

데이터 전처리

- 데이터 전처리
 - 데이터를 그대로 사용하지 않고, 가공해서 모델을 학습 시키는데 좀 더 좋은 형식으로 만들어주는 것
- Feature Scaling
 - 머신 러닝 모델에 사용할 입력 변수들의 크기를 조정해서 일정 범위 내에 떨어지도록 바꾸는 것
 - 경사 하강법을 좀 더 빨리할 수 있게 도와준다
 - 방법
 - min-max normalization
 - 데이터의 최소, 최대값을 이용해서 데이터의 크기를 0과 1사이로 바꿔준다

원래 데이터	$\text{최대값} - \text{최소값}$ $= 210 - 140 = 70$	Normalize 한 데이터
키 (cm)		키 (cm)
180	$\frac{180 - 140}{70} = 0.57$	0.57
170	$\frac{170 - 140}{70} = 0.43$	0.43
<small>최소</small> 140	$\frac{140 - 140}{70} = 0$	0
<small>최대</small> 210	$\frac{210 - 140}{70} = 1$	1
160	$\frac{160 - 140}{70} = 0.29$	0.29

min-max normalization 일반화

중지

x_{new} : Feature scaling 한 데이터

x_{old} : Feature scaling 하기 전 데이터

x_{max} : 데이터 최댓값

x_{min} : 데이터 최솟값

0과 1사이의
숫자로 바꿈

$$x_{new} = \frac{x_{old} - x_{min}}{x_{max} - x_{min}}$$

- 선형 회귀뿐만 아니라 경사 하강법을 사용하는 모든 알고리즘의 속도를 빠르게 해준다.
 - 정규화 또한 가능
- 머신 러닝에서 사용하는 데이터
 - 수치형 데이터 : 나이, 몸무게, 키
 - 범주형 데이터 : 혈액형, 성별 (→ 수치형 데이터로 바꿔줘야!)
 - One-hot Encoding
 - 각 카테고리의 열을 새로운 열로 만들어주는 방법

One-hot Encoding

범주형 데이터에 일정한 관계를 만든지 않으면서 수치형 데이터로 바꿀 수 있음!

혈액형	나이		A형	AB형	B형	O형	나이
A	9	→	1	0	0	0	9
AB	40		0	1	0	0	40
B	35		0	0	1	0	35
O	20		0	0	0	1	20
A	70		1	0	0	0	70