Journal of Knowledge Information Technology and Systems(JKITS), Vol. 15, No. 4, pp. 497-506, August 2020
https://doi.org/10.34163/jkits.2020.15.4.006

Journal of Knowledge Information Technology and Systems

ISSN 1975-7700

http://www.kkits.or.kr

High-School Baseball Pitcher's ERA(Earned Run Average) Prediction Using Multi-Variable Linear Regression Analysis Method

Young-Hwan Oh*

Division of IT Convergency, Korea Nazarene University

ABSTRACT

Recently, studies on artificial intelligence has been actively conducted, and scholars and researchers are constantly trying to create artificial intelligence beyond humans. Machine learning uses mathematics' logical algorithms to learn, analyze and predict various data. It assists decision making based on this analysis. The goal is not to directly write algorithms optimized for software or programs, but to create execution capabilities by letting IT devices or electronic devices themselves learn using large-scale big data and learning algorithms. Linear regression is a regression analysis method that models the linear correlation between one or more independent variables(x) and a dependent variable(y). Generally, a linear regression model is established using a least squares method. Multiple linear regression analysis targets regression models with two or more independent variables. Simple regression analysis can bring bias by estimating with only one independent variable. A baseball player's pitching ability depends on his speed, ball and concentration. In particular, among pitchers' stats, WHIP, H/9(Hits Runs Allowed Per 9 Innings Pitched) and K/9(Strikeout Per 9 Innings Pitched) are important measures of pitching. In this paper, the ERA(Earned Run Average) is predicted using the pitcher's WHIP, H/9 and K/9 as training data. We can predict whether a specific pitcher will be nominated for a professional baseball player or not.

KEYWORDS: Artificial intelligence, Machine-learning, Multi-variable linear regression, Gradient descent algorithm, Tensorflow

ARTICLE INFO: Received 15 June 2020, Revised 9 July 2020, Accepted 10 August 2020.

*Corresponding author is with the Division of IT Convergency, Korea Nazarene University, 456 Wolbong-ro Seobuk-gu Cheonan, 31172, KOREA.
E-mail address: yhoh@kornu.ac.kr

1. 서 론

인공지능(Artificial intelligence)은 인간과 유사한 지능을 갖고 있는 기능을 갖춘 컴퓨팅 시스템이며, 인간의 유사 지능을 기계와 가전 등에 인공적으로 설계와 구현을 한 것이다[1]. 지난 수십 년 동안 인공 지능에 관한 연구와 개발이 활발히 진행 중이고, 학자와 연구원들은 사람을 넘어서는 인공 지능 시스템을 제작하기 위해 끊임없이 노력 중이다.

머신 러닝(Machine learning)은 수학의 논리적인 알고리즘을 이용해 데이터를 학습, 분석하고 예측한다. 이 분석한 내용을 기반으로 의사 결정을 도와준다. 최종적으로는 소프트웨어나 프로그램에 최적화된 알고리즘을 직접 기술해 넣는 것이 아니라 대용량의 빅데이터(Big data)와 학습 알고리즘을 이용하여 IT기기나 전자장치 그 자체가 학습하도록 하여 실행 능력을 만드는 것을 목표로 한다[2]. 최근에는 인간의 뇌가 가진 인지적이고 생물학적 특징, 뉴런 구조를 바탕으로 하는 인공 신경망(Artificial neural network) 연구에 상당한 노력을 하고 있다. 딥 러닝(Deep learning)은 사람의 뇌의 뉴런과 비슷한 정보 I/O 알고리즘을 가져와 방대한 양의 데이터를 저장하고 학습하여 기존의 인공 신경망 시스템에서 발전한 인공 지능 기술이다[3]. 이제 딥 러닝은 현실 세계를 감지하고 분석하는 인공 지능을 개발하는 데 가장 인기 있는 기술로 거듭나고 있다[4].

야구 경기는 투수(Pitcher)와 포수(Catcher)가 각각 야구공을 송구하고 포구하는 것으로부터 시작되며 이 경우가 야구 경기에서 차지하는 비중이 가장 크다[5]. 특히, 투수는 훌륭한 재목으로 성장하기 위해서 투구 스피드(Ball speed)와 제구(Ball control) 그리고 투구 밸런스(Pitching balance)가 좋아야 한다. 야구의 다양한 통계값 중에서 투수와 연관된 스탯(STAT)은 다음과 같다[6].

- ERA(Earned Run Average, 평균자책점) : 투수가 9이닝당 내준 평균 자책점(Earned Runs)
- WHIP(Walks plus Hits divided by Innings Pitched, 안타+볼넷 허용률) : 투수가 1이닝당 허용한 안타수와 볼넷수
- K/9(Strikeout Per 9 Innings Pitched, 탈삼진율) : 투수가 9이닝당 잡은 삼진 개수
- H/9(Hits Runs Allowed Per 9 Innings Pitched, 피안타율) : 투수가 9이닝 당 허용한 피안타 개수
- BB/9(Base on Balls Per 9 Innings Pitched) : 투수가 9이닝 당 볼넷 개수
- K/BB(탈삼진/볼넷) : 삼진 개수를 4구 개수로 나눈 값, 2.0 이상이면 좋은 투수로 판단하고 3.0 이상이면 최고급 투수로 본다.

투수의 투구스피드가 약 140km인 경우 투수판으로부터 홈플레이트까지의 시간은 약 0.4~0.5초이다[7]. 재능이 있고 인기있는 투수로 성장하기 위한 중요한 요소는 투구의 스피드이지만, 투수의 다양한 STAT에서도 뛰어 나야된다. 고교야구 투수의 통계 값은 1차와 2차 드래프트(Draft)에서 고교야구 선수가 프로야구로 지명되는 중요한 지표이다. 또한 프로야구 선수가 되는 객관적인 데이터이다. 다음 <그림 1>은 대한야구소프트볼협회(Korea Baseball Softball Association)에서 발췌한 고교야구 투수들의 각종 데이터의 예이다.

이름	팀명	평균자책점	팀경기	경기수	승	패	승률	타자	타수
이 호(33)	북 고등학교	3.60	6	3	0	0	0.000	22	18
김 진(21)	대 고등학교	4.50	6	5	1	1	0.500	77	62
신 후(1)	북 고등학교	1.29	6	4	1	1	0.500	35	25
이 희(28)	대 고등학교	6.75	6	3	0	0	0.000	18	18
안 건(21)	광 고등학교	9.00	6	3	0	1	0.000	18	17
박 영(50)	세 고등학교	12.60	6	2	0	1	0.000	24	16
곽 철(19)	광 고등학교	5.73	6	4	0	1	0.000	56	43
조 근(19)	북 고등학교	0.00	6	2	0	0	0.000	6	5
홍 기(1)	대 고등학교	3.00	6	4	0	0	0.000	16	12
김 범(37)	대 고등학교	7.43	6	5	1	3	0.250	118	101
김 우(42)	북 고등학교	9.00	6	1	0	0	0.000	5	3
백 진(29)	북 고등학교	18.00	6	2	0	0	0.000	11	6

탈삼진	폭투	보크	실점	자책점	WHIP	피안타율
1	0	0	2	2	1.60	0.278
10	1	0	11	8	1.69	0.290
8	3	0	4	1	1.71	0.320
7	0	0	3	3	1.75	0.389
4	0	0	4	4	1.75	0.353
3	0	0	7	7	1.80	0.313
5	0	0	12	7	1.91	0.349
1	1	0	0	0	2.00	0.200
3	0	0	2	1	2.00	0.333
13	4	1	26	19	2.00	0.347
2	0	0	1	1	2.00	0.333
0	1	0	4	4	2.00	0.000

그림 1. 고교야구 투수 데이터
Figure 1. Pitcher data of High-school baseball

2020년 현재 대한야구소프트볼협회에 등록된 고교야구는 81개 팀이 있으며 황금사자기, 청룡기와 같은 전국대회 왕중왕전과 고교야구 전, 후반기 주말리그 등 많은 경기를 소화하고 있다. 고교야구 주말리그제는 KBSA(Korea Baseball-Softball Association)가 2010년대 초반부터 시행한 제도로서 공부도 하는 우수 엘리트 학생 스포츠라는 생각을 가지고, 전국을 크게 10개 정도의 권역으로 나눠 7-8개 팀을 배정하여 주중을 제외한 주말리그로 열리고 있다[8]. 고교야구 선수는 KBO(Korea Baseball Organization) 프로선수들과 체력, 체격, 기술적인 면에서 커다란 간격이 있기 때문에 이를 줄이기 위한 여러 노력을 하고 있다. 고교 야구 선수가 프로야구 선수가 되는 지명제도(드래프트)는 엘리트 야구 선수로 성장하는 중요한 기회가 된다. 통상적으로 매년 6월과 8월에 각각 1차 드래프트와 2차 드래프트가 개최된다. 이 때 10개의 프로야구 팀은 고교야구 선수의 체격, 성장가능성 등을 주요 골자로 하여 지명을 하지만 2~3년 동안 그들이 이루어 놓은 각종 지표(야구 스탯)을 무시할 수가 없다. 이는 대한야구소프트볼협회 데이터베이스 서버에 전부 저장되어 각 프로팀의 스카우터와 다양한 사용자에게 제공된다.

좋은 투수로 성장하기 위해서는 학생선수가 던질 수 있는 구종에 대한 구속, 제구 그리고 투구 밸런스가 중요하다. 이를 통하여 야구 경기에서 좋은 성적을 낼 수 있다. 투수의 평균자책점, WHIP, 피안타율, 탈삼진율, 방어율은 대표적인 투수의 척도지표이다[9]. 이를 통하여 훌륭한 프로야구 선수가 될 수 있는 지를 가늠하여 볼 수 있다. 또한 프로 스카우터가 이를 바탕으로 드래프트를 통하여 프로야구 신인선수로 발탁한다. 투수의 WHIP, 피안타율, 탈삼진율, 방어율은 서로 연관되어 있으며, 본 논문에서를 이를 학습데이터로 선정하여 다른 투수의 성적을 예측할 수 있다.

본 논문에서는 텐서플로우(Tensorflow)을 이용하여 고등학교 투수의 방어율을 예측하기 위한 다중 선형회귀분석기법을 연구한다. 이를 위하여 투수의 WHIP, 피안타율, 탈삼진율을 학습데이터로 이용하며 예측 기법을 구현하고 성능을 평가한다.

본 논문의 구성은 다음과 같다. 제 2장에서는 관련 연구와 제 3장에서는 학습데이터, 그리고 제 4장에서는 예측 모델을 기술한다. 제 5장에서는 구현과 실험 결과를 논의한다. 마지막으로 제 6장에서 결론을 기술한다.

2. 관련 연구

2.1 다중 선형회귀 분석

선형회귀 분석(Linear Regression Analysis)을 위해서는 우선 선형회귀 모델(Linear Regression Model)을 작성해야 한다. 이 모델은 수학식으로 표현된 함수를 의미하며, 하나 이상의 독립 변수(Independent Variable)와 독립변수(Dependent Variable)와의 상관관계를 의미한다[10]. 선형회귀 분석은 종속변수 y, 독립변수 x, 상수항 b 사이의 관계를 모델링하는 방법이다. 이 때 여러 개의 독립 변수를 다루는 경우를 다중 선형회귀(Multiple Linear Regression)이라고 한다[11].

2.2 텐서플로우

텐서플로우는 2011년에 구글(Google)이 연구개발을 시작하여 오픈 소스(Open source)로 공개한 ML 라이브러리로 기계 학습과 딥 러닝 분야를 전문가와 일반인들이 이용하기 쉽도록 다양한 기능들을 제공한다[12]. 최근에 알파고(Alphago)와 함께 전 세계적으로 관심이 높아진 추세이다. 또한 높은 수준의 프로그래밍 언어로 알려진 파이썬(Python)을 활용하여 연산프로세싱을 처리할 수 있다[13].

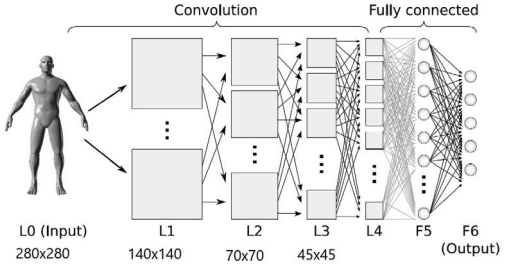

그림 2. 텐서플로우
Figure 2. Tensorflow

3. 학습데이터

고교야구 선수나 프로야구 선수 중 투수 능력은 직구 스피드(Straight speed), 제구(Ball control), 투구 밸런스(Pitching balance)가 가장 중요하다. 특히 투수의 스탯(STAT) 중 WHIP, 피안타율, 탈삼진율은 투수의 중요한 척도가 된다. 이 밖에 평균자책점, K/BB 등 다양하다. 본 연구에서는 투수의 WHIP, 피안타율, 탈삼진율을 학습데이터(Training data)로 하여 방어율을 예측한다. 이를 토대로 하여 특정 투수가 프로에 지명되는 지를 예측할 수 있다[14].

본 장에서는 고교야구 투수의 WHIP, 피안타율, 탈삼진율을 가지고 텐서플로우를 이용하여 예측모델을 산정한다. 이를 위하여 먼저 학습데이터를 결정하고 최소제곱법(Least squares method)을 이용하여 학습률(Learning rate)을 결정한다[15].

3.1 학습데이터

본 실험을 위하여 대한야구소프트볼협회 홈페이지 자료실에서 2019년 고등학교 야구에서 투수로써 경기에 임했던 KBO리그에 프로로 선발된 투수 9명을 임의로 선택하여 표본 자료를 샘플링하였다.

학습은 지도 학습(Supervised learning)과 비지도 학습(Unsupervised learning)으로 나뉜다. 지도 학습은 미리 정해진 데이터를 이용하여 학습을 한다. 이 때 기초가 되는 데이터를 학습 데이터라고 한다[16]. 본 장에서는 종속변수 x1, x2, x3로부터 독립변수 y를 예측 한다. 다음 <표 1>는 독립 변수와 종속 변수를 보여 준다.

표 1. 독립변수와 종속변수
Table 1. Independent and dependent variables

x1(WHIP)	x2(피안타율)	x3(탈삼진율)	y(방어율)
1.00	0.229	10.31	2.63
0.98	0.209	11.25	3.54
1.14	0.213	6.75	3.54
1.53	0.159	10.08	3.20
1.11	0.181	11.06	2.57
1.47	0.256	12.56	3.35
1.40	0.156	10.8	2.10
1.64	0.179	9.82	4.64
1.56	0.197	7.88	2.81

여기에서 x1 데이터는 WHIP, x2 데이터는 피안타율, x3 데이터는 탈삼진율 그리고 y 데이터는 방어율이다.

3.2 산점도

텐서플로우의 matpotlib을 이용하면 학습데이터의 산점도(scatter plot)를 표현할 수 있다. 아래 <그림 3>은 학습데이터에 대한 데이터 프레임 작성에 대한 것이다. 아래 코드를 실행하면 아래와 같이 산점도 그래프가 그려진다. 코드의 내용은 다음과 같다. 먼저 Matplotlib를 import한다. Matplotlib는 파이썬(Python)에서 그래프를 그릴 수 있는 라이브러리다. plt.plot()를 이용하여 scatter 그래프를 그릴 것을 지정한다. plt.plot(x축 데이터, y축 데이터)가 기본 문법이다.

```
# 데이터(y=방어율, x1=WHIP, x2=피안타율, x3=탈삼진율)
y_data = [2.63, 3.54, 3.54, 3.20, 2.57, 3.35, 2.10, 4.64, 2.81]
x1_data = [1.00, 0.98, 1.14, 1.53, 1.11, 1.47, 1.40, 1.64, 1.56 ]
x2_data = [0.229, 0.209, 0.213, 0.159, 0.181, 0.256, 0.156, 0.179, 0.197]
x3_data = [10.31, 11.25, 6.75, 10.08, 11.06, 12.56, 10.8, 9.82, 7.88]

import matplotlib.pyplot as plt
import pandas as pd

df = pd.DataFrame(vectors_set)

pd.tools.plotting.scatter_matrix(df)
plt.tight_layout()
plt.show()

plt.plot(x1_data, y_data, 'ro')
plt.show()

plt.plot(x2_data, y_data, 'b+')
plt.show()

plt.plot(x3_data, y_data, 'cs')
plt.show()
```

그림 3. 학습데이터와 데이터프레임

Figure 3. Training data and Data frame

위의 <표 1>를 산점도 그래프로 나타내면 다음 <그림 4>부터 <그림 6>까지이다.

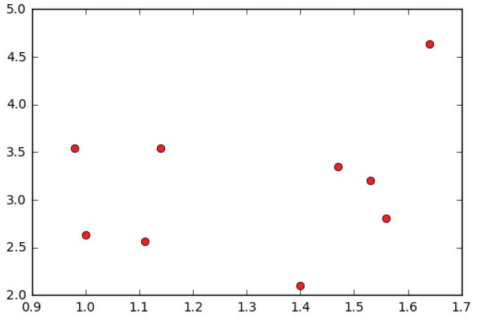

그림 4. x1(WHIP)와 y(방어율)

Figure 4. x1(WHIP) and y(ERA)

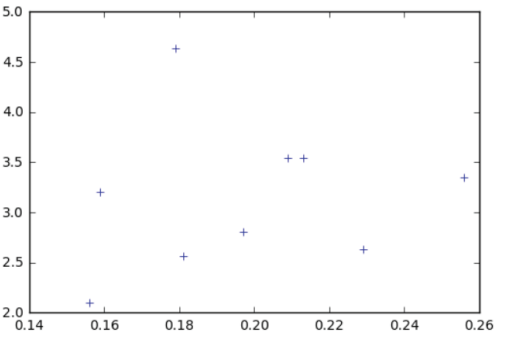

그림 5. x2(피안타율)와 y(방어율)

Figure 5. x2(AVG) and y(ERA)

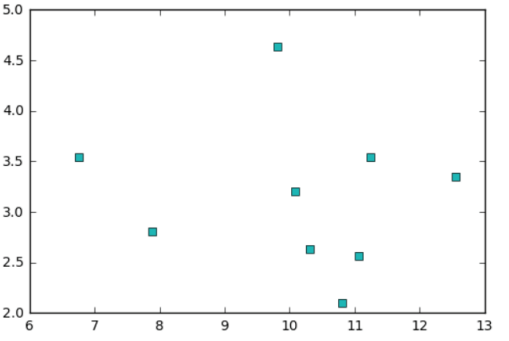

그림 6. x3(탈삼진율)와 y(방어율)

Figure 6. x3(K/9) and y(ERA)

4. 다중 선형회귀분석

본 장에서는 다중 선형회귀분석에 대한 가설 방정식과 최소 제곱법 그리고, 경사 하강법을 설명한다.

4.1 가설 방정식

다중 선형회귀분석은 두 개 이상의 독립 변수들과 하나의 종속변수의 상관관계를 분석하는 기법으로 단순 선형회귀의 확장이다. 3.1절에서 언급한 WHIP, 피안타율, 탈삼진율, 방어율 학습데이터를 가지고 학습을 진행하면 여러 개의 독립변수로 결과값을 예측하기 위한 가설을 매트릭스 곱셈 연산으로 표현할 수 있다.

$$H(X) = XW$$

$$(x_1 \quad x_2 \quad x_3) \cdot \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = (x_1 w_1 + x_2 w_2 + x_3 w_3)$$

즉, 단순 선형회귀분석에서 이용되었던 가설에 x 값을 n개 추가하여 위의 식으로 가설을 만들 수 있다. 본 논문에서는 독립 변수 x의 개수가 3개이므로 이를 수식으로 표현하면 아래와 같다.

$$H(x_1, x_2, x_3) = w_1 x_1 + w_2 x_2 + w_3 x_3 + b$$

그리고 독립 변수가 추가되어 변경된 가설을 이용하면 비용 함수(cost function)식은 가설 부분만 변경하면 된다. 이러한 학습 데이터를 바탕으로 비용함수를 만들어 낸다면 투수의 WHIP, 피안타율, 탈삼진율을 가지고 투수의 방어율을 예측할 수 있다.

$$cost(W,b) = \frac{1}{m}\sum_{I=1}^{m}(H(x_1^{(i)}, x_2^{(i)}, x_3^{(i)}) - y^{(i)})^2$$

4.2 최소 제곱법

본 연구에서는 가장 적합한 가설을 구하기 위하여 여러 개의 가설 중에서 최소 제곱법을 이용한다. 즉, 가설과 학습데이터 사이에 오차가 작다면 적합한 가설이 된다.

H(x) - y가 실제 데이터와 가설 데이터 간의 차이 값이 된다. 이 데이터는 양수 또는 음수값이 둘 다 나오기 때문에 아래와 같이 제곱을 해준다.

$$(H(x) - y)^2$$

m개의 학습데이터가 있을 때, 여러 개의 학습데이터에서 그려진 직선은 예측 값(H(x)) 과 학습데이터 값(y)의 차이를 제곱한 값의 평균이다.

$$cost(W,b) = \frac{1}{m}\sum_{i=1}^{m}(H(x^{(i)}) - y^{(i)})^2$$

즉, 선형회귀는 여러 개의 학습데이터가 주어졌을 때, 비용 값을 제일 적게 하는 가설의 매개변수(W, b)를 찾는 알고리즘이다. 즉, 비용을 최소화하는 W와 b의 해를 구하는 것이 선형회귀 분석의 목적이 된다. 이 경우 일반적으로 경사 하강 알고리즘(Gradient descent algorithm)을 이용하여 최저값을 구한다.

4.3 경사 하강법

경사 하강법은 함수의 기울기를 이용하여 그래프 값의 차이값을 비교한 후 제일 적은 방향으로 이동시키는 방법이다.

$$cost(W) = \frac{1}{2m}\sum_{i=1}^{m}(Wx^{(i)} - y^{(i)})^2$$

$$W := W - \alpha\frac{\partial}{\partial W}cost(W)$$

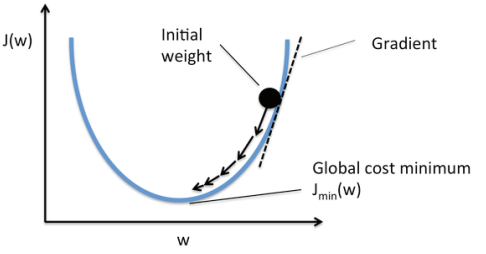

그림 7. 경사 하강법

Figure 7. Gradient descent method

즉, 미분은 함수의 순간 변화율을 뜻하며 한 점에서의 기울기를 의미한다. 임의의 점에서 미분값을 구하면 그래프(가설)의 기울기를 구할 수 있다. 경사 하강법에서 미분 값이 제로값을 찾는 경우

기울기가 양수와 음수를 번갈아 반복됨으로 기울기의 부호를 변경할 때 이동거리를 정해주는 것이 중요하다. 이를 학습률이라고 한다.

5. 구현 및 실험결과

5.1 텐서플로우를 이용한 다중선형회귀분석

본 논문에서는 텐서플로우를 이용하여 데이터셋 형태의 다중선형회귀분석 모델을 작성한다[17]. 먼저 학습 데이터를 가지고 오고 여기에 맞는 최적의 직선을 텐서플로우를 이용하여 실행한다. 행렬을 이용하지 않을 때는 데이터 셋을 각각 한 개씩 만들어 초기화 한다. 변수 또한 데이터 셋에 맞춰 w1, w2, w3 그리고 x1, x2, x3 각각 선언해 주어야 한다.

<그림 8>에서는 투수의 WHIP, 피안타율, 탈삼진율을 학습 데이터로 이용하여 데이터 셋으로 지정한다. 그 다음에 w1, w2, w3과 b을 지정하여 학습을 위해 랜덤 값을 할당한다. 텐서플로우는 가중치와 바이어스에 해당하는 변수를 이용한다. 또한 그 값들이 자동으로 조절하게 하며, 비용함수가 최소화가 되도록 가중치와 바이어스를 조정한다.

```
import tensorflow as tf

#학습데이터(x1=WHIP, x2=피안타율, x3=탈삼진율, y=방어율)
x1_data = [1.00, 0.98, 1.14, 1.53, 1.11, 1.47, 1.40, 1.64, 1.56 ]
x2_data = [0.229, 0.209, 0.213, 0.159, 0.181, 0.256, 0.156, 0.179, 0.197]
x3_data = [10.31, 11.25, 6.75, 10.08, 11.06, 12.56, 10.8, 9.82, 7.88]
y_data = [2.63, 3.54, 3.54, 3.20, 2.57, 3.35, 2.10, 4.64, 2.81]

# placeholders for a tensor that will be always fed.
x1 = tf.placeholder(tf.float32)
x2 = tf.placeholder(tf.float32)
x3 = tf.placeholder(tf.float32)
Y = tf.placeholder(tf.float32)

# tf.Variable were assigned x1_data, x2_data, x3_data and y_data respectiely
w1 = tf.Variable(tf.random_normal([1]), name='weight1')
w2 = tf.Variable(tf.random_normal([1]), name='weight2')
w3 = tf.Variable(tf.random_normal([1]), name='weight3')
b = tf.Variable(tf.random_normal([1]), name='bias')
hypothesis = x1 * w1 + x2 * w2 + x3 * w3 + b
```

그림 8. 학습데이터 정의
Figure 8. Definition of training data

다음 <그림 9>은 비용 함수를 구하기 위하여 가설 데이터와 학습 데이터의 절대값의 제곱의 합을 구해 데이터의 개수만큼 나누도록 한다. 이를 위하여 텐서플로우의 tf.reduce_mean(tf.square(...)) 함수를 이용한다. 4.3에서 기술한 바와 같이 텐서플로우는 비용값을 최소화하기 위해 최적화 알고리즘인 경사 하강법을 활용한다. 여기에서는 학습도를 1e-5로 하여 경사 하강법 알고리즘을 실행한다.

```
# cost/loss function
cost = tf.reduce_mean(tf.square(hypothesis - Y))
# Minimize. Need a very small learning rate for this data set
optimizer = tf.train.GradientDescentOptimizer(learning_rate=1e-5)
train = optimizer.minimize(cost)
```

그림 9. 가설과 비용함수
Figure 9. Hypothesis and cost function

다음 <그림 10>는 텐서플로우의 전역변수값을 초기화하기 위하여 global_variables_initializer() 함수를 이용한다. 본 연구에서는 최적화를 위하여 2000번을 수행하여 학습을 진행하였다.

```
# Launch the graph in a session.
sess = tf.Session()
# Initializes global variables in the graph.
sess.run(tf.global_variables_initializer())
for step in range(2001):
    cost_val, hy_val, _ = sess.run([cost, hypothesis, train],
                                   feed_dict={x1: x1_data, x2: x2_data, x3: x3_data, Y: y_data})
    if step % 100 == 0:
        print(step, "Cost: ", cost_val, "\nPrediction:\n", hy_val)

# 최적화가 완료된 모델에 결과 값을 넣고 결과가 잘 나오는지 확인한다.
print("Your score will be", sess.run(hypothesis,feed_dict={x1: 0.78, x2: 0.134, x3: 13.56}))
print("Your score will be", sess.run(hypothesis,feed_dict={x1: 2.34, x2: 0.286, x3: 7.45}))
```

그림 10. 세션 생성과 최적화
Figure 10. Session creation and optimization

5.2 실행 결과

선형회귀분석을 위한 학습을 실행하면 cost의 값이 점점 최소화되고 가설함수 h = w1*x1 + w2*x2 + w3*x3 + b를 계산하여 실 데이터값과 학습으로 얻어진 데이터 값을 비교하게 된다. w1, w2, w3 값과 b의 값에 랜덤 값을 주게 되고, 학습모델로 봤을 때 이 값들이 수렴하는지 지켜보면 된다. 그 결

과 값으로 각 열은 트레이닝 횟수, cost, w1, w2, w3, b의 값을 나타낸다. 다음 <그림 11>은 실행 결과를 나타낸다.

```
0 Cost:  64.51006
Prediction:
 [-5.473085  -5.9804792 -3.4193273 -4.4501004 -5.6585407 -5.841034
 -5.0339613 -4.138814  -3.2883475]
100 Cost:  42.35315
Prediction:
 [-3.9264994 -4.296399  -2.3925223 -2.9275188 -3.999981  -3.9554076
 -3.4081812 -2.6522613 -2.087902 ]
200 Cost:  27.894917
Prediction:
 [-2.6771774 -2.9360156 -1.5630544 -1.6975746 -2.6602085 -2.432212
 -2.094881  -1.4514172 -1.1181651]
300 Cost:  18.460356
Prediction:
 [-1.667984   -1.8371129  -0.89299476 -0.70402277 -1.5779516  -1.2017872
 -1.034001   -0.48136806 -0.33479416]
.........
1800 Cost:  0.77094865
Prediction:
 [2.400554  2.5927722 1.809556  3.302139  2.7850277 3.758557  3.2432182
 3.4302633 2.8247368]
1900 Cost:  0.76071334
Prediction:
 [2.4336967 2.6288338 1.8316694 3.3348284 2.8205605 3.7989588 3.2780857
 3.4622002 2.850586 ]
2000 Cost:  0.7540205
Prediction:
 [2.4604645 2.6579533 1.8495499 3.3612413 2.849256  3.8315868 3.3062515
 3.4880095 2.8714871]
Your score will be [2.8829074]
Your score will be [4.126597]
```

그림 11. 실험 결과
Figure 11. Experiment results

학습 데이터를 이용한 다중 선형회귀분석을 통하여 한 투수의 x1(WHIP)=0.78, x2(피안타율)=0.134, x3(탈삼진율)=13.56 인 경우, y(방어율)은 2.8829074 이고 또 다른 투수의 x1(WHIP)=2.34, x2(피안타율)=0.286, x3(탈삼진율)=7.45인 경우 y(방어율)은 4.126579로 예측할 수 있다. 이를 토대로 하여 특정 투수의 우수성을 볼 수 있다.

6. 결 론

본 논문에서는 고등학교 투수의 방어율을 예측하기 위한 다중선형회귀 분석기법을 연구하였다. 이를 위하여 딥 러닝 프레임워크인 텐서플로우를 이용하며 다중선형회귀 예측 모델을 설계와 구현하고 성능을 평가하였다. 투수의 WHIP, 피안타율과 탈삼진율은 프로야구의 지명을 위한 중요한 요소들이다. 또한 투수의 방어율은 투수의 STAT 중 여러 통계값과 밀접한 관계가 있다. 본 연구에서는 투수의 WHIP, 피안타율과 탈삼진율 그리고 방어율에 대한 학습데이터를 생성하고 최적화 알고리즘인 경사 하강법을 이용하여 다중 선형회귀분석을 수행하였다. 이를 통하여 고교야구 선수에 대한 평가와 지명(드래프트)에 대해서 도움을 준다. 본 연구의 한계점으로는 경사 하강법에서 학습률에 대한 기준점을 찾기가 어렵다는 사실이다. 학습률이 낮으면 천천이 가는 대신 최소값 근처에서 안정작으로 수렴할 수 있지만, 학습률이 높은 경우 최소값 근처에서 발산하는 현상(Oscillation)이 나타날 수 있어 어려움이 있었다. 추후 연구로는 고교야구 투수의 성적(방어율, 피안타율, 삼진율)에 따른 프로야구 지명 예측에 대한 로지스틱 회귀(Logistic regression) 분석을 계속할 예정이다.

References

[1] G. Simari, and I. Rahwan, *Argumentation in artificial intelligence*, Springer Publishing Company, Jan. 2009.

[2] J. Nilsson, *Introduction to machine learning*, Robotics Laboratory, Department of Computer Science, Stanford University, CA, 1998.

[3] Neural Networks and Deep Learning, M. Nielsen, http://neuralnetworksanddeeplearning.com/, May 2016.

[4] G. Huang, Z. Liu, L. Maaten, and K. Weinberger, *Densely connected convolutional networks*, In Proceeding of IEEE Conf. Computer Vision Pattern Recognition, pp. 4700-4708, Honolulu, USA, Jul. 2017.

[5] Y-J. Lee, and J-T. Kim, *The kinematic*

analysis of the pitching motion for the straight and curve ball, Journal of The Korean Society of Sports Biomechanics, Vol. 12, No. 2, pp. 109-130, 2002.

[6] Baseball Reference, https://www.baseball-reference.com/, Apr. 2020.

[7] R. Whiteley, *Baseball throwing mechanics as they relate to pathology and performance – A review*, Journal of Sports Science and Medicine, Vol. 6, pp. 1-20, 2007.

[8] Korea Baseball Softball Association, http://www.korea-baseball.com/, June. 2019.

[9] FanGraphs Baseball | Baseball Statistics and Analysis, http://https://www.fangraphs.com/, May. 2018.

[10] C. J. Peng, K. L. Lee, and G. M. Ingersoll, *An introduction to logistic regression analysis and reporting*, The Journal of Educational Research, Vol. 96, No. 1. pp. 1-15, 2002.

[11] Simple and Multiple Linear Regression in Python, https://towardsdatascience.com/simple-and-multiple-linear-regression-in-python-c928425168f9, May. 2017.

[12] E. S. Han, *A Study on the application of tensorflow to determine the correctional distance,* Journal of Knowledge Information Technology and Systems, Vol. 15, No. 3. pp. 323-329, 2020.

[13] Tesorflow Tutorials, https://gist.github.com/haje01/202ac276bace4b25dd3f, Jun. 2018.

[14] Y. H. OH, *High-school baseball Pitcher's pitching speed prediction using linear regression analysis method*, Journal of Knowledge Information Technology and Systems, Vol. 14, No. 4. pp. 381-390, 2019.

[15] S. J. Miller, *The method of least squares*, Mathematics Department, Brown University, Rhode Island, 2006.

[16] J. Dougherty, R. Kohavi, and M. Sahami, *Supervised and unsupervised discretization of continuous features*, Proceeding of the Twelfth International Conference on Machine Learning, pp. 194-202, 1995.

[17] Tensorflow, https://www.tensorflow.org/, June. 2019.

다중선형회귀분석 기법을 이용한 고교야구투수의 방어율 예측

오영환

나사렛대학교 IT융합학부 교수

요 약

최근에 인공 지능에 관한 연구가 활발히 진행 중이고, 학자와 연구원들이 인간을 넘어서는 인공 지능을 제작하기 위해 끊임없이 노력 중이다. 기계 학습은 수학의 논리적인 알고리즘을 이용해 각종 데이터를 학습, 분석하고 예측한다. 이 분석한 내용을 기반으로 의사 결정을 도와준다. 최종적으로 소프트웨어나 프로그램에 최적화된 알고리즘을 직접 기술해 넣는 것이 아니라 대용량의 빅데이터와 학습 알고리즘을 이용하여 IT기기 또는 전자장치 그 자체가 학습하도록 하여 실행 능력을 만드는 것을 목표로 한다. 선형 회귀는 하나 이상의 독립 변수 x와 종속 변수 y와의 선형 상관 관계를 모델링하는 회귀분석 기법으로 일반적으로 최소제곱법을 이용해 선형 회귀 모델을 정립한다. 다중 선형회귀분석은 독립 변수가 2개 이상인 회귀 모형을 분석 대상으로 한다. 단순 회귀분석은 하나의 독립 변수를 가지고 추정함으로써 편의현상(bias)를 가지고 올 수 있다. 야구선수의 투수 능력은 속구 스피드, 제구, 집중력이 좌우한다. 특히 투수의 스탯 중 WHIP, 피안타율, 탈삼진율은 투수의 중요한 척도가 된다. 본 논문에서는 투수의 WHIP, 피안타율, 탈삼진율을 학습

데이터로 하여 방어율을 예측한다. 이를 토대로 하여 특정 투수가 프로에 지명되는 지를 예견할 수 있다.

감사의 글

본 논문은 2020년 나사렛대학교 학술연구비 지원에 의해 연구되었음

Young Hwan Oh received the bachelor's degree in the Department of Computer Engineering from the Inha University in 1991. He received the M.S. degree and the Ph.D. degree in the Department of Computer Science and Engineering from Inha University in 1997 and 2001, respectively. From 1992 to 1994, he was a researcher at Korea Power Corporation. He has been a professor in the Department of Information and Communication at Korea Nazarene University since 2002. His current research interests include artificial intelligence, big data, database system. He is a member of the KKITS.

E-mail address: yhoh@kornu.ac.kr