**Chapter 1. What Is Machine Learning?**

[기계 학습은] 컴퓨터가 명시적으로 프로그래밍하지 않고 학습 할 수있는 학습 영역입니다.

컴퓨터 프로그램은 P로 측정 한 T에서의 성능이 경험 E로 향상되면 일부 작업 T와 성능 측정 P에 대한 경험 E로부터 학습을한다고합니다.

즉, 프로그램은 경험된 E로부터 학습을 한다.

스팸필터를 예로

스팸 이메일 (예 : 사용자가 신고) 및 일반 (비 스팸, "햄"이라고도 함) 이메일

시스템이 학습에 사용하는 예제를 학습 세트라고합니다. 각 교육 예제를 교육 인스턴스 (또는 샘플)라고합니다.

Task T는 새 이메일에 대해 스팸 플래그를 지정

Expericence E는 학습 데이터

Performance P 정확하게 분류 된 전자 메일의 비율

전통적인 프로그래밍 기법을 이용한 스팸필터와

머신러닝을 이용한 스팸필터의 차이

데이터를 통해 기계가 대신 학습 해준다

대량의 데이터에서 보이지 않는 패턴들에 대해서도 발견할 수 있게되는 data mining

요약하자면

1 기존 솔루션에 많은 수작업 튜닝 또는 긴 규칙 목록이 필요한 문제 : 한 가지 학습 알고리즘으로 코드를 단순화하고 성능을 향상시킬 수 있습니다.

2 기존의 접근법을 사용하여 전혀 좋은 해결책이없는 복잡한 문제 : 최상의 기계 학습 기술로 해결책을 찾을 수 있습니다.

3 변화하는 환경 : 기계 학습 시스템은 새로운 데이터에 적응할 수 있습니다.

4 복잡한 문제 및 많은 양의 데이터에 대한 통찰력 확보.

머신러닝 시스템의 종류

1 인간 감독 (감독, 감독되지 않음, 반 감독, 강화 학습)

Whether or not they are trained with human supervision(supervised, unsupervised, semisupervised, and Reinforcement Learning)

2 비행 중 점진적으로 학습 할 수 있는지 여부 (온라인 대 배치 학습)

Whether or not they can learn incrementally on the fly (online versus batch learning)

3 단순히 새로운 데이터 포인트를 알려진 데이터 포인트와 비교하는지 여부에 관계없이, 또는 교육 데이터에서 패턴을 감지하고 과학자와 마찬가지로 예측 모델을 구축 할 수 있습니다 (인스턴스 기반 대 모델 기반 학습)

Whether they work by simply comparing new data points to known data points, or instead detect patterns in the training data and build a predictive model, much like scientists do (instance-based versus model-based learning)

* 1. Whether or not they are trained with human supervision 1)supervised 지도학습(감독학습) 라벨(레이블) 값을 주는 학습방법 트레이닝 셋의 메일에 대해 정상 또는 비정상이라는 답을 부여 -분류(Classification) 스팸이냐 아니냐 구분 -회귀(Regression) 주식과 같이 변동하는 연속적인 데이터들에 대해 예측하는 방식 k-Nearest Neighbors Linear Regression  Logistic Regression  Support Vector Machines (SVMs)  Decision Trees and Random Forests  Neural networks2  2)unsupervised 비지도학습(무감독학습) 라벨 값을 주지 않는 방법 - 군집화(Clustering) 알고리즘 - Visualization 시각적으로 2D나 3D로 나타낼 수 있음, 차원축소는 많은 정보에 대해 손실없이 데이터를 간략하게 하는 것 dimensionality reduction 속도 빠르고 디스크와 메모리 덜사용 성능도 좋을 수 있음 비정상 탐지(anomaly detection)새로운 데이터가 들어왔을 때 기존에 있는 데이터와 비교하여 비정상인지 감지 - 연관규칙 학습(Association rule learning) 데이터셋 속성들 사이에 관계를 발견  -Clustering  k-Means  Hierarchical Cluster Analysis (HCA)  Expectation Maximization  -Visualization and dimensionality reduction  Principal Component Analysis (PCA)  Kernel PCA Locally-Linear Embedding (LLE) t-distributed Stochastic Neighbor Embedding (t-SNE)  -Association rule learning  Apriori  Eclat  3)semisupervised 준지도학습 지도학습과 비 지도학습 섞어놓은 것 지도학습과 비지도학습의 결합 알고리즘 deep belief networks (DBNs)는 RBMs라는 비지도학습에 기반하여 학습되며 전체시스템은 지도학습 기법을 통해 fine-tuned된다  4)Reinforcement  강화학습은 agent가 주변환경 관찰, 환경에 맞는 선택 후 행동 수행, 정확한 행동은 보상rewards, 아니면 벌점penalty 부여 예를 들어 알파고나 로봇

1. Whether or not they can learn incrementally on the fly 머신러닝 시스템을 구분하는 다른 기준으로 incoming data로부터 점진적으로 학습할 수 있는 시스템인지 아닌지에 대한 분류 1)batch 배치 학습에서는 이용가능한 모든 데이터를 학습해야 하는 점진적으로 학습할 수 없는 시스템이다. (많은 시간과 컴퓨팅 리소스를 사용하므로 일반적으로 오프라인으로 수행됨) 시스템이 학습 되면 더이상의 학습이 없이 적용  새로운 데이터를 학습 시키려면 기존 데이터 또한 학습 -> 기존 시스템 중단 -> 새로운 시스템으로 교체 So 전체 프로세서는 자동화 처리 될 수 있지만 새로운 데이터 학습시킬 때 시간과 자원 소모가 크고 변동이 심한 데이터는 적합하지 않다.  2)online 온라인 학습에서는 개별적으로 또는 미니 배치라고하는 작은 그룹별로 데이터 인스턴스를 순차적으로 공급하여 점진적으로 시스템을 교육합니다. 각 학습 단계는 빠르고 저렴 하므로 시스템은 새로운 데이터가 도착하면 즉시 그것에 대해 알 수 있습니다. 한정된 컴퓨팅 리소스에 대해서도 좋은 학습방법 새로운 데이터에 대해서만 학습을 하기때문에 기존의 것들은 버릴 수 있다. 가장 중요한 파라미터 중 하나인 learning rate(어떻게 변화하는 데이터를 빠르게 적용할 수 있나) 높으면 -> 빠르게 적용 ->기존데이터 빠르게 잊음 낮으면 -> 느리게 학습 -> 새로운 데이터 노이즈 값에 덜 반응 가장 큰 과제는 안좋은 데이터가 유입될 때 시스템 성능이 서서히 감소 -> 모니터링 잘 해야하며 성능 저하되는 즉시 학습 중단  - 온라인 학습 역시 오프라인으로 수행(incremental learning으로 생각)
2. 중요!!!!!!!!! non linear 모델에서 잘됨 Whether they work by simply comparing new data points to known data points, or instead detect patterns in the training data and build a predictive model, much like scientists do 또 다른 구분 방법은 일반화하는 방법 대부분의 기계 학습 과제는 예측에 관한 것입니다. 이것은 여러 가지 훈련 예제가 주어지면 이전에 보지 못한 예를 일반화 할 수 있어야 함을 의미 진정한 목표는 새로운 instance에서 잘 수행하는 것 1)instance-based (knn알고리즘) 알려진 스팸 전자 메일과 동일한 전자 메일을 플래그하는 대신 스팸 필터는 알려진 스팸 전자 메일과 매우 유사한 전자 메일을 플래그하도록 프로그래밍 할 수 있습니다. 이를 위해서는 2 개의 전자 메일 간의 유사성 측정이 필요합니다. 두 이메일 간의 (매우 기본적인) 유사성 척도는 공통된 단어의 수를 세는 것일 수 있습니다. 알려진 전자 메일과 공통된 단어가 많은 경우 시스템에서 전자 메일을 스팸으로 표시합니다.  이를 인스턴스 기반 학습이라고합니다. 시스템은 예제를 핵심으로 학습 한 다음 유사성 측정을 사용하여 새로운 사례로 일반화합니다  기존의 학습 데이터에 새로운 데이터가 추가되었을 때 데이터의 유사성을 측정하여 분류 하는것  2)model-based 일련의 예제에서 일반화하는 또 다른 방법은 이러한 예제의 모델을 작성한 다음 해당 모델을 사용하여 예측을 만드는 것입니다. 이를 모델 기반 학습이라고합니다.  instance 의 경우 각각의 트레이닝 instance들과 유사성을 측정하는 것이고 model은 생성된 모델에 새로운 instance 값이 들어올때 해당 영역으로 분류를 하는 것

모델을 만드는 법 돈이 행복과 관련이 있는지 예제

Main challenges of machine learning

불충분한 학습 데이터 양 - 데이터 양이 많아질수록 정확성 올라감 대표성이 없는 데이터 - 누락 데이터 추가하면 모델이 이상해짐

품질이 낮은 데이터 - 오류, 이상, 노이즈 데이터

관련없는 기능 - 교육 자료에 관련성이 충분하고 관련성이없는 자료가 너무 많으면 시스템에서 학습 할 수 있습니다.

학습 데이터 오버피팅 - 지나친 일반화는 우리 인간이 너무 자주하는 일이며 불행히도 조심하지 않으면 기계가 같은 함정에 빠질 수 있습니다 오버피팅

학습 데이터 언더피팅 -

스탭핑 백

테스트 및 검증