Explanatory Regression and Predictive Regression

고태훈 (taehoonko@dm.snu.ac.kr)

Explanatory vs. Predictive

Explanatory regression

- 목표: 타겟변수와 입력 변수들의 관계를 설명
- Fit the data well
- Understand the contribution of input variables to the model.

Predictive regression

- 목표: 새로운 데이터의 타겟변수값을 예측
- Train the model on training set
- Evaluate the performance of the model using validation(or test) set.

Explanatory regression

- **설명적 목적의 회귀모델은 출력변수 Y에 입력변수 X들이 어떠한 영향을 미치는지 살펴보는 데에 목적이 있음**
  - 학습된 회귀모델이 얼마나 많은 정보를 설명하는가? ➔ "Goodness-of-fit"
  - 각 입력변수들이 통계적으로 출력변수에 유의한 영향을 미치고 있는가? ➔ t-stastics
  - 이외에 여러 가지 통계검정을 목적으로 사용

Explanatory regression

표 2. 회귀분석에 의한 단순매개효과 검증 결과

	b	*SE*	*t*	*p*
총효과와 직접효과				
감정이 태도에 미치는 총효과:	1.00	.19	5.30	.00
감정이 긍정적인 사고에 미치는 직접효과:	.70	.14	5.02	.00
긍정적인 사고가 태도에 미치는 직접효과(감정의 영향을 통제함):	.60	.29	2.04	.05
감정이 태도에 미치는 직접효과(긍정적인 사고의 영향을 통제함):	.58	.27	2.15	.04

간접효과		Boot *SE*	LL 95% CI	UL 95% CI
간접효과에 대한 부트스트랩 검증 결과				
Effect	.42	.24	.08	1.05

Note. *b*는 비표준화된 회귀계수를 나타냄.

정선호, 서동기. (2016.3). 회귀분석을 이용한 매개된 조절효과와 조절된 매개효과 검증 방법. 한국심리학회지: 일반, 35(1), 257-282.

Explanatory regression

❖ 회귀계수 β_i ➔ X_i가 Y에 미치는 영향

- β_i의 부호
  - $\beta_i > 0$: X_i가 Y에 긍정적인 영향
  - $\beta_i < 0$: X_i가 Y에 부정적인 영향
- $|\beta_i|$
  - X_i가 Y의 값에 공헌한 정도
  - ex) X_1와 X_2의 스케일(scale)이 같은 경우 $\beta_i > 0$, $\beta_2 > 0$, $|\beta_1| > |\beta_2|$라면, X_1이 X_2보다 Y값 증가에 더 큰 영향을 미침

    ➔ 입력변수들의 영향력을 비교해서 보고자 하는 경우에는 반드시 **feature scaling**을 해야 함

참고: Feature scaling

❖ Standardization (Z-score normalization)

- The features will be rescaled so that they'll have the properties of a standard normal distribution with $\mu = 0$ and $\sigma = 1$.

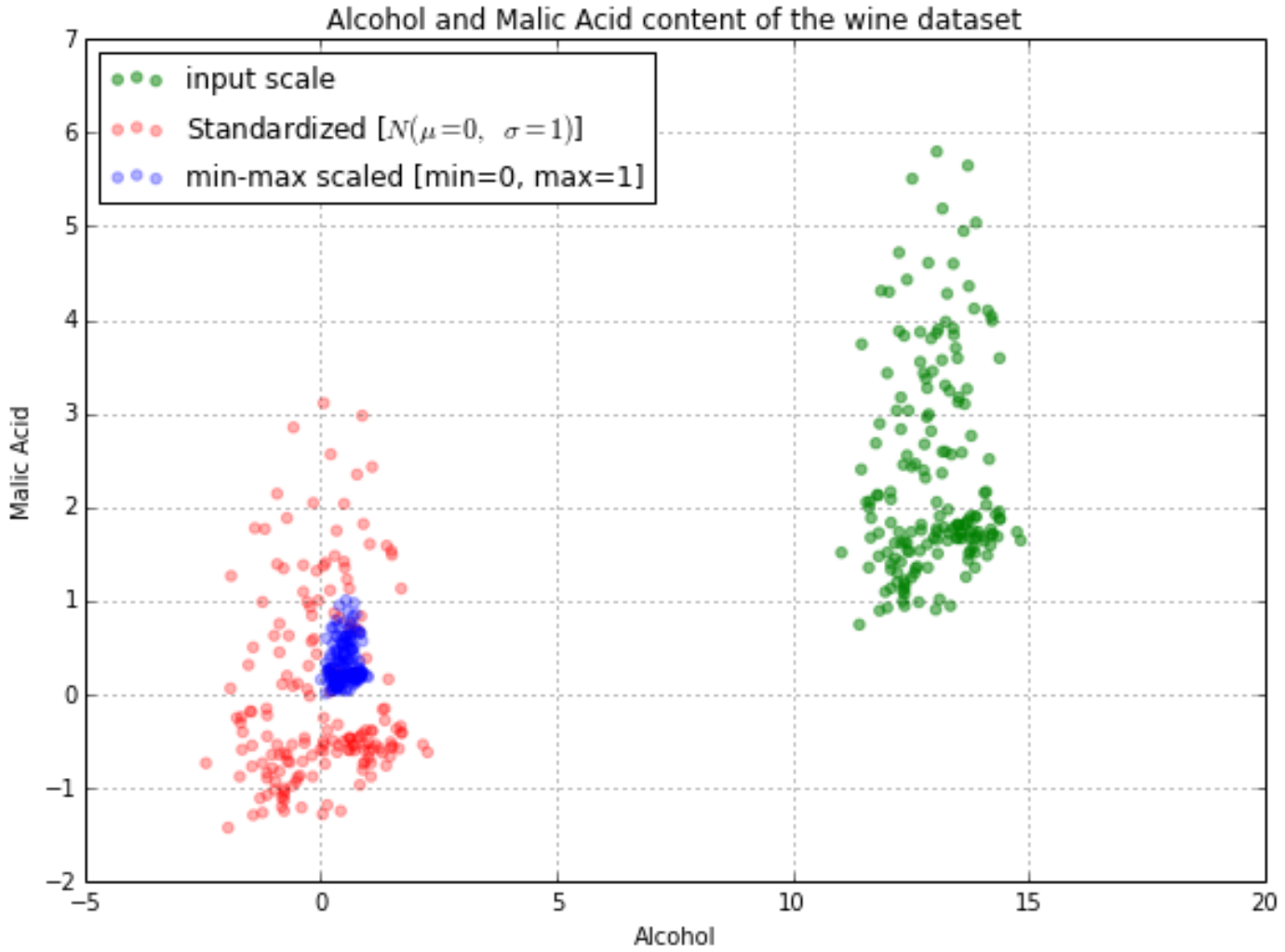
$$Z = \frac{X - \mu(X)}{\sigma(X)}$$

❖ Min-Max scaling (Normalization)

- The feature is rescaled to a fixed range [0,1]

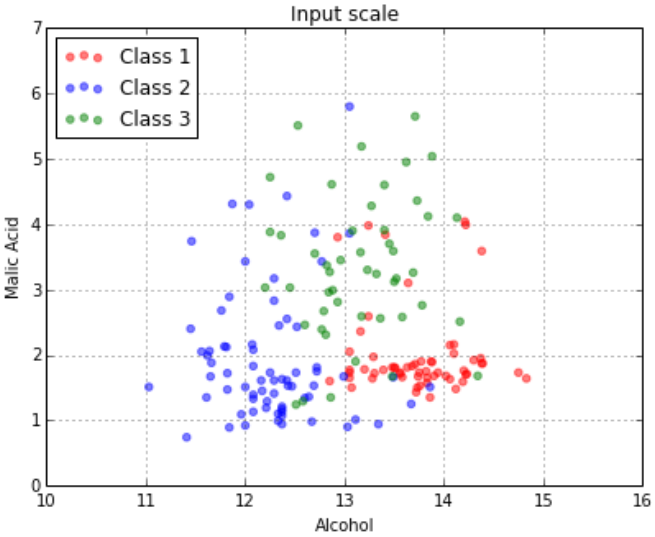
$$X_{norm} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

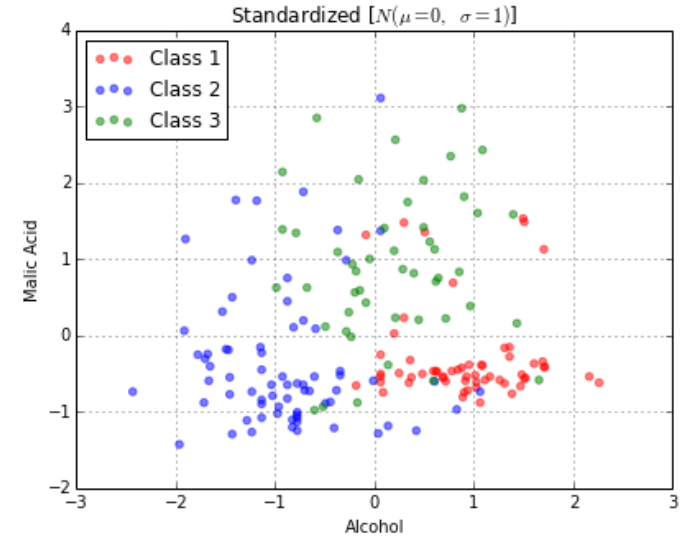
참고: Feature scaling

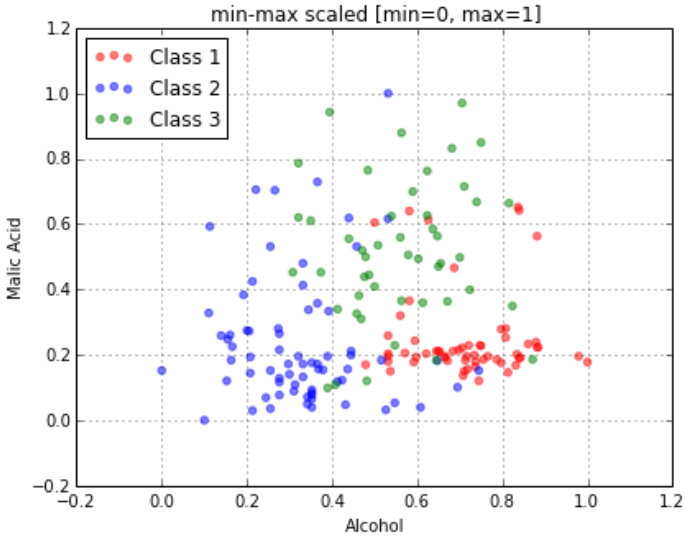


http://sebastianraschka.com/Articles/2014_about_feature_scaling.html

참고: Feature scaling







http://sebastianraschka.com/Articles/2014_about_feature_scaling.html

참고: Feature scaling

- **데이터 포인트 간의 거리/유사도를 기반으로 작동하는 알고리즘**
  - ex) k-means clustering, hierarchical clustering, k-nearest neighbors, support vector machine (SVM) with radial basis function (RBF) kernel
  - 스케일이 큰 입력변수가 큰 영향을 미치는 distance/similarity metrics를 사용하는 경우 feature scaling이 필요할 수 있음: Euclidean distance, RBF kernel, etc.
- **Gradient descent/ascent 기반의 최적화를 통해 모델을 학습하는 알고리즘**
  - ex) Logistic regression, neural networks, support vector machines (SVMs)
  - Gradient가 대부분 입력변수의 스케일에 비례하므로 feature scaling을 통해 학습 속도를 빠르게 할 수 있음
- **Zero-centered data라는 가정을 기반으로 하는 알고리즘**
  - ex) Principal component analysis (PCA), SVM with radial basis function (RBF) kernel
  - 위의 학습 방법은 기본 가정이 zero-centered data (즉, 각 변수의 평균이 0) 에서 출발

Explanatory regression

❖ 선형회귀모델의 적합도 (Goodness-of-fit)

- R^2: 데이터 전체의 변동 중 선형회귀모델이 설명하는 변동의 비율
  - $0 \leq R^2 \leq 1$
  - 1에 가까울수록 회귀모델이 데이터를 잘 설명한다는 뜻
  - ex) $R^2 = 0.91$인 경우, 전체 데이터 변동성의 91%를 선형회귀모델이 설명

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})}{\sum(y_i - \bar{y})}$$

Explanatory regression

❖ 선형회귀계수의 검정

- 가설 $H_0 : \hat{\beta}_i = 0 \quad H_1 : \text{not } H_0$
- 검정통계량 $$t = \frac{\hat{\beta}_i - 0}{\text{standard error}(\hat{\beta}_i)}$$
- 2-sided p-value를 계산
  - p-value < α ➔ (1- α) 유의수준에서 귀무가설을 기각, 즉 $\hat{\beta}_i \neq 0$

    ➔ $\hat{\beta}_i$이 통계적으로 Y에 유의한 영향을 미치고 있음
  - ex) $\hat{\beta}_1$에 대한 p-value = 0.04 < 0.05라면, 95% 유의수준에서 $\hat{\beta}_1$이 통계적으로 Y에 유의한 영향을 미치고 있음

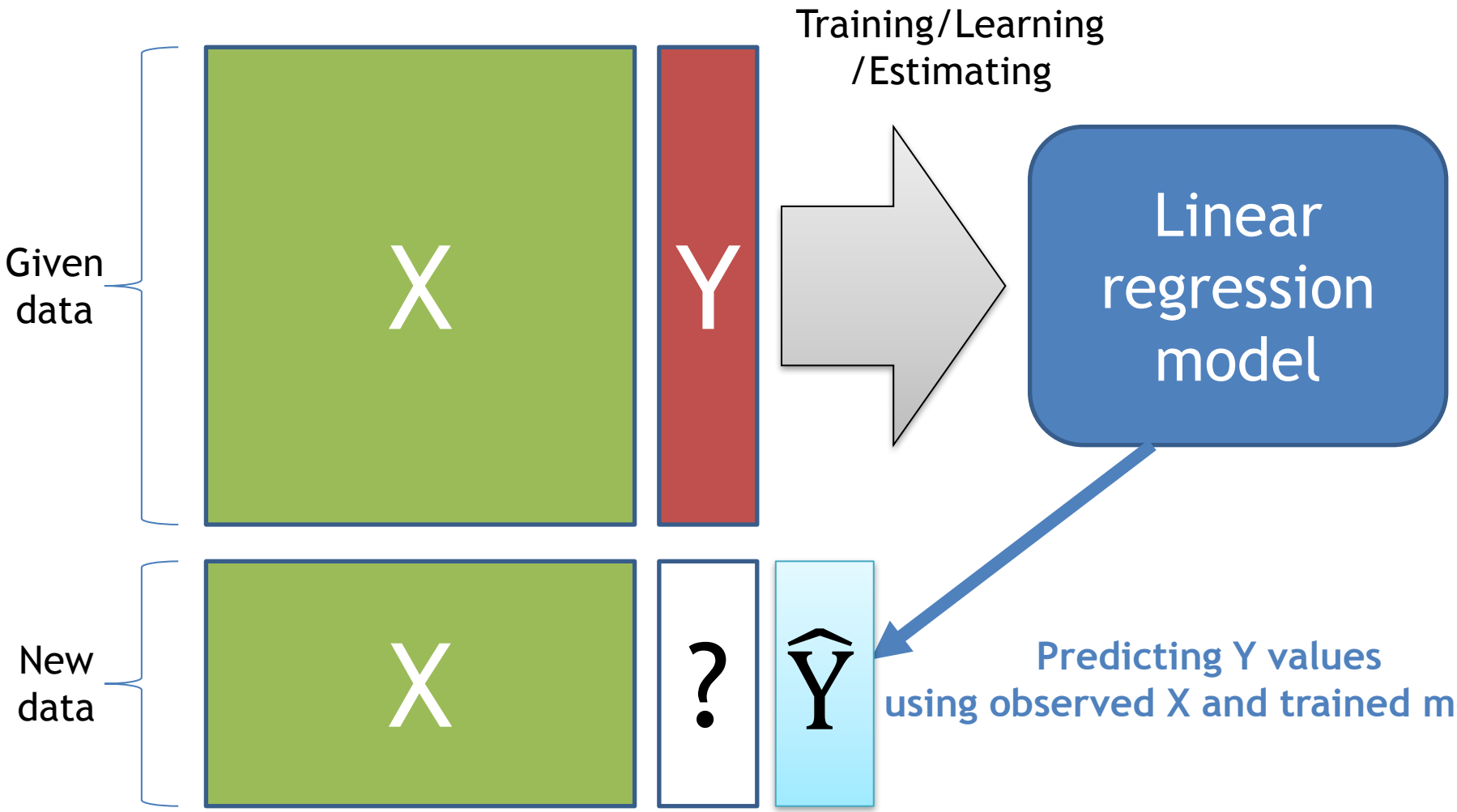
Predictive regression

❖ 주어진 데이터로부터 학습된 회귀모델로 미래를 예측

- 예측하고자 하는 대상
  - 현상이 관측되었을 때 ➔ 입력변수 X_1, X_2, …, X_p 의 값이 기록된 레코드가 등장하였을 때,
  - 이 현상에 의한 결과를 예측 ➔ 출력변수 Y를 예측
- 주어진 데이터: 현상과 이에 의한 결과
  - 관측된 현상과 이에 대한 결과가 기록되어 있는 데이터

    ➔ 입력변수 X_1, X_2, …, X_p와 출력변수 Y의 값이 기록된 레코드들

Predictive regression

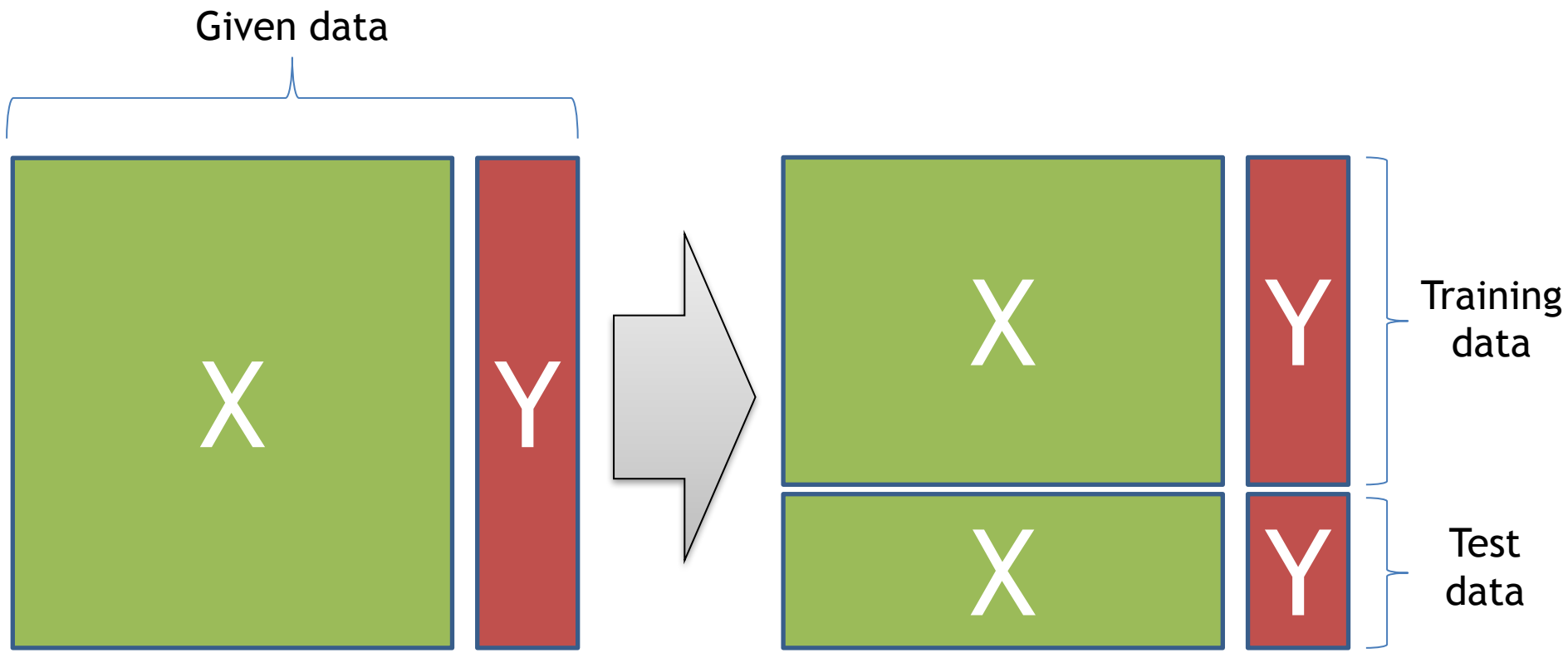


Predictive regression

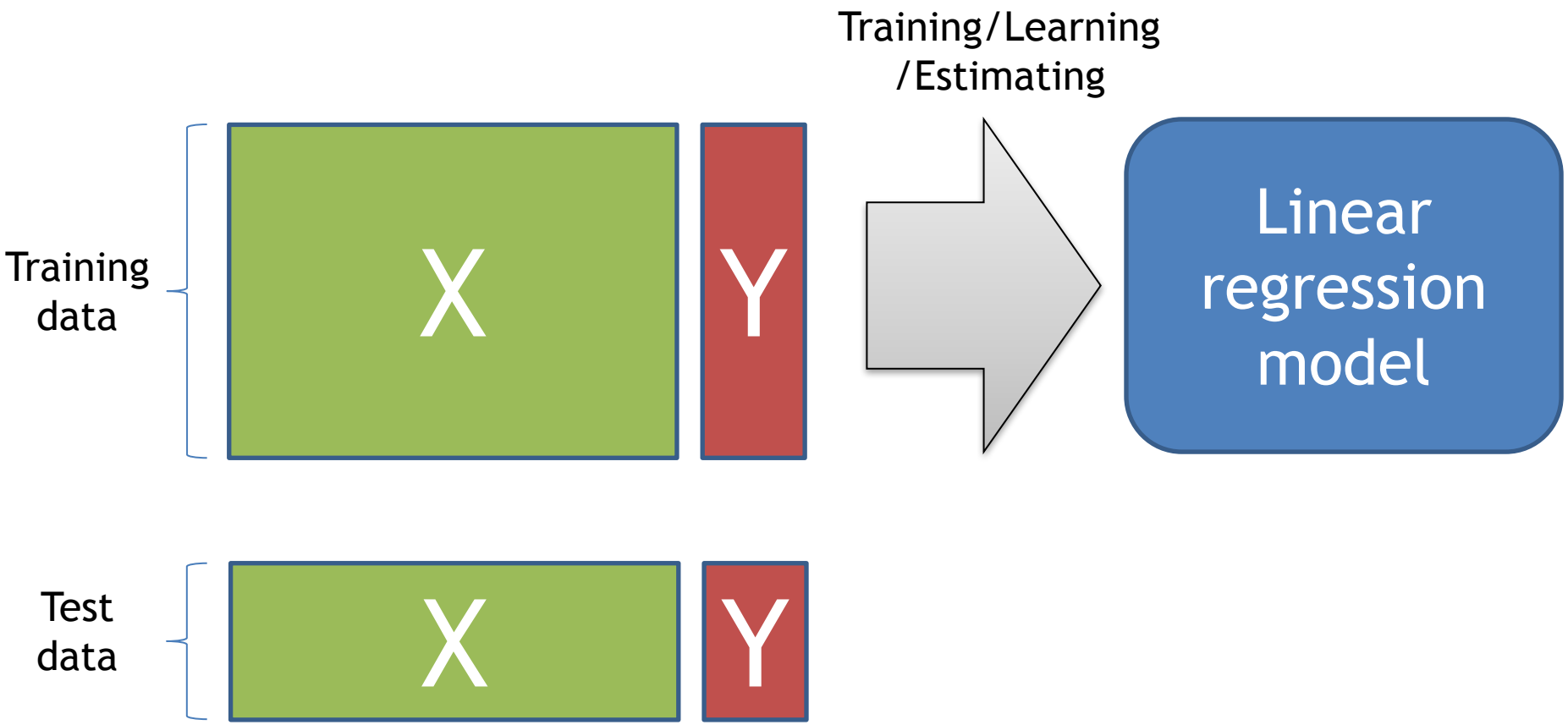
❖ 선형회귀모델의 성능평가는 Train error와 Test error로 구분

- 주어진 데이터를 다음 두 가지로 나눔
  - 학습데이터: Training data / Training set
  - 검증데이터: Validation data / Validation set / Test data / Test set
  - (학습데이터):(검증데이터) = 7:3 / 6:4 / 5:5 / …
- 학습데이터로 선형회귀모델을 학습한 후,
  - 학습데이터에서의 Y와 $\hat{Y}$를 비교: Train error
  - 검증데이터에서의 Y와 $\hat{Y}$를 비교: Validation error / Test error
  - **Train error는 모델의 적합도, Test error는 모델의 예측성능**

Predictive regression



Predictive regression



Predictive regression