"머신러닝"은 인공지능(AI)의 한 분야로, 데이터 분석을 위한 모델 생성을 자동화 하여 소프트웨어가 데이터를 바탕으로 학습하고 패턴을 찾아냅니다. 이를 통해 사람의 개입을 최소화 하고 빠르게 의사 결정을 내릴 수 있도록 지원합니다.

많은 양의 데이터를 처리하고 분석하는 대부분의 산업에서는 머신러닝을 적극적으로 활용하고 있습니다. 대부분 실시간 데이터를 분석하고 인사이트를 얻음으로써 기업은 보다 효과적으로 기회를 포착하고 경쟁 우위를 획득할 수 있습니다(예: 금융서비스, 정부, 의료서비스. 교육, 운송, 기업 마케팅, 영업, 에너지 등)

미래를 상상하여 만든 영화에서만 볼 수 있을것이라 여겨지던 것들이 이제는 Google 검색 결과 같이 평범한 사물까지도 머신러닝의 산물이 되고 있습니다. 넷플릭스(Netflix)는 머신러닝을 사용하여 개인 맞춤형 영화 추천 서비스를 제공하고 있습니다. eHarmony는 머신러닝을 통해 사랑까지 수량화하여 예측합니다. 은행들은 사이버 감시를 비롯해 사기 및 악용 여부를 모니터링하는 데까지 머신러닝을 활용하고 있습니다.

위와 같이 인상적인 머신러닝 적용 사례가 많았습니다. **가장 기억에 남는 머신러닝 적용사례 한 가지는**머신러닝을 웨어러블 의료기기를 통해 실현한 적용사례입니다. 헬스케어 기술 및 컨설팅 기업인 Geneia의 예입니다.

빅 데이터와 머신러닝을 함께 사용함으로써 헬스케어 기업이 보다 낮은 비용으로 환자를 치료하도록 돕고 있습니다.

임상 데이터, 실험 데이터, 생리 데이터, 보험 통계 데이터, 소비자 데이터 등 여러 가지 다른 소스에서 데이터를 수집하고 그 결과 Theon이라고 하는 데이터 통합 플랫폼을 개발하여 머신러닝 기법을 적용하는 형태로 이루어졌다고 합니다.

Geneia는 머신러닝을 사용하여 값을 서로 비교한 후 누락된 값을 발견할 뿐만 아니라 주성분 분석을 실시하여 데이터 패턴, 즉 추이와 인과관계를 알려주는 군집 형태를 살펴볼 수 있습니다.

머신러닝을 통해 이전에는 미쳐 발견하지 못했던 데이터 패턴을 찾아낼 수 있습니다.

또한 웨어러블 의료 기기와 센서로 훨씬 더 많은 정보를 수집할 수 있게 되었습니다. 대기 정보, 습도, 냉장고를 개방하거나 화장실을 사용한 횟수 등 우리는 개인에 대한 이해뿐만 아니라 모집단의 동향에 대해서도 더욱 정확하게 파악할 수 있게 되었습니다.

Geneia는 머신러닝을 이 모든 데이터에 적용한 뒤로 데이터 정확도가 높아지면서 과거 임상 평가나 실험 값에 비해 조기에 예측하여 신속히 개입할 수 있게 되며, 특히 이러한 현상은 변화를 더욱 빠르게 나타내는 연속형 데이터 세트에서 더욱 두드러졌습니다.

활용 예를 들면, 울혈성 심부전을 앓고 계신 할머니가 집에서 회복 중이신 경우에는 웨어러블 의료 기기가 체중이나 맥박 산소 수치, 또는 호흡률의 변화를 빠르게 감지할 수 있습니다. 이렇게 수집되는 데이터는 모두 Theon 애플리케이션에 입력되어 건강 상태가 더 악화되기 전에 미리 감지하고 알려줄 수 있습니다.”라고 설명했습니다.

노령 인구층을 위한 헬스케어 서비스 제공업체가 충분하지 않기 때문에 이전과는 다른 방법으로 환자를 진료할 수 있는 새로운 방법이 절실합니다. 머신러닝이 적용된 웨어러블 의료 기기는 사람들의 건강을 증진하여 수명을 늘릴 뿐만 아니라 환자가 집과 같이 가장 편한 곳에서 가족과 함께 요양할 수 있도록 하는 데 커다란 기여를 할 것입니다 앞으로도 지속적으로 활용 가능한 도메인이며 머신러닝이 사회에 긍정적인 영향을 미칠 수 있는 한 분야라고 생각합니다.

규모에 상관없이 모든 산업군에 종사하는 기업들에서 머신러닝을 사용하는 사례가 점차 늘어나면서 앞으로 머신러닝은 “비즈니스 경쟁에서 빠지지 않는 필수 기법”이 될 것이라고 봅니다.

아무리 훈련이 잘 된 모델이라도 제약 조건을 무시할 수는 없을 것입니다. 그러나 머신러닝을 적용하여 이전까지 서로 무관한 데이터 세트(환경, 사회 경제, 생체 측정, 금융, 유전체, 농업 등)가 통합되는 미래를 기대해 봅니다.

NLP 기술로 이메일함에서 스팸을 걸러낼 수 있는 감지 모델 (spam detection)을 만들수 있다. 이미 Gmail 같은 서비스에는 이러한 기술들이 널리 쓰이고 있다고 들었습니다. 또한 기본적인 통계 모델 Naive Bayes Classifier를 사용할 수있다.

가장 간단한 0이냐 1이냐의 이분법 (binary classification)을 logistic regression이라는 통계 모델을 가지고 학습시키는 예시를 볼 수 있다. Spam detection도 이분법으로 나눌 수 있고 문제는 우리가 어떠한 데이터를 가지고 있냐이다. 영화 리뷰를 긍정이냐 부정이냐로 나눌 때에는 부정 (negative; 0)과 긍정 (positive; 1)의 학습/평가 데이터 셋을 가진다. 스팸 이메일 역시 이러한 학습 데이터가 필요하다. **스팸이 아닌 이메일** (negative; 0), 그리고 **스팸 이메일** (positive; 1). 이러한 데이터를 가지고 있다면 바로 binary classification 문제로 정의하고 해결책을 찾아볼 수 있다(지도학습)

데이터가 없다면, 가장 쉬운 방법은 지금 이메일함에 들어가서 몇 백개의 이메일을 가져온 다음에 스팸인지 아닌지 직접 분류를 하는 것 이를 **human annotation**또는**data labeling** 라고 한다. 통계 모델을 학습시킬 정도의 규모가 될 때까지 직접 사람이 학습 데이터를 만들어 내는 것이다.

세상에는 labeling 되지 않은 데이터가 엄청나게 많습니다. 이메일 같은 경우에도 여러분의 편지함에 엄청나게 쌓여 있지만 어떤 것이 스팸인지 아닌지는 labeling 하기 전에는 알 수가 없다. 이러한 labeling 과정이 보통 머신러닝 모델 개발에 가장 큰 비용과 시간이 들어간다. 그렇기 때문에 크라우드소싱을 이용한 crowdworker 서비스를 통해 노동력이 싼 개발 도상국의 인력을 이용하기도 한다.

네이버나 구글 같은 서비스들은 유저들이 스팸이라고 신고도 데이터가 될 수 있겠다고 생각한다.

또한, **anomaly detection**이라는 테크닉이다. anomaly는 변칙이라는 뜻인데, 데이터 전체를 보았을 때 다른 데이터 포인트들과 아주 동떨어진 것들을 찾아내는 것, 이는 데이터 자체를 보기 때문에 label이 따로 필요하지 않다. 다만 데이터의 특성에 따라 성능의 차이가 많이 날 수 있고, 제대로 평가하기가 힘들 수 있다. 예를 들어, 이메일 유저에게 유용한 이메일들이 대부분 비슷한 주제로만 이루어져 있다면, 갑자기 다른 주제인 스팸을 찾아내는 건 쉬울 수 있다. 하지만 저의 이메일들이 매우 다양한 주제를 가지고 있다면 이러한 모델로 스팸을 찾아내는 것은 어려울 수 있다(비지도 학습).

대표적인 5가지의 task

1. classification은 지도학습의 일종으로, 기존에 존재하는 데이터와 category와의 관계를 learning하여 새로 관측된 데이터의 category를 판별하는 문제이다. 스팸 필터를 예로 들어들어보자. 스팸 필터의 데이터는 이메일이고, category, 혹은 label, class는 spam메일인지 일반 메일인지를 판별하는 것이 될 것이다. 스팸필터는 먼저 스팸 메일, 그리고 일반 메일을 learning을 한 이후, 새로운 데이터 (혹은 메일)이 input으로 들어왔을 때 해당 메일이 스팸인지 일반 메일인지 판별하는 문제를 풀어야하며, 이런 문제를 classification이라고 한다. classifier:Decision Tree, Regression Tree and Ensemble method, Naïve Bayes, KNN

2. clustering 은 비지도학습의 일종으로, label 데이터 없이 주어진 데이터들을 가장 잘 설명하는 cluster를 찾는 문제이다. Classification을 하기 위해서는 데이터와 각각의 데이터의 label이 필요하지만, 실제로는 데이터는 존재하지만 그 데이터의 label이나 category가 무엇인지 알 수 없는 경우가 많기 때문에 classfication이 아닌 다른 방법을 통해 데이터들을 설명해야하는 경우가 발생한다. 각각의 데이터에 대한 정보는 아무 것도 없는 상태에서 주어진 데이터들을 가장 잘 설명하는 클러스터를 찾아내는 것이 클러스터링의 목적이다. 따라서 클러스터링은 대부분 Optimization 문제를 푸는 경우가 많다.

-K-means: 거리 기반으로 세우고 그것을 푸는 알고리즘, 클러스터를 정의하는 방법에는 여러 가지가 있을 수 있지만, 가장 간단한 정의 중 하나는 클러스터 내부에 속한 데이터들이 서로 ‘가깝다’라고 정의하고, ‘가장 가까운’ 내부 거리를 가지는 클러스터를 고르는 것이다

-Gaussian Mixture Model:확률과 확률분포를 기반으로 세우고 그것을 푸는 알고리즘

3. regression, 회귀분석은 데이터 변수들간에 함수관계를 파악하여 통계적 추론을 하는 기술이다.

좀더 쉽게 설명하자면, 독립변수에 대한 종속변수값의 **평균**을 구하는 방법입니다.

h() 가 위에서 말한 조건에 따른 평균을 구하는 함수이며 우리는 이것을 보통 ‘회귀 모델’이라고 부릅니다.

이 함수는 어떤 조건(x1, x2, x3, …)이 주어지면 각 조건의 영향력(beta1, beta2, beta3, …)을 고려하여 해당 조건에서의 평균값을 계산해 주는 것인데, 뒤에 붙는 e 는 ‘오차항’을 의미한다. 측정상의 오차나 모든 정보를 파악할 수 없는 점 등 다양한 현실적인 한계로 인해 발생하는 불확실성이 여기에 포함된다.  이런 잡음은 이론적으로 보면 평균이 0이고 분산이 일정한 정규 분포를 띄는 성질이 있다. 회귀 분석을 한다는 것은 이 h() 함수가 무엇인지를 찾는 과정을 의미한다.

시간에 따라 변화하는 데이터나 어떤 영향, 가설적 실험, 인과 관계의 모델링등의 통계적 예측에 이용될 수 있다.

4.ranking: Learning to rank란

item의 정확한 score를 regression하는게 아니라, optimal ordering of list of items.

종류:Pointwise|Pairwise|Listwise

각각은 loss function에 들어가는 document의 개수에 따라 구분된다.

pointwise는 1개, Pairwise는 2개, Listwise는 pointwise는 document가 query에 해당하는 relevancy를 학습하고, 실제 inference할 때는 각 document의 relevancy 순으로 출력, 즉 관건은, documents끼리 independent하다는 것, classical classification/regression을 사용, pairwise는 a pair of documents를 보고 ordering을 매김, the number of inversions를 줄이게끔 loss function을 정의하여 사용, relative order를 예측하는 게 좀 더 ranking의 본질에 가깝다. 단점은 training/inference가 시간복잡도가 크고 사실 낮은 ranking에 존재하는 pair끼리의 계산이나 상위 ranking에 존재하는 pair끼리의 계산이 동일하게 들어가는 점이 단점,  listwise는 entire list of documents를 보고 optimal ordering을 찾는 것, 예를 들면, NDCG를 utility function으로 보고 학습, 장점이 training/inference의 시간복잡도가 낮다

**Recommender system과 Ranking의 차이점은?**

-ranking은 ordering이 결과임, 개별 item의 score도 predicted rating이 아니라, ordering으로서의 score이며 utility로서의 score가 아님

-ranking은 user의 input(query나 category선택, 지리적 정보 등)이 중요한 역할

5. recommendation: Recommendation problem은 여태까지 사용자가 item에 대해 evaluate한 history data를 기반으로 사용자가 아직 사용하지 않은 item에 대한 사용자의 평가를 예측하는 문제라고 할 수 있다. 추천과 랭킹 문제는 마케팅을 포함한 다양한 분야에서 오랜 세월 관심을 가져왔던 분야이다. 특히 광고를 제작하는 사람들 입장에서는 적은 비용으로 최대한의 효율을 낼 수 있는 타겟광고는 그야말로 금덩이나 다름없는 영역이라 할 수 있을 것이다. 추천은 그만큼 제한된 자원을 최대한 효율적으로 분배할 수 있는 방법이기도 하며, 사람들의 지갑을 더 열게 할 수 있는 중요한 문제인 것이다. Netflix와 왓챠는 내가 점수를 매긴 별점을 바탕으로 내가 좋아할만한 영화나 드라마를 추천해준다. 아마존, 이베이, Gmarket 등의 온라인 쇼핑몰들 역시 내가 클릭했던 상품들의 history를 기반으로 내가 좋아할만한, 사고싶어할만한 상품들을 추천해주거나, 이 상품들을 묶어서 하나의 작은 할인 패키지를 구성하기도 한다. 페이스북은 내가 좋아요를 누른 포스트들을 바탕으로 내가 좋아할만한 페이지를 추천하고, 타겟 광고를 내보낸다. 우리가 지금은 너무나 자연스럽게 받아들이는 이 사실들은 전부 머신러닝에 의해 가능해진 것들이다.