머신러닝(배포용)

2018년 7월 28일 토요일 오전 5:59

1. 머신러닝이란?

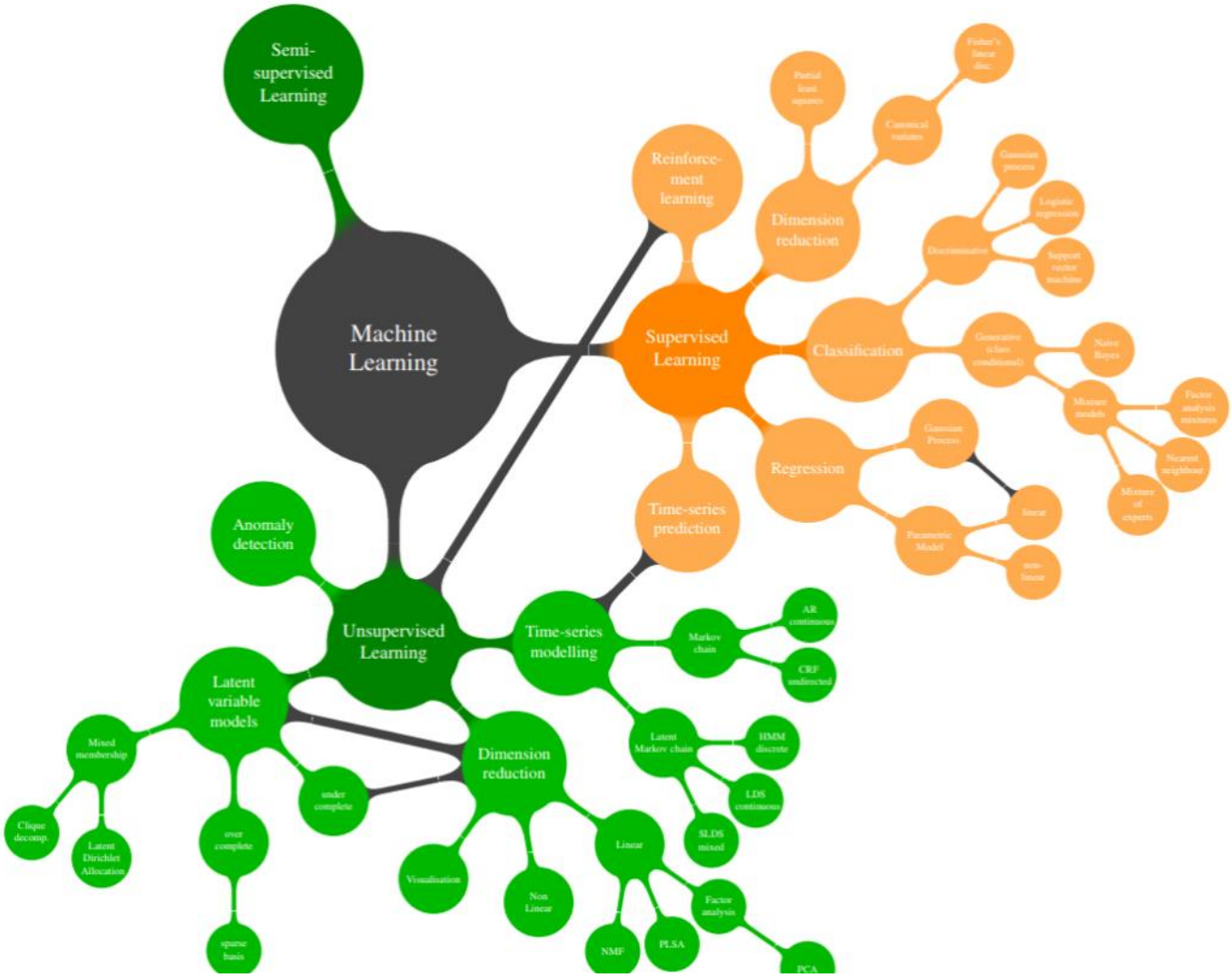
인공 지능의 한 분야로, 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야를 말한다

인공지능 > 머신러닝 > 딥러닝

"데이터" 에서 "정보"를 발견하는 방법

- 이때 사람의 개입 없이 머신 스스로 학습
  - 데이터: 학습 데이터(training data)
  - 정보: 학습 결과, 모델(model)

2. 머신러닝의 종류



지도 학습(Supervised Learning)

회귀(Regression)

분류(Classification)

비지도 학습(Unsupervised Learning)

군집화(Clustering)

군포 추정(Underlying Probability Density Estimation)

준지도 학습(Semi-supervised Learning)

학습 데이터가 약간의 레이블을 가지고 있음

일부는 지도 학습이고, 또 일부는 비지도 학습이다.

강화 학습(Reinforcement Learning)

최종 출력이 바로 주어지지 않고 시간이 지나서 주어지는 경우

1-1. 세부 종류들

예측

Linear regression, Regression tree, Kernel regression, Support vector regression

분류

Logistic regression, Decision tree, Nearest-neighbor classifier, Kernel discriminate analysis, Neural network, Support Vector Machine, Random forest, Boosted tree

차원(변수) 축소

Principal component analysis, Non-negative matrix factorization, Independent component analysis, Manifold learning, SVD

그룹화

k-means, Hierarchical clustering, mean-shift, self-organizing maps(SOMs)

선행학습(Pre-training), 2차 분류

Deep Learning(Stacked Restricted Boltzmann Machine, Stacked Auto-Encoders 등을 사용한 Multi layers Neural Nets, Non-linear Transformation)

데이터 비교

Bipartite cross-matching, n-point correlation two-sample testing, minimum spanning tree

1-2. 머신러닝의 응용 분야

a. 클래스 분류(Classification)

특정 데이터에 레이블을 붙여 분류할 수 있다.

예를 들어, 스팸 메일 분류, 필기 인식, 증권 사기 등에 사용하는 경우를 말한다.

b. 클러스터링-그룹 나누기(Clustering)

값의 유사성을 기반으로 데이터를 여러 그룹으로 나눌 수 있다.
예를 들어, 사용자의 취향을 그룹으로 묶어 사용자 취향에 맞는 광고를 제공하는 경우를 의미

c. 추천(Recommendation)
특정 데이터를 기반으로 다른 데이터를 추천하는 것이다.
예를 들어, 사용자가 인터넷 서점에서 구매한 책들을 기반으로 다른 책을 추천하는 경우를 의미

d. 회귀(Regression)
과거의 데이터를 기반으로 미래의 데이터를 예측하는 것이다.
판매 예측, 주가 변동 등을 예측하는 경우를 의미

e. 차원 축소(Dimensionality Reduction)
데이터의 특성을 유지하면서 데이터의 양을 줄이는 것이다.
어렵게 말하면 "특성을 유지한 상태로 고차원의 데이터를 저차원의 데이터로 변환하는 것"이다.
데이터를 시각화하거나 구조를 추출해서 용량을 줄여 계산을 빠르게 하거나 메모리를 절약할 때 사용한다.
(간단히 말해서 어떤 데이터가 있을 때, 특징을 추출하는 것을 의미한다. 예를 들어, 사람의 얼굴을 분석한다고 하자. 사람 얼굴 이미지는 용량이 꽤 되는 고차원의 데이터이다. 이러한 이미지에서 눈의 크기, 굴곡, 코의 위치, 코의 크기, 입의 위치 등을 분석해서 숫자로 추출해내면 이를 저차원의 데이터라고 부른다. 이렇게 만드는 것을 차원 축소라고 한다.)

f. 초과 학습(초과 적합)
초과 학습(Overfitting)이란 훈련 전용 데이터가 학습되어 있지만 학습되지 않은 새로운 데이터에 대해 제대로 된 예측을 못하는 상태를 의미한다.
즉, 배운 것밖에 해결하지 못하는 상황을 의미한다.
이러한 상황이 일어나는 원인은 다음과 같다.
    데이터가 너무 적은 경우
    모델에 비해 문제가 너무 복잡한 경우
데이터가 적을 경우의 근본적인 해결방법은 데이터의 수를 늘리는 것이다.
하지만 현실에서 수집할 수 있는 데이터의 수는 한정되어 있을 수 밖에 없다.
그리고 모델에 비해 문제가 너무 복잡한 경우에는 다른 모델을 선택해야 할 것이다.

1-3. 지도학습에서 사용되는 대표적인 모델

a. 서포트 벡터 머신 (support vector machine)
패턴 인식, 자료 분석을 위한 지도 학습 모델이며, 주로 분류와 회귀 분석을 위해 사용한다. 두 카테고리 중 어느 하나에 속한 데이터의 집합이 주어졌을 때, SVM 알고리즘은 주어진 데이터 집합을 바탕으로 하여 새로운 데이터가 어느 카테고리에 속할지 판단하는 비확률적 이진 선형 분류 모델을 만든다. 만들어진 분류 모델은 데이터가 사상된 공간에서경계로 표현되는데 SVM 알고리즘은 그 중 가장 큰 폭을 가진 경계를 찾는 알고리즘이다.

b. 은닉 마르코프 모델 (Hidden Markov model)
통계적 마르코프 모델의 하나로, 시스템이 은닉된 상태와 관찰가능한 결과의 두 가지 요소로 이루어졌다고 보는 모델이다. 관찰 가능한 결과를 야기하는 직접적인 원인은 관측될수 없는 은닉 상태들이고, 오직 그 상태들이 마르코프 과정을 통해 도출된 결과들만이 관찰될 수 있기 때문에 '은닉'이라는 단어가 붙게 되었다. 음성 인식, 필기 인식, 동작 인식,품사 태깅, 악보에서 연주되는 부분을 찾는 작업, 부분 방전, 생물정보학과 같이 시간의 영향을 받는 시스템의 패턴을 인식하는 작업에 유용한 것으로 알려져있다.

c. 회귀 분석 (Regression)

통계학에서, 회귀분석(regression analysis)은 관찰된 연속형 변수들에 대해 두 변수 사이의모형을 구한뒤 적합도를 측정해 내는 분석 방법이다. 회귀분석은 시간에 따라 변화하는 데이터나 어떤 영향, 가설적 실험, 인과 관계의 모델링등의 통계적 예측에 이용될 수 있다.

d. 신경망 (Neural network)

인공신경망은 생물학의 신경망(특히 뇌)에서 영감을 얻은 통계학적 학습 알고리즘이다. 인공신경망은 시냅스의 결합으로 네트워크를 형성한 인공 뉴런(노드)이 학습을 통해 시냅스의 결합 세기를 변화시켜, 문제 해결 능력을 가지는 모델 전반을 가리킨다. 좁은 의미에서는 역전파법을 이용한 다층 퍼셉트론을 가리키는 경우도 있지만, 인공신경망은 이에 국한되지 않는다.

e. 나이브 베이즈 분류 (Naive Bayes Classification)

특성들 사이의 독립을 가정하는 베이즈 정리를 적용한 확률 분류기의 일종으로 1950 년대 이후 광범위하게 연구되고 있다. 텍스트 분류에 사용됨으로써 문서를 여러 범주 (예: 스팸, 스포츠, 정치)중 하나로 판단하는 문제에 대한 대중적인 방법으로 남아있다.

1-4. 머신러닝의 실제 사용 사례

▸ 사기 방지: 1억 5,000만 개의 디지털 월릿을 통해 연간 2,000억 달러 이상의 결제를 처리하는 페이팔(PayPal)은 온라인 결제업계의 선두 주자다. 이 정도 규모에서는 사기 비율이 낮다해도 그 비용은 상당하다. 창업 초기에는 월별 사기 피해 금액이 1,000만 달러에 이르렀다.

▸ 타겟팅 디지털 디스플레이: 광고 기술 기업 Dstillery는 실시간 입찰 플랫폼에서 타겟 디지털 디스플레이 광고를 진행하도록 한다. 개인의 브라우징 내역, 방문, 클릭 및 구매에 대해 수집된 데이터를 사용해 한 번에 수백 개의 광고 캠페인을 처리하며 초당 수천 건의 예측을 실행한다.

▸ 콘텐츠 추천: Comcast는 TV 서비스 고객을 위해 각 고객의 이전 시청 습관을 기반으로 한 실시간으로 개인 맞춤화된 콘텐츠를 추천한다. 수십억 개의 내역 기록을 사용해 각 고객별로 고유한 취향 프로필을 작성한 다음, 공통적인 취향을 가진 고객을 클러스터로 묶는다. 그런 후 각 고객 클러스터를 대상으로 가장 인기있는 콘텐츠를 실시간으로 추적 및 추천한다.

▸ 자동차 품질 개선: Jaguar Land Rover의 신형 차량에는 60개의 온보드 컴퓨터가 탑재되며 이 컴퓨터는 2만 개 이상의 메트릭스를 기준으로 매일 1.5GB의 데이터를 생성한다. 고객이 차량을 실제로 어떻게 다루는 지를 파악해서 얻은 정확한 사용 데이터를 통해 설계자는 부품 고장과 잠재적 안전위험을 예측할 수 있다. 조건에 맞는 차량 엔지니어링에도 도움이 된다.

▸ 유망 잠재 고객에 집중: 마케터들은 최적의 판매와 마케팅 기회, 그리고 최적의 제품을 판단하기 위한 도구로 구매 성향 모델을 사용한다. 라우터부터 케이블 TV 박스에 이르기까지 방대한 제품을 보유한 Cisco의 마케팅 분석팀은 몇 시간 만에 6만 개의 모델을 교육시키고 1억 6,000만 명의 잠재 고객을 확보했다.

▸ 의료 보건 서비스 개선: 환자의 재입실은 의료 보험 공단과 민간 보험사가 재입실 비율이 높은 병원에 불이익을 줄 수 있다. 건강한 상태를 유지할 가능성이 충분히 높은 환자만 퇴원시키는 역량이 병원의 재무에 큰 영향을 미치게 된다. Carolinas Healthcare System은 환자의 위험 점수를 계산하고 병원 사례 관리자는 이를 바탕으로 퇴원 결정을 내린다.