변수학습 선형회귀와 그 활용

Variable Learning Linear Regression and its Application

김 재 윤, 강 석 원, 송 왕 철, 이 상 준
제주대학교, 제주대학교, 제주대학교, 제주대학교

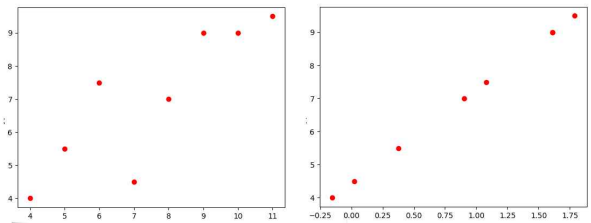
Jae-Yun Kim, Seok-Won Kang, Wang-Cheol Song, Sang-Joon Lee
Jeju National Univ., Jeju National Univ., Jeju National Univ., Jeju National Univ.,

요약

선형회귀는 여러 학문에서 변수들 사이의 관계를 설명하고자 할 때 쓰는 방법이다. 하지만 선형 회귀는 입력에 따른 출력이 무조건 선형 구조를 이룰 수밖에 없기 때문에 데이터셋이 곡선형을 이루거나 출력간의 차이가 지속적으로 증가하거나 감소하지 않는 경우 오차가 커질 수밖에 없다. 이를 보완하기 위해 변수를 학습시켜 선형 구조를 이루기 위해 변수가 어느 위치로 가야 하는지를 다시 학습시키는 방법을 사용하여 같은 데이터셋을 가지고 일반 선형회귀와 변수학습 선형회귀 간의 차이를 비교했고, cost가 줄어들며 데이터셋에 좀 더 근접한 모델이 나왔다.

I. 서론

Tensorflow에서 일반 선형회귀를 사용하면 각 특징들이 출력에 미치는 영향(Weight)과 보정값을 텐서변수로 선언해 학습시킨 출력과 예측된 출력의 차이를 나타내는 지표인 cost를 줄이는 방향으로 학습하게 된다. 이런 텐서변수의 특성을 활용해서 입력을 모두 텐서변수로 치환한 후 학습을 시켜 텐서변수가 선형이 되도록 변수의 위치를 찾아갈 수 있게 한다.



▶▶ 그림 1. 일반 데이터셋과 변수학습 후 데이터셋의 그래프

Ⅱ. 비교 분석

1. 비교 자료

두 알고리즘을 비교하기 위해 대학생의 수면시간 예측 모듈을 각각의 알고리즘으로 만들어 비교해보았다.

수면 데이터셋으로는 수면에 영향을 끼칠 수 있는 요인으로 연령, 직업이 있고 미묘한 차이를 보이는 성별이 있으나[1][2], 직업으로는 대학생으로 고정하고 연령을 1년 단위로 조사해 이를 통해 수면 시간을 예측하는 모델을 만들었다.

2. 자료 수집 방법

2017년 4월 4일부터 2017년 4월 20일까지 SNS를 통해 무작위 사람들의 연령, 성별을 입력 받고, 추가적으로 영향을 끼칠만한 요인으로 전날 수면을 취한 시간 및 당일날 잠에 든 시간을 선택했고, 알람 없이 일어난 날의 수면시간을 조사하였다.

3. 자료 분석

데이터셋 안에 있는 모든 입력을 텐서변수로 치환하고 각각의 weight와 곱하고 보정값을 더해 입력한 결과가 가능한 일치하도록 weight과 입력 변수를 조정한다.

$$E = \sum w(i) * x(i) \ + b \tag{1}$$

(1)식을 이용해 학습시킨 데이터를 이용해 특정 입력으로 예측을 하게 되면 입력이 데이터셋을 벗어났을 때 예측이 불가능한 문제가 나타나기 때문에 이를 보완하기 위해서 데이터 셋 안의 모든 입력을 정렬된 배열로 만들고 입력이 배열의 두 원소 사이에 있다면

$$X = \alpha + (\beta - \alpha) * (x_{value} - array(x))/(array(x) - (array(x+1)) \tag{2}$$

(2)식을 적용해 X을 weight에 곱하는 것으로 예측값을 낼 수 있다.

Ⅲ. 결과

변수학습 선형회귀를 사용했을 때 cost가 1.95000에서 1.41566으로 떨어지는 결과가 나왔고 아래 표들은 22세 남학생에 전 날 24시 잠에 들어 7시간동안 잔 것을 기준으로 입력을 하나씩 변화시켰을 때 잠잘 시간을 예측한 결과다.

표 1. 잠에 든 시간에 따른 수면시간 비교(단위 : 시간)

	일반선형회귀	변수학습 선형회귀
22시	7.720	7.411
23시	7.716	7.376
24시	7.713	7.762
1시	7.709	8.036
2시	7.705	6.845
3시	7.702	6.215
4시	7.698	7.664

표 2. 나이에 따른 수면시간 비교(단위 : 시간)

	일반선형회귀	변수학습 선형회귀
20세	8.097	8.086
21세	7.905	7.532
22세	7.713	7.762
23세	7.520	7.693
24세	7.328	7.413

표 3. 전날 수면시간에 따른 수면시간 비교(단위 : 시간)

	일반선형회귀	변수학습 선형회귀
4시간	9.979	11.497
5시간	9.224	9.662
6시간	8.468	8.607
7시간	7.713	7.762
8시간	6.957	7.764
9시간	6.202	6.720
10시간	5.446	5.349

표 4. 성별에 따른 수면시간 비교(단위 : 시간)

	일반선형회귀	변수학습 선형회귀
남성	7.713	7.762
여성	7.775	7.770

Ⅳ. 결론

변수학습 선형회귀 데이터셋의 형태가 직선이 아닌 데이터들을 타겟으로 학습했을 때 일반 선형회귀보다 더 낮은 cost와 입력값의 차이가 출력값과의 차이가 일정한 비율을 가지는 일반 선형회귀와는 다르게 신뢰할만한 예측값을 보여주었다. 하지만 표 1에서 변수학습 선형회귀는 결과가 데이터셋에 치중되어있어 자는 시간대는 연관이 없다는 결론을 도출해낼 수 있는 일반 선형회귀와는 달리 아무런 정보를 얻을 수 없었다는 한계가 있다. 두 모듈 간의 이식성이 아주 좋기 때문에 일반선형회귀로 전체적인 증감을 파악하고 변수학습 알고리즘으로 현실에 더 가가까운 예측을 얻는 것을 권장한다.

■ 참 고 문 헌 ■

[1] 박형진, 이현자 “한국인의 수면시간과 수면자세”, 한국생활과학회 학술대회 논문지, pp.169-170, 2012.

[2] 은기수, 차승은, “한국의 일상생활에서 남녀의 수면시간과 관련 요인 탐색,” 통계연구 제 15권 제2호, pp. 82-103, 2010.

영문 : “This research(or This presentation or This work) was financially supported by The Project Management Center Cultivating Smart Grid & Clean Energy Manpowers(CK-1), JNU”

국문 : “이 연구(이 발표, 이 작품 등)는 제주대학교 스마트그리드와 청정에너지 융복합산업 인력양성사업단(CK-1)의 지원을 받아 수행되었습니다.”