(과소적합), 모델이 너무 복잡(과대적합) 하지 않아야 한다.

테스트와 검증

모델을 평가하고 상세하게 튜닝하기 위해서 갖가지 전략을 이용할 수 있다.

데이터를 훈련 세트와 테스트 세트 두 개로 나누어 훈련 세트를 사용해 모델을 훈련하고 테스트 세트를 사용해 모델을 테스트 한다. 테스트 세트에서 모델을 평가함으로써 오차에 대한 추정값을 얻는다. 훈련 오차가 낮지만, 테스트 오차가 높다면 이는 모델이 훈련 데이터에 과대 적합된 것이다.

조금 다른 얘기로, test dataset 은 독립적으로 유지되어야 한다.

이 부분이 중요한데, 더 나은 모델을 만들기 위해 테스트 셋을 활용한다면 테스트 셋에 최적화된 모델이 만들어지기 때문이다. 새로운 데이터에 모델이 예상만큼 잘 작동하지 않을 수 있다.

구체적으로,

train dataset / validation dataset / test dataset 으로 나누어 모델 학습 및 평가, 실제 새로운 데이터에 모델을 적용하는 방식을 많이 활용한다.

validation dataset 은 모델 최종 평가 이전 시범 평가라고 생각하면 된다. validation dataset 은 모델이 학습 할 때 train dataset 과 같이 들어가지만, 모델의 학습에는 영향을 주지 않는다. 즉, 파라미터 업데이터에는 영향을 주지 않는다. validation set을 활용하면 과대, 과소 적합의 상황을 모니터링 할 수 있기에 유용하다.

validation set 의 크기가 너무 작다면, 모델 시범 평가가 제대로 이루어지지 않을 수 있다. 이러한 가능성을 방지하기 위해 작은 validation set 을 여러 개 사용해 반복적인 교차 검증 (cross validation) 을 수행하는 전략이 있다. validation set 마다 나머지 데이터에서 훈련한 모델을 해당 validation set에서 평가한다. 모든 모델의 평가 점수를 평균하면 훨씬 정확한 성능을 측정할 수 있다. 하지만 연산의 복잡성이나, 단순 연산 수가 증가하는 경향이므로 학습 시간과 모델 복잡성에 약점이다.