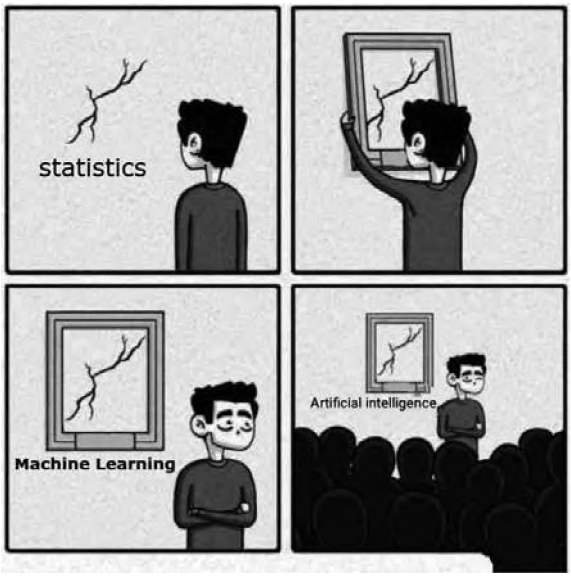
서범석 통계기획팀 과장

설명가능한 인공지능?

요새 여러 분야에서 머신러닝 방법론이 인기를 끌고 있다. 머신러닝은 알고리즘을 이용하여 데이터로부터 유의미한 패턴을 찾아내는 방법론을 총칭하는데, 워낙 인기를 끌다보니 그에 대한 반발 또한 큰 것 같다. 대부분의 반발은 알고리즘이 찾아낸 패턴을 신뢰할 수 없다는 주장이다. 이러한 주장의 기저에는 사람이 이해할 수 있어야 신뢰할 수 있다는 가정이 깔려있다. 머신러닝이 아무리 예측을 잘한다 한들 과정과 결과를 이해할 수 없다면 활용이 제한될 수밖에 없다. 그런데 정말 머신러닝의 방법론은 사람이 전혀 이해할 수 없는 것일까? 필자는 통계학을 공부하면서 설명가능한 머신러닝을 주제로 학위를 받았다. 슬기로운연구생활 원고 작성을 부탁받아, 그동안 연구했던 내용을 바탕으로 머신러닝이란 무엇인지, 설명가능한 머신러닝을 구현하기 위해 학계에서 어떤 연구들이 진행되고 있는지 간략히 소개해보고자 한다.

인공지능, 머신러닝, 그리고 통계모델링

우선 용어부터 정리를 해보자. 요새 인공지능이란 수식어가 인기다. 컴퓨터로 무언가를 계산했다하면 다들 인공지능이란 단어를 붙이는 것 같다. 인공지능을 어떻게 정의할지에 대해서는 논란이 있겠지만, 인공지능을 구현하기 위한 방법론으론 머신러닝이 대세를 이룬 듯 하다. 다시 말하지만 머신러닝이란 말 그대로 컴퓨터(머신)를 이용해서 데이터로부터 패턴을 찾아내는(러닝) 방법들을 말한다. 전통적으로 데이터에서 패턴을 찾아내는 방법론을 통계학으로 부르던 것과 개념상 같다. 그럼에도 불구하고 많은 학자들이 머신러닝을 통계학과 구분해서 사용을 하는데, 그 이유는 상당부분 앞으로 설명할 인공신경망이라는 모형 때문이다. 인공신경망 모형이 워낙 전통적인 통계모형과는 다른 접근법을 취하며 주로 컴퓨터공학자들을 중심으로 발전되어 왔기 때문일 것이다.



sandserif comics

설명을 쉽게 하기 위해 다음과 같은 상황을 가정해보자. 우리는 일종의 의사결정 과정을 다음과 같은 함수식으로 표현할 수 있다.

$$y = m(x)$$

여기서 y와 x를 부르는 용어는 분야마다 다른데 여기선 간단히 출력변수와 입력변수라고 하자.[1] 우리는 입력변수 x를 고려하여 어떤 결과 y를 얻고자 한다. 그것은 여러 상황을 고려해서 점심에 짜장면을 먹을지 짬뽕을 먹을지를 결정하는 문제일 수도 있고, 오늘의 경제지표를 보고 내일의 주가지수를 짐작하는 문제일 수도 있다.

많은 분야 이론가들은 연구자의 직감을 이용하여 m을 설명하는 반면[2], 통계학자들은 이 문제를 풀기위해 과거의 데이터 (y, x)를 살펴본다. 과거의 데이터를 가장 잘 설명하는 함수 m을 찾는 것이 이들의 목표이다. 과거에 비가 오는 날마다 짬뽕이 더 맛있게 느껴졌다면 '비오는 날에는 짬뽕'이라는 공식을 만들어 내는 것과 같다. 여기까지만 보면 인공지능, 머신러닝, 통계학이 모두 비슷한 말일 수 있겠다.
그렇다면 m을 과연 어떻게 찾아내야 하는걸까?

예측의 정확성 vs 수리적 미학

전통적으로 m을 찾기 위한 다양한 방법들이 제시되어 왔다. 혹자는 이들중 예측을 위한 방법론을 머신러닝, 분석을 위한 방법론을 통계학으로 나누어 설명하곤 한다. 그러나 이는 언뜻 보아도 매우 틀린 말이다. 통계학의 관심사는 늘 예측의 정확도를 높이는 것이었다. 통계를 배우면 가장 먼저 오차와 분산을 배우지 않는가.
m을 추정하기 위한 방법으로 보간법, 회귀법, 다항식 근사 등 여러가지를 생각해 볼 수 있는데, 역시 우리에게 가장 익숙한 방법은 다음과 같이 선형회귀모형을 이용하는 것이다.

$$m(x) = a + bx + e \ ,$$

e 는 평균이 0인 통계분포를 따름

선형회귀모형은 이해가 쉽고 여러가지 유용한 특성을 갖는다. 먼저 함수식이 간단해서 함수 m의 평균적인 의사결정과정(a+bx)을 명확히 인지할 수 있고 m의 불확실성(e) 또한 통계분포를 가정하여 모형화 할 수 있다. 덕분에 우리는 입력변수의 변화가 출력변수에 미치는 영향을 바로 알 수 있고, 통계분포 가정하에 여러가지 가설검정(hypothesis testing)도 해볼 수 있다. 다만 이 모형은 치명적인 단점이 있는데, y와 x의 관계가 복잡하다면 정확도가 떨어진다는 것이다. 점심에 짜장면을 먹을지 짬뽕을 먹을지를 결정하는 데만 해도 다이어트 중인지, 어제 기름진 음식을 먹었는지, 어제 기름진 음식을 먹었고 다이어트 중이지만 오늘은 치팅데이인지 등 복잡할 수 있다. 이 경우 선형모형은 의사결정과정을 지나치게 단순화하는 문제가 있다.

이러한 복잡한 상황에 이용할 수 있는 통계적 방법들이 바로 Kernel Regression, B-spline, LOWESS(locally weighted scatterplot smoothing) 등 비선형회귀와 비모수 방법론[3]들이다. 비록 선형회귀모형처럼 데이터의 패턴을 쉬운 분포함수로 표한하는 것은 불가능하지만, 비모수 방법론을 이용하면 복잡한 데이터의 패턴을 보다 높은 정확도로 예측하는 것이 가능하다. 다만 비모수 회귀방법은 모형에 제한을 두지 않고 m을 추정하려다 보니, m 함수의 형태와 분포를 가정하는 모수적 방법보다 훨씬 많은 데이터가 필요하고 계산도 오래걸린다.

인공신경망 모형

컴퓨터공학자 중심으로 발전된 인공신경망 모형은 이러한 문제를 대부분 해결했다. 인공신경망 모형은 다음과 같이 여러겹의 은닉층 h와 하나의 출력층 g를 합성하여 모형을 구성한다.

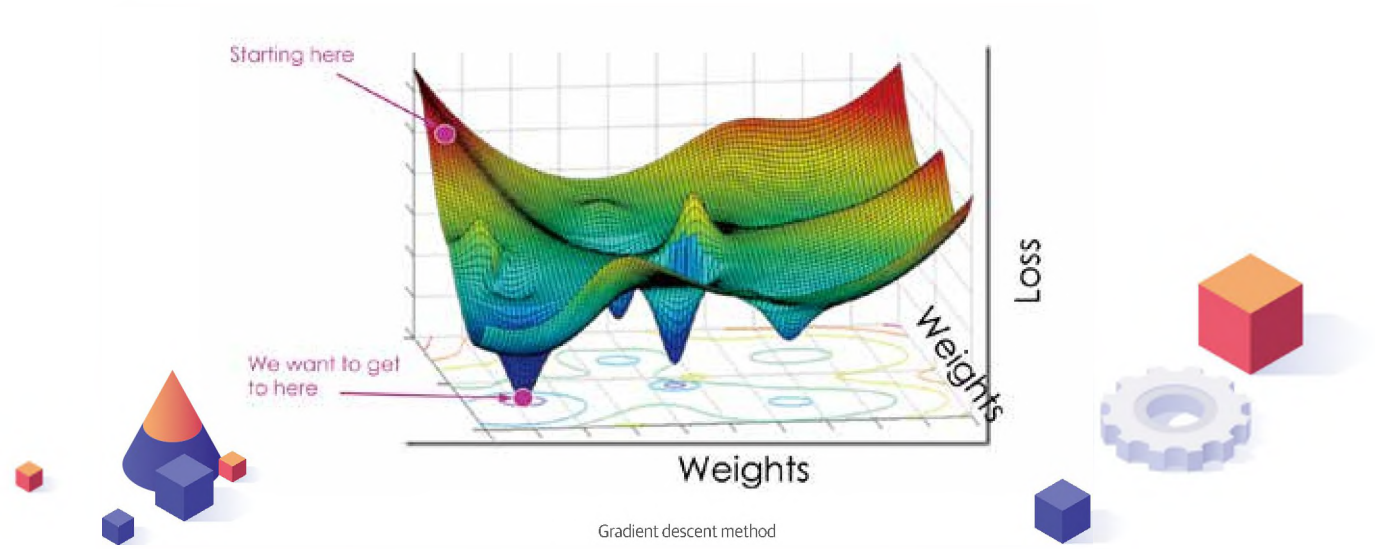
$$m(x) = g \circ h_L \circ \cdots \circ h_1(x)$$
$$z_l = h_l(z_{l-1}) = \sigma(uz_{l-1} + b), \quad \text{for } l = 1, \ldots, L \text{ and } z_0 = x.$$
$$g(z_L) = linear \text{ 또는 } softmax \text{ 함수}$$

여기서 활성화 함수라고 불리는 함수의 역할이 중요한데 보통 ReLU함수()를 이용한다. 활성화 함수는 선형방정식 wz+b를 비선형으로 바꿔주는 역할을 하게 된다. 이 때문에 비선형함수를 합성해서 만든 인공신경망 모형은 비선형의 m 함수를 잘 추정할 수 있다.

결과만 얘기하면 인공신경망 모형으로 추정한 m은 조각별 선형함수(piecewise linear function)의 형태를 갖게 된다. 이는 사진을 모자이크로 표현하는 것과 비슷하다. 더 많은 작은 조각을 이용할 수록 사진을 세밀하게 근사하는 것이 가능해지는데, 인공신경망 모형에서는 이러한 효과를 얻기 위해 은닉층의 수와 차원을 늘리면 된다. 다만 주어진 데이터를 너무 세밀하게 추정하다보면 불필요한 노이즈까지 추정하게 되는 과적합(over fitting)의 문제가 나타난다. 혹자는 과적합의 문제를 인공신경망 모형의 큰 단점으로 지적하는데, 과적합의 문제는 선형회귀 모형에서도 동일하게 나타난다. 다만 선형모형으로는 노이즈까지 추정하는 복잡한 모형을 구성하는 것이 어려울 뿐이다.

인공신경망 모형의 구조는 예측정확도를 높이는데도 탁월하지만, 모형을 자유롭게(flexible) 구성할 수 있다는 장점도 있다. 인공신경망 모형은 많은 파라미터를 포함하는 만큼 구조를 구성하는 것이 자유롭다. 입력변수를 연결하는 은닉층들을 적절히 연결하거나 분리하면 이미지, 텍스트는 물론 시계열 데이터의 특성을 모형에 쉽게 반영할 수 있다. 선형모형을 구성하며 고민하듯이 lag 변수를 하나 더 넣을지 말지를 고민할 필요가 없게 된다.

[1] y와 x 모두 다변량 혹은 일변량 변수일 수 있다. 여기서는 가장 흔히 생각할 수 있는 여러 변수(x)를 고려하여 하나의 의사결정(y)을 도출하는 상황을 가정하자.
[2] 이때도 귀납적 검증을 위해 통계적 방법론을 이용하지만 데이터마이닝 측면의 통계적 방법론은 검증보다는 새로운 지식의 발견(knowledge discovery)에 초점이 있다.
[3] 데이터의 분포(parametric distribution)를 미리 가정하지 않고 분석하는 통계적 방법들을 비모수(non-paraetric) 방법론이라고 한다.



Gradient descent method

인공신경망 모형과 전통적 회귀모형의 가장 큰 차이는 사실 모형을 어떻게 추정하는가에 있다. 전통적 회귀모형은 함수 m을 단순한 형태로 가정하기에 추정해야 할 미지수(parameter)의 수(p)가 데이터의 샘플 수(n)에 비해 매우 작은 상황이 된다. 이것은 미지수의 수(p)보다 방정식의 수(n)가 많은 연립방정식을 푸는 문제와 같다. 즉, 해가 존재하지 않게 된다. 따라서 우리는 모든 방정식을 만족시키는 답을 찾는 것이 아닌 모든 방정식에서 오차를 최소화하는 답을 찾는 것으로 문제를 바꿔 풀게 된다. 그러면 이 문제는 이제 단 하나의 최적해를 갖게 된다.

인공신경망 모형은 어떨까? 인공신경망 모형은 여러 차원의 출력값을 갖는 은닉층을 여러 겹으로 합성한 함수이다. 따라서 대부분의 경우 추정해야 할 미지수의 수가 데이터 샘플 수를 초월하는 상황이 되고, 이는 미지수가 방정식보다 많은 연립방정식을 푸는 문제와 같다. 즉, 해가 무수히 많이 존재하는 상황이 된다. 이 문제를 어떻게 풀면 좋을까? 답은 간단하다. 그저 작은 오차를 갖는 아무 해답이나 찾으면 되는 것이다. 어짜피 우리는 최적해가 뭔지 모른다. 적당한 해를 찾기 위해 데이터를 조금씩 대입해가면서 수치적 방법(gradient descent method)으로 하나의 해를 찾게 된다. 그렇게 찾은 적당한 답은 대부분의 경우 다른 방법으로 찾은 답보다 훨씬 정확하다.

인공신경망 모형은 은닉층의 수와 차원을 적절히 조절하면 어떤 복잡한 함수 m도 잘 근사할 수 있는 것으로 알려져 있다. 이를 Universal Approximation Theorem이라고 부른다.

설명가능한 인공신경망 모형

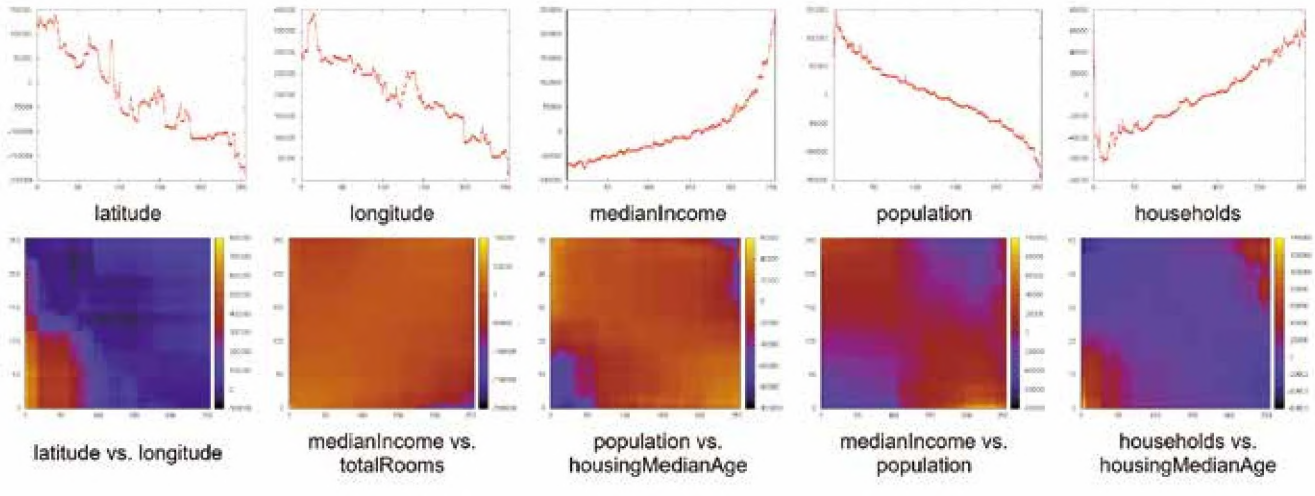
인공신경망 모형이 예측을 잘한다는 것은 널리 알려져 있다. 그러면 이제 복잡한 인공신경망 모형을 어떻게 설명할 수 있을지를 알아보자. 사람이 어떤 프로세스나 시스템(여기서는 함수 m)을 이해한다는 것은 사실 매우 철학적인 문제이다. 여기에 관해서는 많은 논의가 있는데, 사람이 '시스템이 결정한 결과의 원인을 이해할 수 있고 그 시스템의 결과를 일관적으로 예측'할 수 있으면 보통 설명가능한(interpretable) 시스템이라고 이야기할 수 있겠다. 우리는 선형모형의 회귀계수를 보는 것에 익숙하지만, 사실 모형에 대한 설명은 시각화 방법이 될 수도 있고 수치적 지표가 될 수도 있고 하나의 문장이 될수도 있다.

철학적인 문제들은 차치하고 구체적으로 이 문제를 학계에서 어떻게 접근하고 있는지 알아보자. 여러 연구가 진행 중이지만 크게 보면 두 가지 접근법으로 나누어 살펴볼 수 있다. 첫 번째는 복잡한 인공신경망 구조에 제약을 가하여 설명이 가능한 구조로 모형을 구성하는 것이다. 두 번째는 복잡한 인공신경망의 결과를 사람이 이해할 수 있는 방법으로 사후적으로 해석하는 방법이다. 예를 들어 모형의 전부를 설명하진 못하지만 특정 의사결정에 대해서 그 의사결정의 이유를 보여주는 방법이 그것이다.

각 접근법의 대표적 예를 살펴보면, 먼저 첫 번째 방법으로는 일반가법모형(GAM, Generalized Additive Model)이 있다. 앞서 언급했듯이 인공신경망 모형은 구조를 자유롭게 구성할 수 있는데, 그 중 출력변수에 미치는 영향을 입력변수 한개의 효과(main effect)와 두개의 상호작용 효과(interaction effect) 까지만 고려하여 모형을 구성한 것이 GAM이다.

$$m(x) = \sum_i f_i(x_i) + \sum_{i,j} f_{ij}(x_i, x_j)$$

여기서 는 모두 여러겹의 은닉층을 갖는 인공신경망의 형태를 갖는다. 이렇게 모형을 구성하면 선형모형이 보여주는 일변수 및 이변수의 회귀효과를 비선형으로 추정하고 시각화하여 보여줄 수 있다.



GAM. 위는 한개 변수의 main effect (출력변수의 예측값이 y축에 나타남). 아래는 두개 변수의 interaction effect (여기서는 출력변수의 예측값이 Heatmap 형태의 색으로 나타남)

두 번째 방법론으로는 LIME(Locally Interpretable model-agnostic explanation) 또는 sensitiviry analysis의 방법론이 널리 이용된다. LIME은 복잡한 인공신경망 모형을 부분적으로(locally) 근사하는 선형모형을 찾아서 이를 인공신경망 모형의 설명에 이용한다. 이차함수의 변곡점을 설명하기 위해 접선을 구한 뒤 접선의 기울기가 0이니 입력변수가 변해도 출력변수는 변하지 않을 것이라고 설명하는 것과 비슷하다. 이 방법은 특정 입력변수 값에 대해서 변수의 중요도(feature importance)와 그 영향을 회귀계수와 비슷하게 설명할 수 있는 장점이 있지만, 모형 전체를 설명하지는 못하는 단점[4]이 있다. Sensitivity analysis도 비슷한 방법인데, 특정 입력변수의 값을 바꾸어서 결과가 어떻게 달라지는 지를 보는 것이다. 인공신경망 모형에 대해서 시나라오 테스트를 하는 것과 비슷하고 LIME과 비슷한 장단점을 갖는다.

필자는 인공신경망 모형을 큰 단위로 쪼개서 각각을 선형모형으로 근사한 뒤 이를 혼합모형(mixture model)으로 구성하는 혼합선형모형(mixture of linear model)을 연구하였다. 관심이 있다면 학위논문을 지식도서관에서 찾아볼 수 있다. 설명이 가능한 머신러닝 분야는 아직 새로운 분야로 많은 연구들이 다양한 관점에서 진행되고 있다. 인공신경망 모형의 점근적 통계분포(asymptotic distribution)를 찾아서 좀 더 엄밀한 통계학의 프레임 안에서 분석하려는 시도도 있고, 인공신경망 해석을 위해 별도의 인공신경망 모형을 구성하려는 시도도 있다. 어떤 실체를 해석하고 설명하는 방법이 사람마다 다르듯, 모형을 해석하려는 노력도 매우 다양한 방향에서 진행되고 있다.

[4] 모형 전체를 설명하기 위해서는 모든 입력변수에 대해서 부분근사모형을 구해야 하는데 이는 실현가능하지 않다(infeasible).

마지막으로

인공신경망 모형이 마치 모든 데이터 분석의 만병통치약이라는 것은 절대 아니다. 실제 패턴이 단순한 데이터의 경우 편리하고 신속한 선형모형을 두고 복잡한 모형을 이용할 이유는 없다. 다만 더욱 많은 데이터를 이용하여 다양한 예측을 하고자 하는 현대사회에서 인공신경망 모형은 패턴을 분석하기 위한 매우 매력적인 툴임에는 틀림없다. 현재 활발히 진행되고 있는 설명이 가능한 머신러닝 방법론을 고려하면 그동안 머신러닝 방법론에 대해 거부감이 컸던 금융, 의료, 행정적 의사결정 등 다양한 분야에서 그 활용성이 커질 것으로 기대한다.

사후적 설명모형을 이용하여 구한 변수 중요도(Feature Importance) 예시