

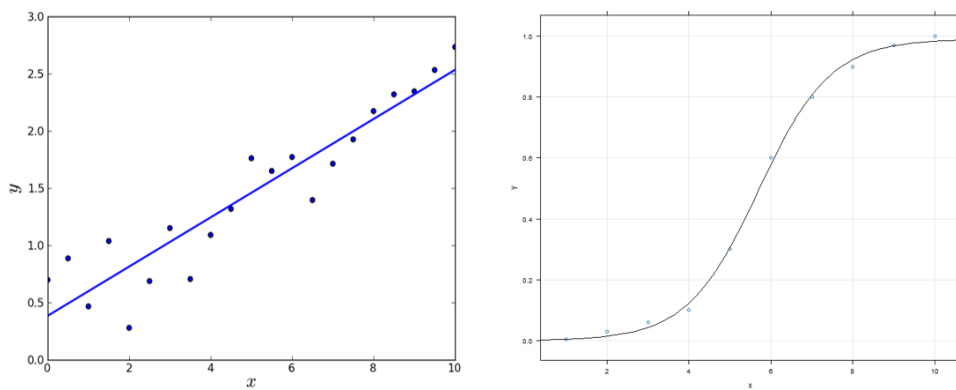
선형 회귀 모델에서 '선형'이 의미하는 것은 무엇인가?

by gimmesilver Nov 14, 2016

이전에 쓴 '회귀 모델에서 회귀란 무엇인가?'라는 글을 보고 어떤 분이 선형 관계에 대해 질문한 것을 보고 '선형'에 대해서도 설명이 필요하겠다고 생각하여 글을 썼습니다. 좀 더 자세한 글을 쓸 수 있다면 좋겠지만 제가 비선형 회귀 모델의 이론적인 배경에 대해서는 자세히 알지 못하기 때문에 이와 관련된 내용은 기초적인 내용만 정리했습니다.

1. '선형(linear)'은 x의 1차 다항식을 의미하는 것이 아니다.

인터넷에 돌아다니는 많은 자료들을 보면 선형 회귀 모델을 잘못 설명하고 있습니다. 이런 자료들을 보면 공통적으로 선형 회귀 모델의 의미를 독립 변수(x)와 종속 변수(y) 간의 관계가 일차식으로 표현되는 것이라고 설명하고 있습니다. 그래서 독립 변수를 제곱근이나 로그 함수 등을 이용해서 적당히 변환하면 비선형 모델을 선형 모델로 만들 수 있다고 설명합니다. 그러나 이것은 '선형'의 의미를 잘못 이해하고 있는 것입니다. 즉, 아래 그림에서 왼쪽 그림은 선형 모델이고 오른쪽 그림은 비선형 모델이라고 설명하는 경우가 많은데 잘못된 예시입니다(둘 다 선형 회귀 모델로 표현 가능합니다).



<그림 1> 선형 회귀 모델과 비선형 회귀 모델로 각각 검색했을때 나오는 대표적인 이미지 예

선형 회귀 모델은 '회귀 계수(regression coefficient)를 선형 결합으로 표현할 수 있는 모델'을 말합니다. 즉, 독립 변수가 일차식이나 이차식이나 로그 함수식이나가 아니라 우리가 추정할 대상인 파라미터가 어떻게 생겼느냐의 문제입니다. 가령 아래 함수들은 모두 선형 회귀식입니다.

- $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$
- $y = \beta_0 x^{\beta_1}$
- $y = \frac{e^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}}{1 + e^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}}$

생각하는 것이기 때문에 이 기준으로 보면 선형 회귀식입니다. 게다가 위 식들은 모두 '선형성'에 직접적으로 관련되지 않은 변수인 x 와 y 를 적절히 변환할 경우 모두 선형 회귀식으로 표현이 가능합니다.(아래 식에서 \log 는 모두 자연 로그)

- $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \Rightarrow \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
- $y = \beta_0 x^{\beta_1} \Rightarrow \log(y) = \log(\beta_0 x^{\beta_1}) \Rightarrow \log \beta_0 + \beta_1 \log(x) \Rightarrow y^* = \beta_0^* + \beta_1 x^*$
- $y = \frac{e^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}}{1 + e^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}} \Rightarrow \frac{y}{1-y} = e^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3} \Rightarrow \log\left(\frac{y}{1-y}\right) = y^* = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

비록 위와 같이 데이터 변환을 통해 선형 회귀식으로 표현가능한 회귀식을 좀 더 엄밀하게는 linearizable regression model(우리말로 하면 선형 가능 회귀 모델?)이라고 구분하여 부르기도 하지만 포괄적으로 봤을 때는 그냥 선형 회귀 모델이라고 부르는 것이 맞습니다. 왜냐하면 이런 linearizable regression model의 파라미터를 추정할 때도 일반적인 선형 회귀 모델과 동일한 기법을 사용하기 때문입니다.

2. 그럼 비선형 회귀 모델은 무엇인가?

비선형 모델은 데이터를 어떻게 변형하더라도 파라미터를 선형 결합식으로 표현할 수 없는 모델을 말합니다. 이런 비선형 모델 중 단순한 예로는 아래와 같은 것이 있습니다. 이 식은 아무리 x, y 변수를 변환하더라도 파라미터를 선형식으로 표현할 수 없습니다.

$$y = \frac{\beta_1 x}{\beta_2 + x}$$

선형 회귀 모델은 파라미터 계수에 대한 해석이 단순하지만 비선형 모델은 모델의 형태가 복잡할 경우 해석이 매우 어렵습니다. 그래서 보통 모델의 해석을 중시하는 통계 모델링에서는 비선형 회귀 모델을 잘 사용하지 않습니다.

그런데 만약 회귀 모델의 목적이 해석이 아니라 예측에 있다면 비선형 모델은 대단히 유연하기 때문에 복잡한 패턴을 갖는 데이터에 대해서도 모델링이 가능합니다. 그래서 충분히 많은 데이터를 갖고 있어서 variance error를 충분히 줄일 수 있고 예측 자체가 목적인 경우라면 비선형 모델은 사용할만한 도구입니다. 기계 학습 분야에서는 실제 이런 비선형 모델을 대단히 많이 사용하고 있는데 가장 대표적인 것이 소위 딥 러닝이라고 부르는 뉴럴 네트워크입니다.

3. 결론

정리하자면, 선형 회귀 모델은 파라미터가 선형식으로 표현되는 회귀 모델을 의미합니다. 그리고 이런 선형 회귀 모델은 파라미터를 추정하거나 모델을 해석하기가 비선형 모델에 비해 비교적 쉽기 때문에, 데이터를 적절히 변환하거나 도움이 되는 feature들을 추가하여 선형 모델을 만들 수 있다면 이렇게 하는 것이 적은 개수의 feature로 복잡한 비선형 모델을 만드는 것보다 여러 면에서 유리합니다.

반면 선형 모델은 표현 가능한 모델의 가짓수(파라미터의 개수가 아니라 파라미터의 결합 형태)가 한정되어 있기 때문에 유연성이 떨어집니다. 따라서 복잡한 패턴을 갖고 있는 데이터에 대해서는 정확한 모델링이 불가능한 경우가 있습니다. 그래서 최근에는 모델의 해석보다는 정확한 예측이 중요한 분야의 경우 뉴럴 네트워크와 같은 비선형 모델이 널리 사용되고 있습니다.

모델

선형

비선형

댓글 21



구독자 1,296

구독하기

집 그리고 우리 가족 & 40평대 아파트 인테리어

건강하고 행복하게 지내세요. 아버지 어머니 형 그리고 나. - 우리 집은 항상 돈에 골려다녔던 것 같았다. 물론 나는 집에서 막내라서 많은 혜택을 보면서 살았다고 생각...

by Sabum Byun

32평 오래된 아파트 인테리어, 온라인 집들이

이사 갈 집 인테리어를 앞두고 원하는 취향과 디자인은 확실히 있는데 예산의 압박으로 개별 인테리어(셀프 리모델링)를 하게 되었다. 다음 회에 자세한 금액은 얘기하겠...

by 파인쥬스

따로家치, 3대 6인 가족이 사는 집

광고 전원주택 5개월 차가 전하는 온라인 집들이 내 인생은 뇌종양 이전과 이후로 나뉜다. 대학 졸업하고 취업하고 결혼하고 아이 낳고. 대한민국 가장으로 평범하게 살...

by 오종현

인공지능 기술/원리의 이해

이 강의는 인공지능 기술의 '원리'를 이해함으로써 인공지능 기술 활용을 위한 기본 소양을 습득하는 것을 목표로 합니다. 다소 길지만 끝까지 완독하시면 현 인공지능 기...

by 송경빈

XGBoost 사용하기

지루하고, 재미없기 짝이 없지만 꾸준한 조희수를 보장할 것 같은 글 소개 시작은 캐글(kaggle)이었다. 캐글이 무엇인지 처음 읽는 분들을 위해서 잠깐 설명하자면, <캐...

by 소고

블록체인 보고서 총망라

2015년~2019년에 발행된 126개의 국내외 보고서 특정 주제에 대한 사회적 관심의 정도는 연구 기관의 보고서 수와 비례하는 게 일반적입니다. 블록체인이 시대적 화두...

by 카카오 정책산업 연구

월라 × *brunch*

오디오북출판