

디지털미디어랩 머신러닝 여름캠프 3주차

(4) Multivariable Linear Regression과 실습

지난 시간

- 머신러닝(ML)이란 무엇인가?
- 지도 학습 (Supervised Learning)
 - 회귀 (Regression)
 - 분류 (Classification)
- 비지도 학습 (Unsupervised Learning)
- Linear Regression 모델 생성방법
 - Hypothesis
 - Cost Function

지난 시간

- Hypothesis

$$H(x) = Wx + b$$

- Cost Function

$$Cost(W, b) = \sum (H(x^{(i)}) - y^{(i)})^2$$

- Linear Regression의 목표는 비용이 최소인 W, b 값을 찾는 것

$$\underset{W, b}{\text{Minimize}} Cost(W, b)$$

목차

- Multivariable Linear Regression
- Hypothesis
- Cost Function
- Biking Sharing Demand 실습

Multivariable Linear Regression

3가지 변수(feature)

x_1 (quiz 1)	x_2 (quiz 2)	x_3 (midterm 1)	y (final)
73	80	75	152
93	88	93	185
89	91	90	180
96	98	100	196
73	66	70	142

Hypothesis

$$H(x) = Wx + b$$



$$H(x_1, x_2, x_3) = w_1x_1 + w_2x_2 + w_3x_3 + b$$

Cost Function

$$Cost(W, b) = \sum (H(x^{(i)}) - y^{(i)})^2$$



$$Cost(w_1, w_2, w_3, b) = \sum (H(x_1^{(i)}, x_2^{(i)}, x_3^{(i)}) - y^{(i)})^2$$

Example

x_1 (quiz 1)	x_2 (quiz 2)	x_3 (midterm 1)	y (final)
73	80	75	152
93	88	93	185
89	91	90	180
96	98	100	196
73	66	70	142

$$H_1(x_1, x_2, x_3) = 0.7x_1 + 0.2x_2 + 1.1x_3$$

$$\text{Cost}(0.7, 0.2, 1.1, 0)$$

$$= (152 - 149.6)^2 + (185 - 185)^2 + (180 - 179.5)^2 + (196 - 196.8)^2 + (142 - 141.3)^2$$

$$= 7.14$$

Example

x_1 (quiz 1)	x_2 (quiz 2)	x_3 (midterm 1)	y (final)
73	80	75	152
93	88	93	185
89	91	90	180
96	98	100	196
73	66	70	142

$$H_1(x_1, x_2, x_3) = 0.7x_1 + 0.2x_2 + 1.1x_3$$

- w_1, w_2, w_3 은 각 feature 들이 y 에 미치는 영향의 정도. (Weight)
- Ex) Midterm 에서 높은 점수를 받은 사람은 Final에도 높은 점수를 받을 것이다.

Orange 실습 1 – 시험점수 예측

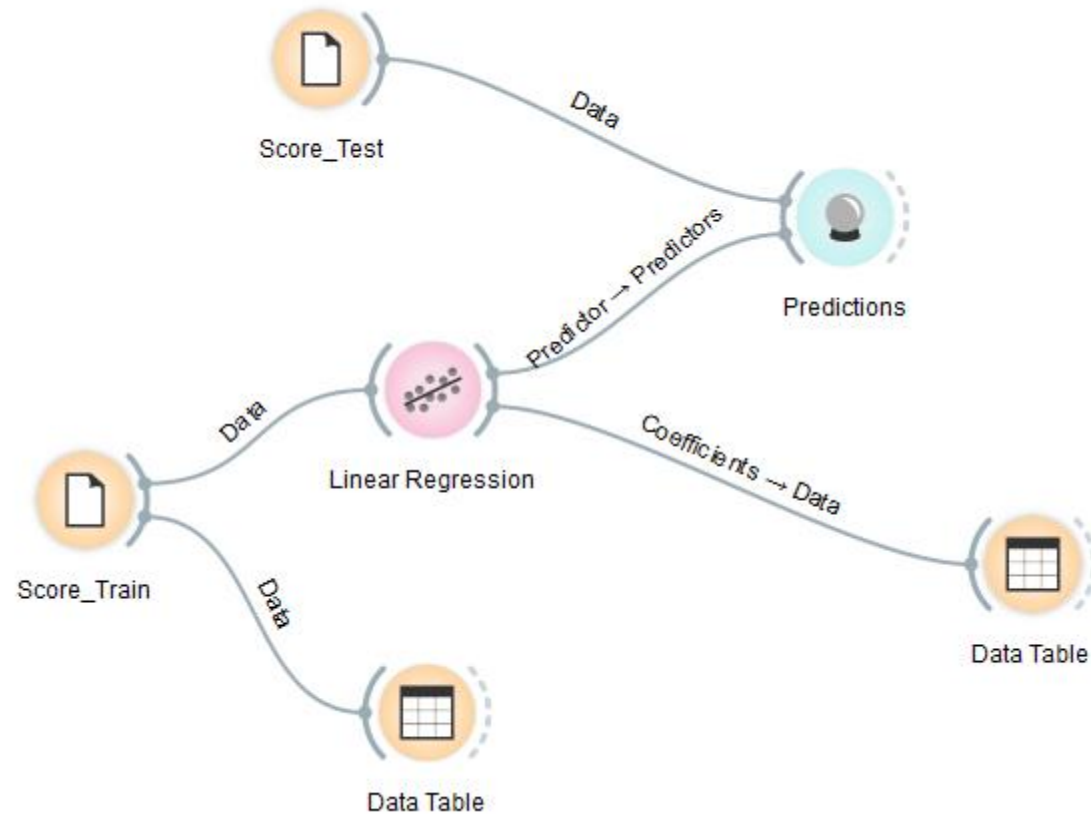
- week3_train.xlsx

x ₁ (quiz 1)	x ₂ (quiz 2)	x ₃ (midterm 1)	y (final)
73	80	75	152
93	88	93	185
89	91	90	180
96	98	100	196
73	66	70	142

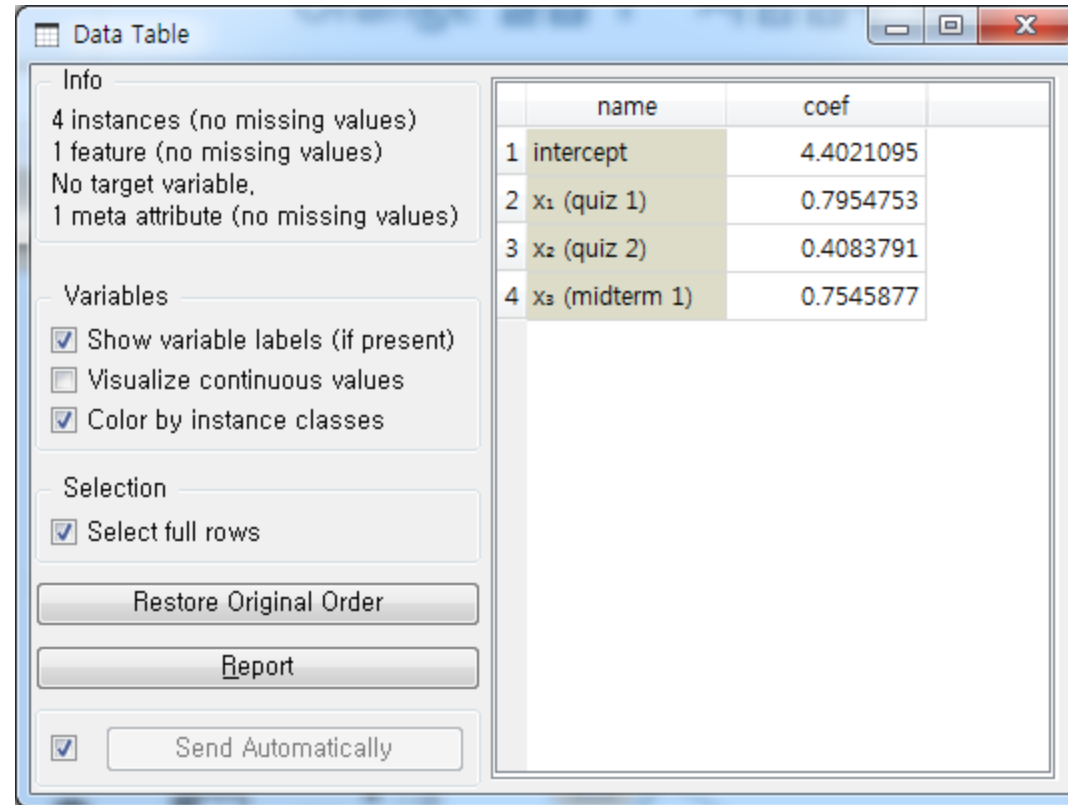
- week3_test.xlsx

x ₁ (quiz 1)	x ₂ (quiz 2)	x ₃ (midterm 1)	y (final)
100	100	100	
50	50	50	
69	96	77	
0	0	0	
25	50	50	

Orange 실습 1 - 시험점수 예측



Orange 실습 1 – 시험점수 예측



Data Table

Info

4 instances (no missing values)
1 feature (no missing values)
No target variable.
1 meta attribute (no missing values)

Variables

Show variable labels (if present)
 Visualize continuous values
 Color by instance classes

Selection

Select full rows

Restore Original Order

Report

Send Automatically

	name	coef
1	intercept	4.4021095
2	x ₁ (quiz 1)	0.7954753
3	x ₂ (quiz 2)	0.4083791
4	x ₃ (midterm 1)	0.7545877

Orange 실습 1 - 시험점수 예측

Info
Data: 5 instances,
Predictors: 1
Task: Regression
Restore Original Order

Data View
 Show full data set

Output
 Original data
 Predictions
 Probabilities
Report

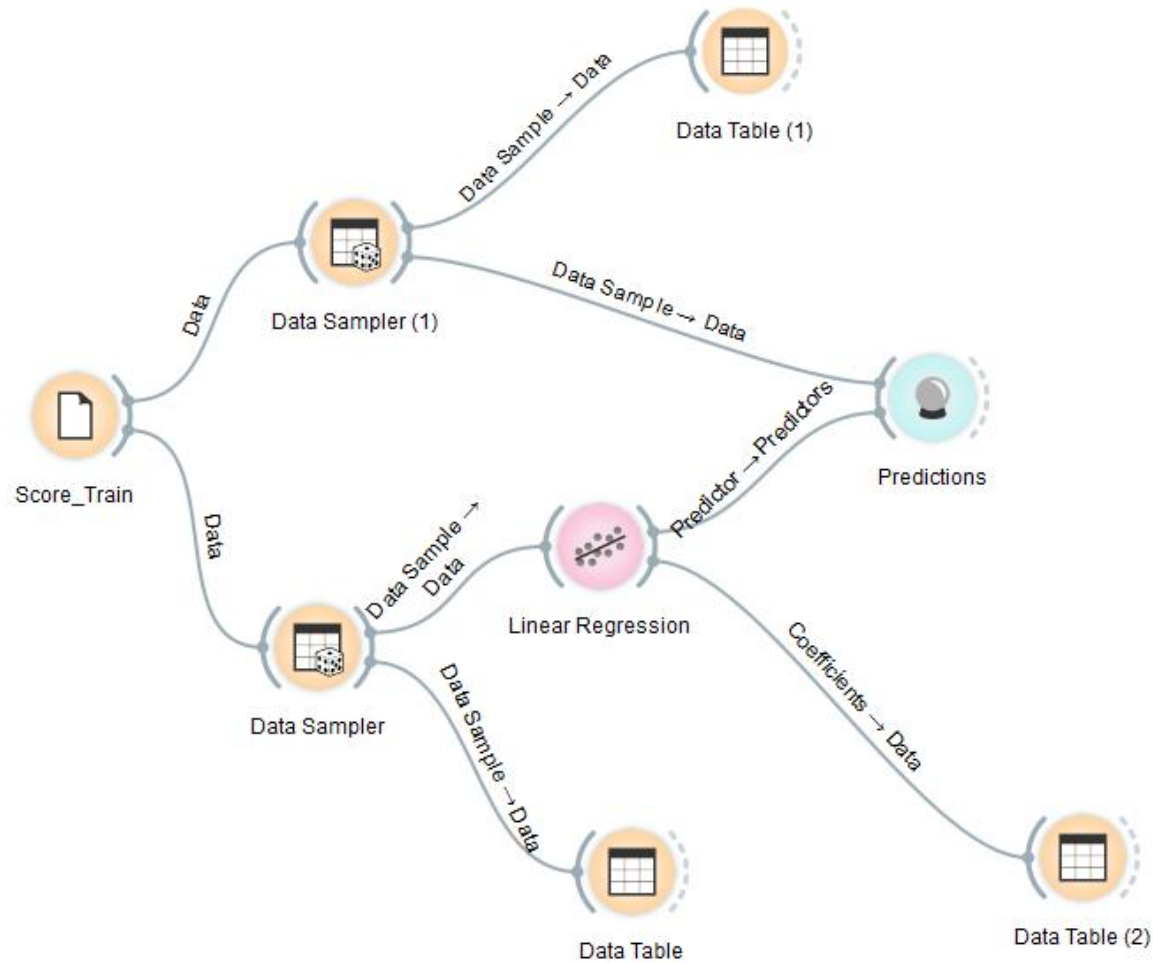
	Linear Regression	y (final)	x ₁ (quiz 1)	x ₂ (quiz 2)	x ₃ (midterm 1)
1	200.246	?	100.000	100.000	100.000
2	102.324	?	50.000	50.000	50.000
3	156.598	?	69.000	96.000	77.000
4	4.402	?	0.000	0.000	0.000
5	82.437	?	25.000	50.000	50.000

Orange 실습 1 – Bike Sharing Demand

Bike Sharing Demand

- datetime - hourly date + timestamp
- season - 1 = spring, 2 = summer, 3 = fall, 4 = winter
- holiday - whether the day is considered a holiday
- workingday - whether the day is neither a weekend nor holiday
- weather -
 - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
 - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
 - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
 - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp - temperature in Celsius
- atemp - "feels like" temperature in Celsius
- humidity - relative humidity
- windspeed - wind speed
- count - number of total rentals

Orange 실습 1 – Bike Sharing Demand



Orange 실습 1 – Bike Sharing Demand

Info
Data: 3266 instances,
Predictors: 1
Task: Regression
[Restore Original Order](#)

Data View
 Show full data set

Output
 Original data
 Predictions
 Probabilities

Report

	Linear Regression	count	season	holiday	workingday	weather	temp	atemp	
1	338.366	537.000	2.000	0	1	1.000	24.600	30.305	26.
2	216.632	219.000	4.000	0	0	2.000	13.940	15.150	49.
3	357.216	619.000	4.000	0	0	1.000	26.240	31.060	41.
4	112.350	157.000	2.000	0	1	3.000	20.500	24.240	88.
5	381.126	497.000	3.000	0	1	2.000	34.440	37.120	36.
6	268.105	143.000	2.000	0	0	1.000	30.340	34.090	58.
7	150.760	185.000	3.000	0	1	1.000	26.240	28.790	89.
8	131.041	132.000	2.000	0	1	2.000	21.320	25.000	83.
9	123.126	12.000	4.000	0	1	1.000	9.020	10.605	64.
10	157.500	79.000	4.000	0	0	1.000	13.940	17.425	66.
11	156.382	78.000	4.000	0	1	3.000	22.140	25.760	94.
12	156.007	43.000	4.000	0	0	1.000	9.020	13.635	55.
13	89.329	10.000	4.000	0	1	1.000	13.940	18.180	87.
14	155.144	6.000	2.000	0	1	2.000	22.140	25.760	77.
15	195.911	240.000	2.000	0	1	1.000	25.420	30.305	69.
16	77.027	90.000	1.000	1	0	1.000	10.660	12.120	60.
17	192.284	157.000	4.000	0	0	1.000	10.660	11.365	48.
18	376.760	373.000	3.000	0	0	1.000	34.440	37.120	36.
19	130.281	2.000	3.000	0	1	3.000	22.960	26.515	94.
20	198.182	184.000	2.000	0	1	2.000	26.240	30.305	73.
21	238.149	400.000	4.000	0	0	2.000	14.760	16.665	46.
22	88.276	135.000	2.000	0	1	1.000	22.140	25.760	94.
23	99.546	35.000	1.000	0	1	1.000	8.200	9.090	51.

참고 자료

모두를 위한 머신러닝/딥러닝

<http://hunkim.github.io/ml/>

$$H(x) = Wx + b$$