

# 머신러닝(Machine Learning)

데이터 처리

# 목 차

01 데이터 전처리와 스케일 조정

02 원핫 인코딩(가변수)

03 구간 분할(bining)

04 원본 특성에 다항식을 추가하기

05 비선형 변환

# 01 데이터 전처리와 스케일 조정

## ▶ 표준화(StandardScaler)

(가) 각 특성(feature)의 평균을 0, 분산을 1로 변경

## ▶ RobustScaler

(가) 같은 스케일을 갖는다.

(나) 평균과 분산 대신 중간 값(median)과 사분위 값(quantile)을 사용.

## ▶ MinMaxScaler - 정규화

(가) 모든 특성이 정확하게 0과 1사이에 위치하도록 데이터를 변경

## ▶ Normalizer

(가) 유클리디안 길이가 1이 되도록 데이터 포인트를 조정.

# 01 데이터 전처리와 스케일 조정

## ▶ 표준화(StandardScaler)

(가) 각 특성(feature)의 평균을 0, 분산을 1로 변경

(나) StandardScaler 공식

$$X_{new} = \frac{X - X_{mean}}{X_{std}}$$

# 01 데이터 전처리와 스케일 조정

## ▶ RobustScaler

(가) 통계적 측면에서 StandardScaler과 유사.

(나) 평균과 분산 대신 median과 quantile을 사용.

- \* outlier의 영향을 받지 않음.

- \* 이상치가 있고 이를 제거하지 않을 경우, RobustScaler가 최선

(다) RobustScaler 공식

$$X_{new} = \frac{X - X_{median}}{(q3 - q1)}$$

# 01 데이터 전처리와 스케일 조정

## ▶ 정규화(MinMaxScaler)

(가) 모든 특징이 정확하게 0과 1사이에 위치하도록 데이터를 변경

(나) MinMaxScaler 공식

$$X_{new} = \frac{X - X_{min}}{(X_{max} - X_{min})}$$

# 01 데이터 전처리와 스케일 조정

## ▶ Normalizer

(가) 유클리디아 길이가 1이 되도록 데이터 포인트를 조정.

- \* 지름이 1인 원에 데이터 포인트 투영한다.

- \* 데이터의 방향(또는 각도)만이 중요할 때 많이 사용.

# 01 데이터 전처리와 스케일 조정

## ▶ 스케일링의 목적

(가) 표준화된 스케일로 각 특징을 다시 표현하여 너무 큰 값으로 인한 잠재적인 수치 불안정성을 방지한다.



## 02 원 핫 인코딩(가변수)

### ▶ Label Encoding : 범주형 데이터를 숫자로 변경

(가) 머신러닝 알고리즘에서 범주형 데이터는 숫자로 변경되어야 한다.

### ▶ 원핫 인코딩 작업 절차

(가) 범주형 변수는 정수 값으로 변경 (라벨 인코딩)

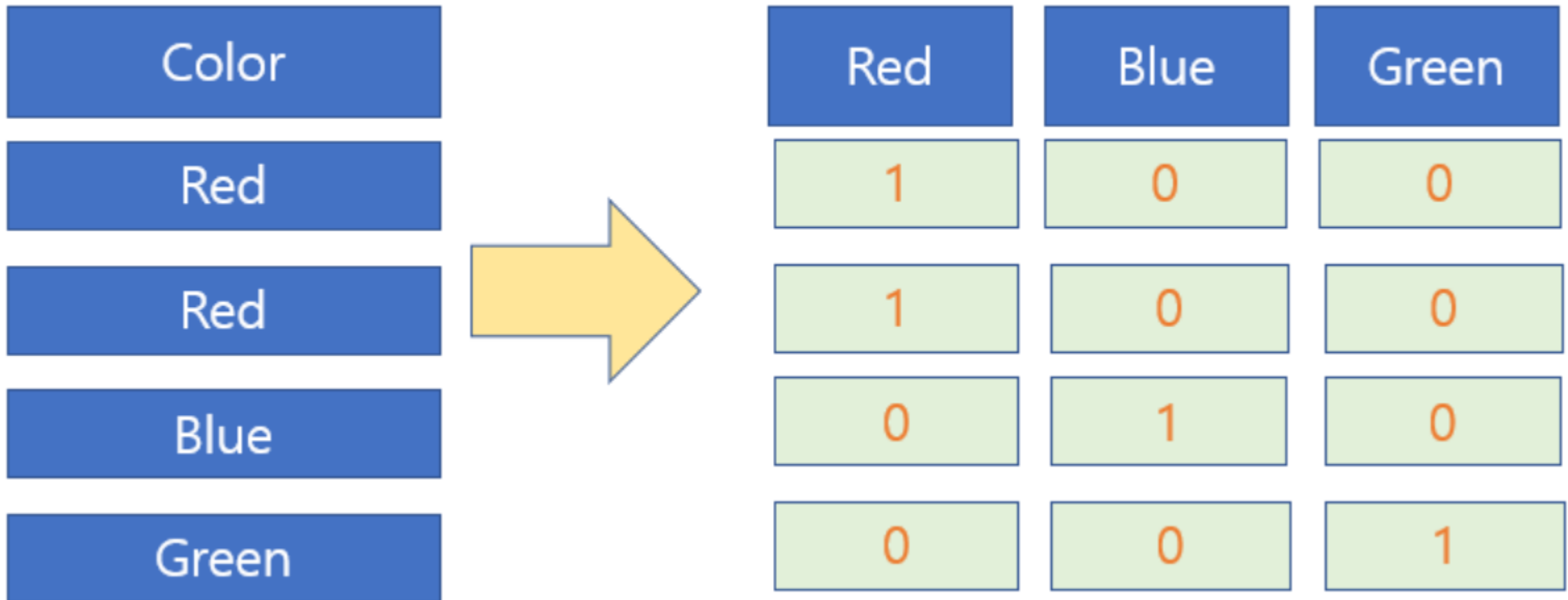
(나) 각각의 정수 값은 해당되는 위치에 1로 표시. 나머지는 0으로 표시

## 02 원 핫 인코딩(가변수)

- ▶ Label Encoding 이 범주형 구분을 숫자로 변경
- ▶ 범주형 변수를 표현하는데 가장 널리 쓰이는 방법
- ▶ 가변수는 범주형 변수를 0 또는 1 값을 가진 하나 이상의 새로운 특성으로 바꾼 것.
- ▶ 통계학에서 사용하는 더미 코딩(dummy coding)과 비슷하지만 완전히 같지 않음.

## 02 원 핫 인코딩(가변수)

### ▶ 범주형 데이터를 이진 벡터(0,1)로 표현



## 03 구간 분할(bining)

- ▶ 연속형 데이터에 가장 강력한 선형 모델을 만드는 방법 중 하나.
- ▶ 한 특성을 여러 특성으로 나누는 구간 분할(bining)

## 04 원본 특성에 다항식을 추가하기

- ▶ preprocessing 모듈의 PolynomialFeatures를 이용 가능.

## 05 비선형 변환

- ▶  $\log$ ,  $\exp$ ,  $\sin$  같은 수학적함수를 이용하여 특성 변환.
- ▶  $\log$ ,  $\exp$  함수는 데이터의 스케일을 변경해 선형 모델과 신경망의 성능의 향상 시킴.