Machine Learning

Introduction

김 광 수

2024년 2학기

통계학 vs 기계학습

- 두 용어가 같은 것을 지칭하는가?
- 통계학 : 모형과 데이터로부터 모집단, 모수에 대한 추론, 해석에 초점을 맞춤.
- 기계학습 : 모형과 데이터로부터 새로운 현상에 대한 예측성능에 보다 초점을 맞춤.
- 지향점에는 다소 차이가 있을 수 있으나 사용하는 도구(tool)에는 공통점이 많고, 무엇보다 데이터를 다룬다는 점에서 큰 유사성을 가진다.

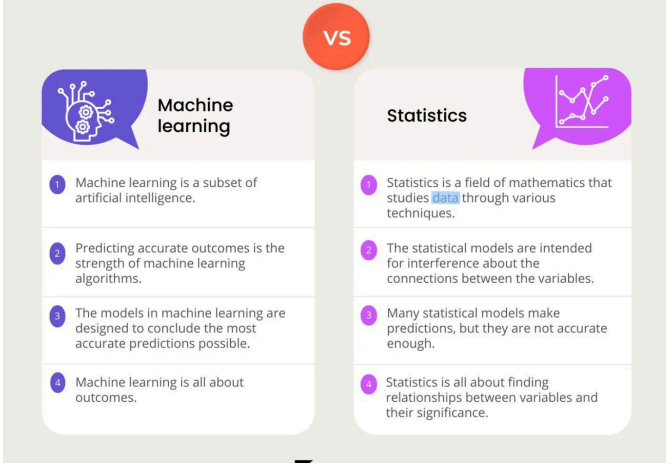

Figure 1: Statistics vs machine learning

https://www.turing.com/kb/introduction-to-https://images.prismic.io/turing/65980ec7531ac2845a272849_Machine_learning_vs_Statistics_a080e3c811.webp?auto=format,compress-for-machine-learning

통계적 학습의 역사

- 기계학습이라는 키워드는 비교적 최근에 생겨났으나 기저에 깔린 개념들은 그 역사가 매우 깊다. 사전적 정의는 “경험을 통해 자동으로 개선하는 컴퓨터 알고리즘의 연구”

- 19세기 초에 최소제곱법에 근거한 선형회귀모형 등장 (천문학의 문제를 해결하기 위해 제안됨).

- 1936년, 판별분석이 제안되었고 1940년에 로지스틱모형이 소개되었다.

- 1970년 초, 선형모형과 로지스틱 모형들을 모두 포괄하는 일반화선형모형의 개념이 소개되었다.

- 1980년대, 계산성능의 향상으로 비선형모형등의 적용이 가능해졌다.

- 1980년 중반, 나무(tree)모형에 근거한 회귀/분류 방법들이 제안되었다.

- 1980년 중반, 선형회귀모형의 한계를 극복한 일반화가법모형이 소개되었다.

- 이후로, 기계학습의 출현으로 통계적 기계학습이 통계학의 한 분야로 자리잡으며 성장해 오고 있다.

- 통계학(혹은 통계적 학습)의 역사는 계산학/컴퓨터공학의 발전과 따로 생각할 수 없다. 데이터저장/공유/처리 능력의 향상, 접근 가능한 소프트웨어의 보급 등이 통계학의 발전을 더욱 풍성하게 만들어 주고 있다.

- 기계학습은 다층퍼셉트론(Multi-layer perceptron), 서포트벡터기계(Support vector machine), 심층신경망(Deep neural networks),
강화학습(Reinforcement learning)의 발전과 궤를 같이하고 있다.

기계학습의 영역

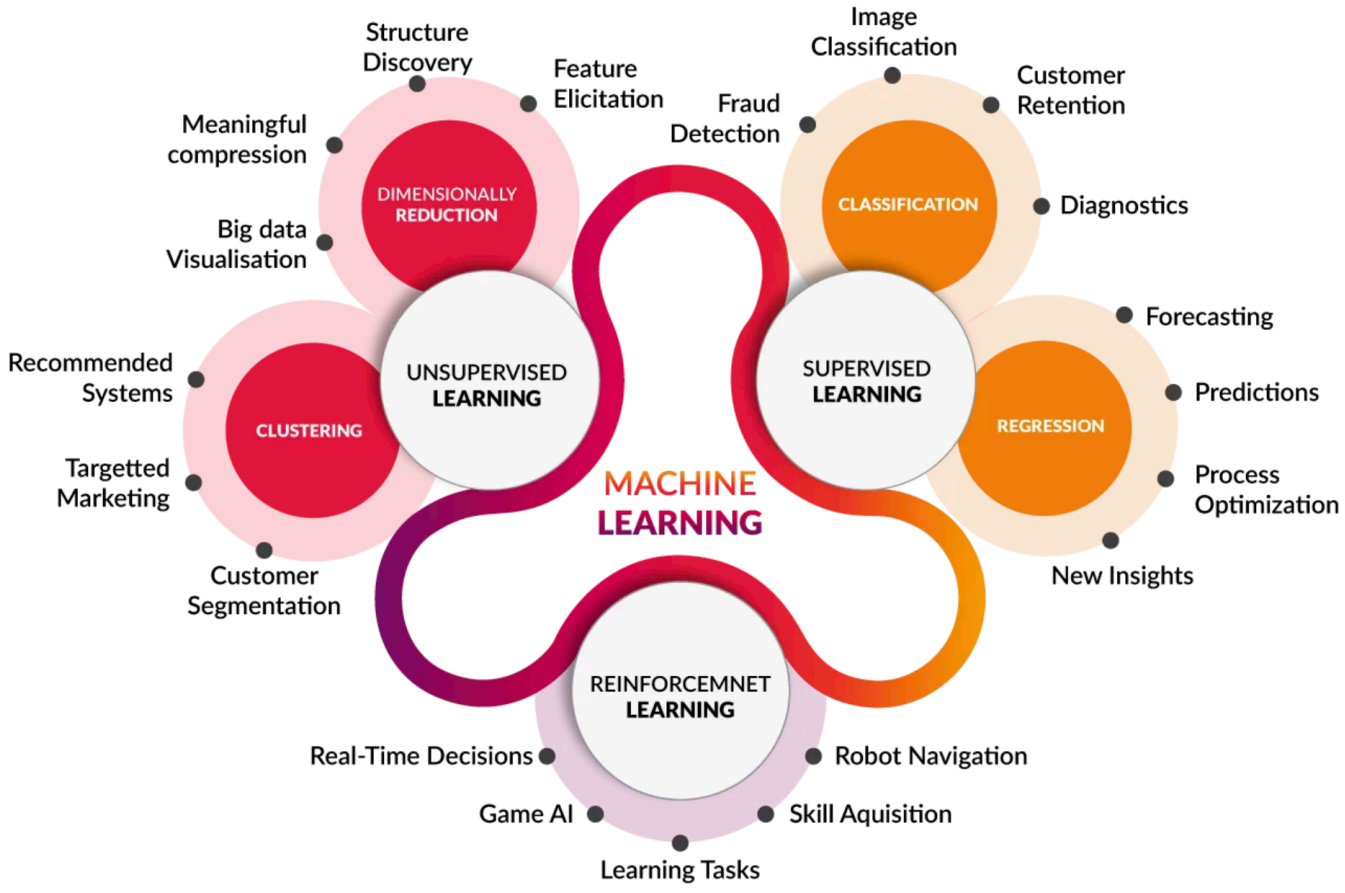

Image source: http://www.cognub.com/index.php/cognitive-platform/

Figure 2: Mahine learning

(회귀분석을 예제로 한) 통계적 학습이란 ?

- Advertising data

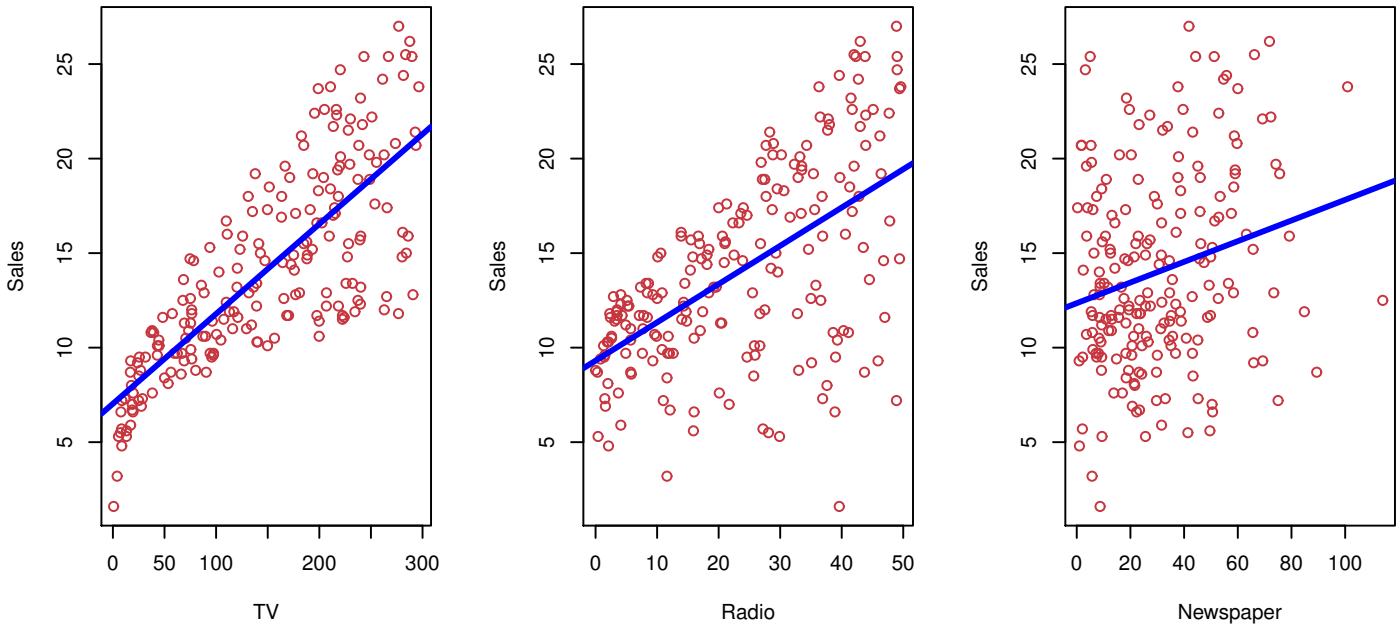

Figure 3: Statistical learning

- 일반적으로 $X_1, \dots, X_p$와 같은 p개의 예측변수(predictor)와 반응변수(response) Y가 다음과 같은 관계를 가진다고 가정했을 때,

$$Y = f(X) + \epsilon$$

  f를 데이터에 근거하여 추정해내는 방법론들을 일컬어 통계적 학습이라 한다.

- f를 추정해내기 위한 접근법은 매우 다양하게 제안되어왔다.

- 목적에 따라 f의 형태, 혹은 추정방법들을 적절히 선택해야 한다.

Why estimate f?

- 예측(prediction) : f의 추정치를 $\hat{f}$이라 하면, 주어진 예측변수 값 X에 대하여 반응변수에 대한 예측값을 다음과 같이 얻을 수 있다.

$$\hat{Y} = \hat{f}(X)$$

- 예측에만 목적이 있을 경우 $\hat{f}$의 정확한 함수형태를 아는 것은 중요치 않다. 정확한 예측을 담보하는 것이 최우선이다. 또한 Y에 대한 예측은 평균값이 최선이 된다는 것을 알 수 있다.

$$\text{argmin}_{c(X)} E[(Y - c(X))^2] = E[Y \mid X] = f(X)$$

- 통계적 학습의 궁극적인 목표는 낮은 예측오차 ($E[(Y - \hat{Y})^2]$)이며 이는 $f(X)$의 추정, 예측문제로 귀결됨.

- 예측정확도(예측오차)의 분해

- Reducible error : $\hat{f}$가 f에 가능하면 가까워지도록 적절한 학습방법을 채택함으로써 줄일 수 있는 오차.

- Irreducible error : ϵ에 의한 오차로, $\hat{f}$가 f를 완벽히 추정하더라도 피할 수 없는 오차.

$$E[(Y - \hat{Y})^2] = E[(f(X) - \hat{f}(X))^2] + Var(\epsilon)$$

Why estimate f?

- 추론(inference) : 예측변수와 반응변수의 관계 혹은 예측변수가 반응변수에 미치는 영향 등을 보다 정확히 이해하기 위함.

$$\hat{Y} = \hat{f}(X)$$

- 추론에 목적이 있을 경우 $\hat{f}$의 함수형태는 어느 정도 표현 가능해야 한다.

- 어떤 예측변수들이 반응변수와 연관되는가?

- 각 예측변수들이 반응변수와 어떻게 연관되는가?

- 변수들간의 관계가 선형함수로 충분히 요약되는가? 참고로 여기에서는 예측변수로부터 반응변수의 평균을 예측하는 문제임.

How do we estimate f?

- 모수적 방법 (Parametric method)

  - f의 형태를 구체적으로 가정한 후, 데이터로부터 모형의 학습(혹은 훈련)을 통해 f를 추정.

  - 대표적인 것이 선형회귀모형(linear regression model)

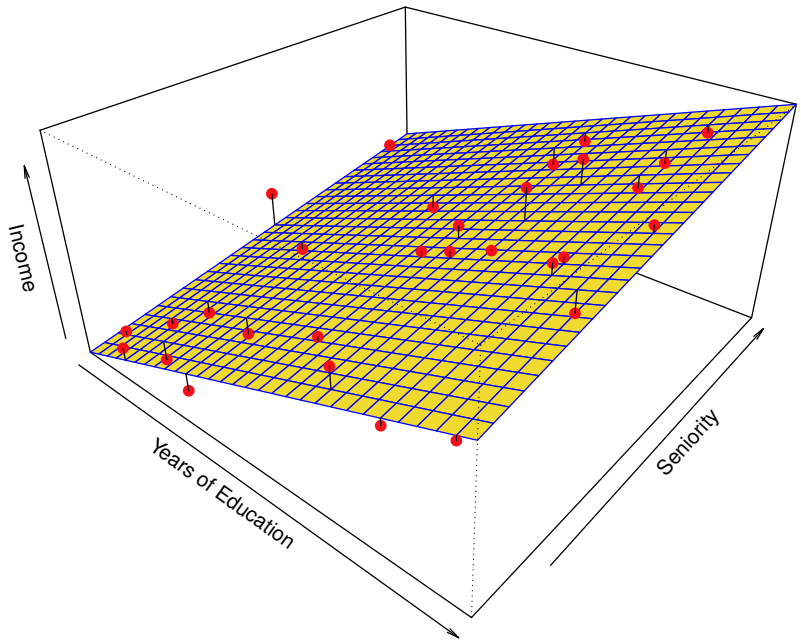

Figure 4: 모수적 방법-Income 자료에 선형회귀모형 적합

How do we estimate f?

- 비모수적 방법 (Nonparametric method)

  - f의 형태에 구체적인 가정 없이 데이터로부터 모형의 학습(혹은 훈련)을 통해 f를 추정.

  - 보통 $\hat{f}$이 지나치게 rough하지 않도록 추정함.

  - 함수형태에 대한 가정을 완전히 피할 수는 없으며, 모형의 유연성(flexibility) 증가로 인해 데이터 적합성이 좋아질 수 있다는 장점이 있음.

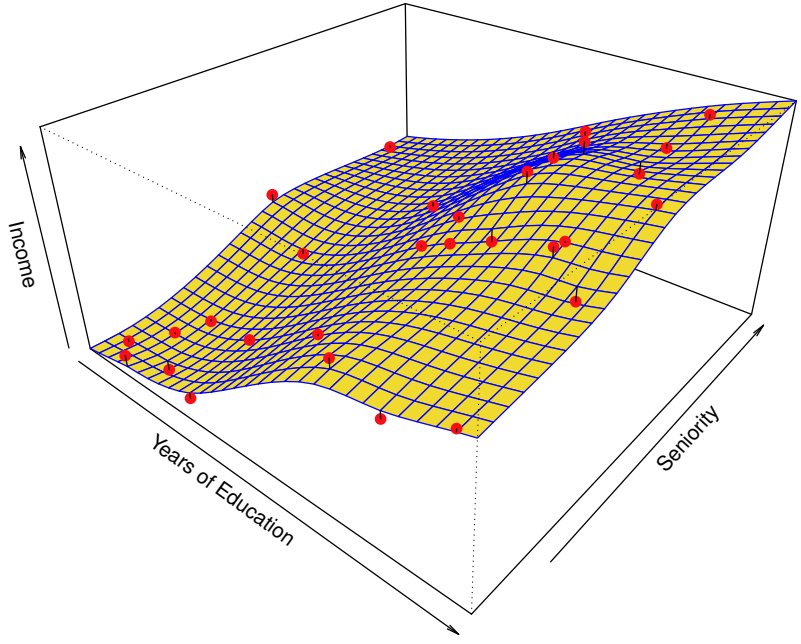

Figure 5: 비모수적 방법-Income 자료에 비모수모형 적합

Prediction accuracy vs Model interpretability

- 위 두 개념 사이의 적절한 타협

- 일반적으로 정확도가 높은 모형들의 해석가능성은 더 낮은 편임.

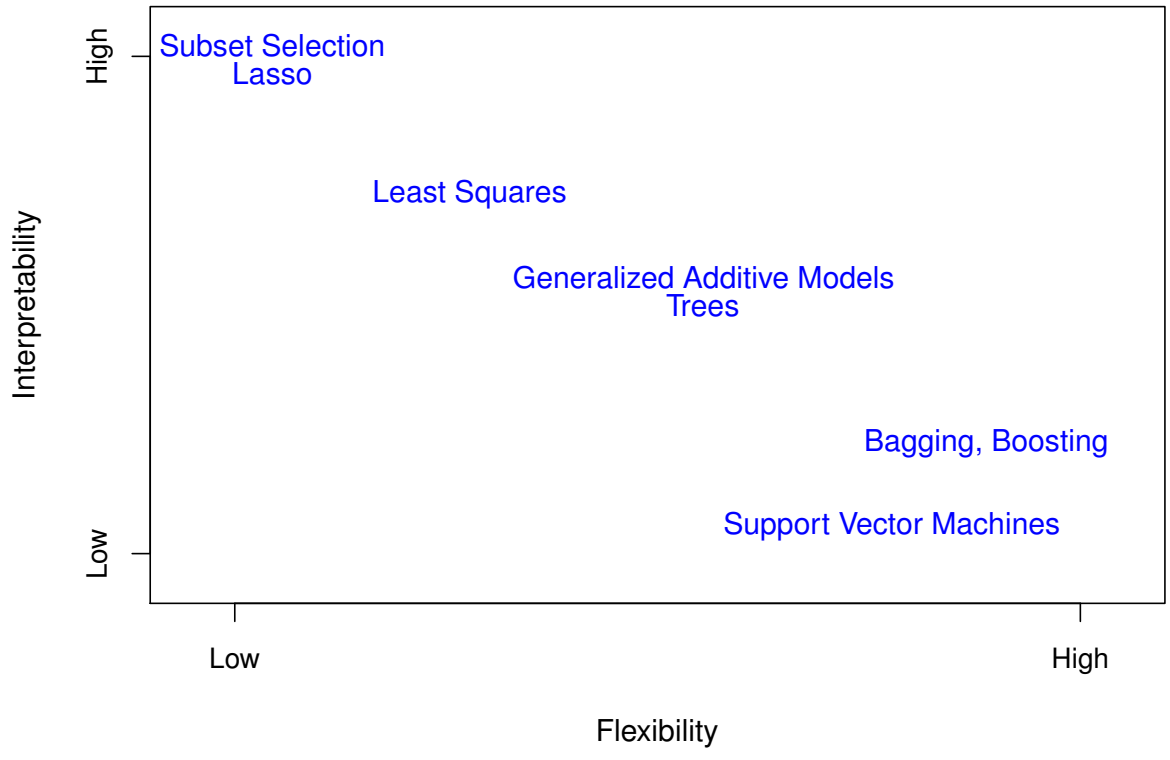

Figure 6: trade-off

지도학습 vs 비지도학습

- Label/response의 존재여부에 따라 구분.

- 지도학습 (supervised learning) : 회귀분석(regression), 분류분석(classification) 등, 반응변수 존재.

- 비지도학습 (unsupervised learning) : 군집분석(clustering), 연관분석 등은 반응변수가 없음.

회귀분석 vs 분류분석

- 반응변수의 형태에 따라 구분
- 회귀분석 : 반응변수가 양적(quantitative) 변수
- 분류분석 : 반응변수가 질적(qualitative) 변수

모형 평가

- 여러 학습법이 소개되는 이유? 모든 데이터에서 항상 가장 우수한 모형은 존재하지 않기 때문

  There is no free lunch in statistics.

  - 이론적으로 모든 데이터에서 가장 좋은 성능을 보이는 것은 존재하지 않음.

- 다양한 학습법의 평가 및 비교를 위한 측도가 필요.

- 목적에 따라 측도는 달라질 수 있다.

- 모형의 수준

  - 대범주: 회귀모형, 분류모형

  - 소범주: 단순선형회귀, 같은 구조를 가지는 딥 네트워크

  - 세부범주: 같은 (소범주) 모형 안에서 계수가 다른 것들을 구분

평가 측도 (Measuring the Quality of Fit) : 회귀모형

- 평균제곱오차 (MSE : mean squared error)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

- 훈련(training) MSE vs 평가(test) MSE

  - 훈련 데이터는 추정에 사용되며 평가 데이터는 예측의 정확도를 위해서 훈련 데이터와 구별되는 데이터이다.

- 훈련자료의 MSE를 가능하면 최소화시키는 학습법을 찾는 것이 항상 좋은가?

  - 훈련은 훈련자료의 MSE를 줄이는 것을 목표로 하면서도 평가 데이터에서도 높은 성능을 보여야 한다.

과적합 문제 (Overfitting)

- 훈련 MSE를 최소화시키는 모형이 평가 MSE를 줄이는 데 항상 유리한 것은 아니다.

- 훈련 MSE가 과도하게 작아지는 모형 적합시, 평가 MSE가 증가하는 현상이 발생하기도 한다.

- 이는 모형이 현 훈련 자료의 적합에 지나치게 집중되어 예측에의 변동성을 증가시키기 때문으로, 이를 일컬어 과적합 (over-fitting) 되었다고 말한다.

예시

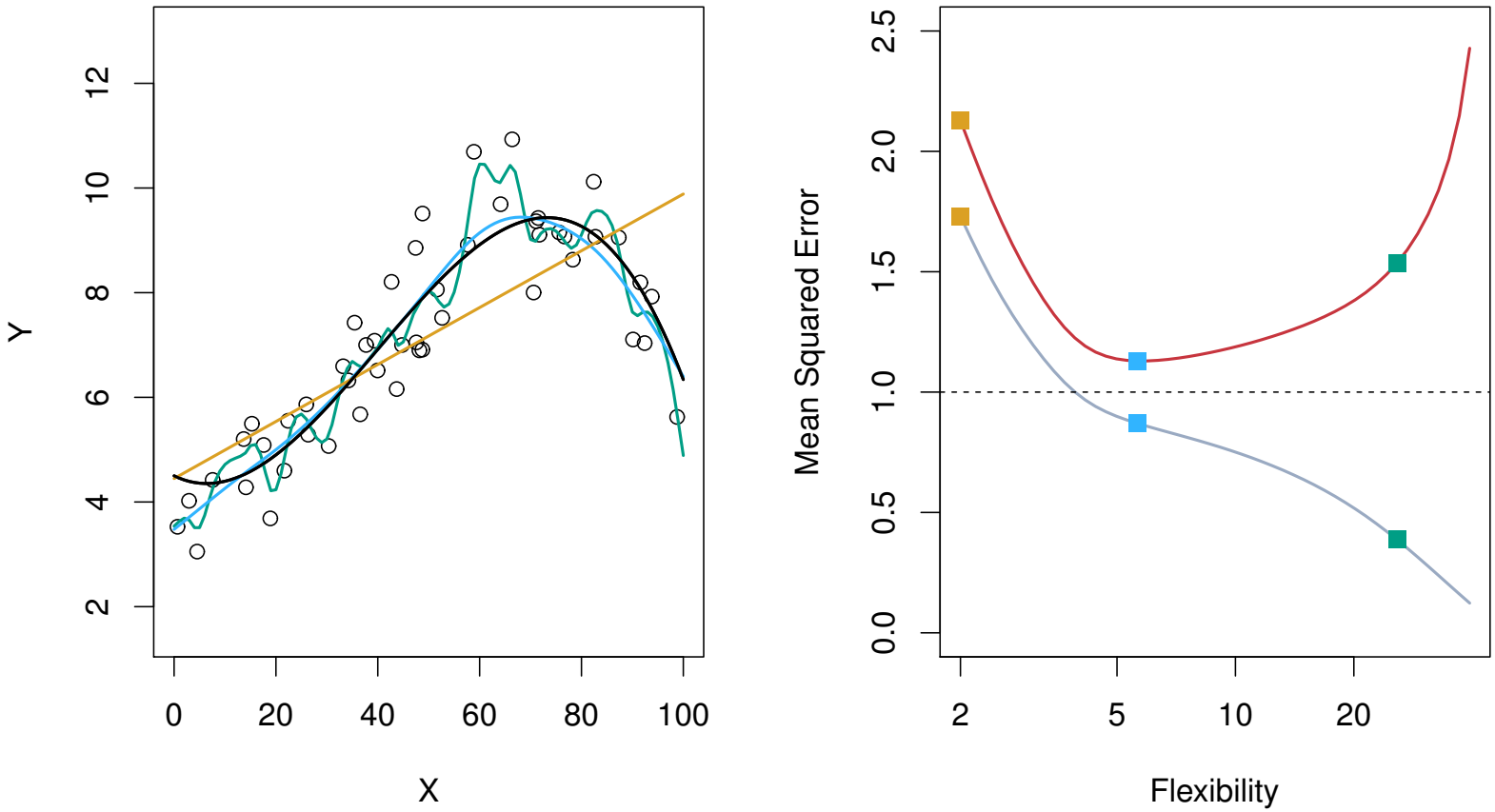

Figure 7: 적합모형과 MSE

예시

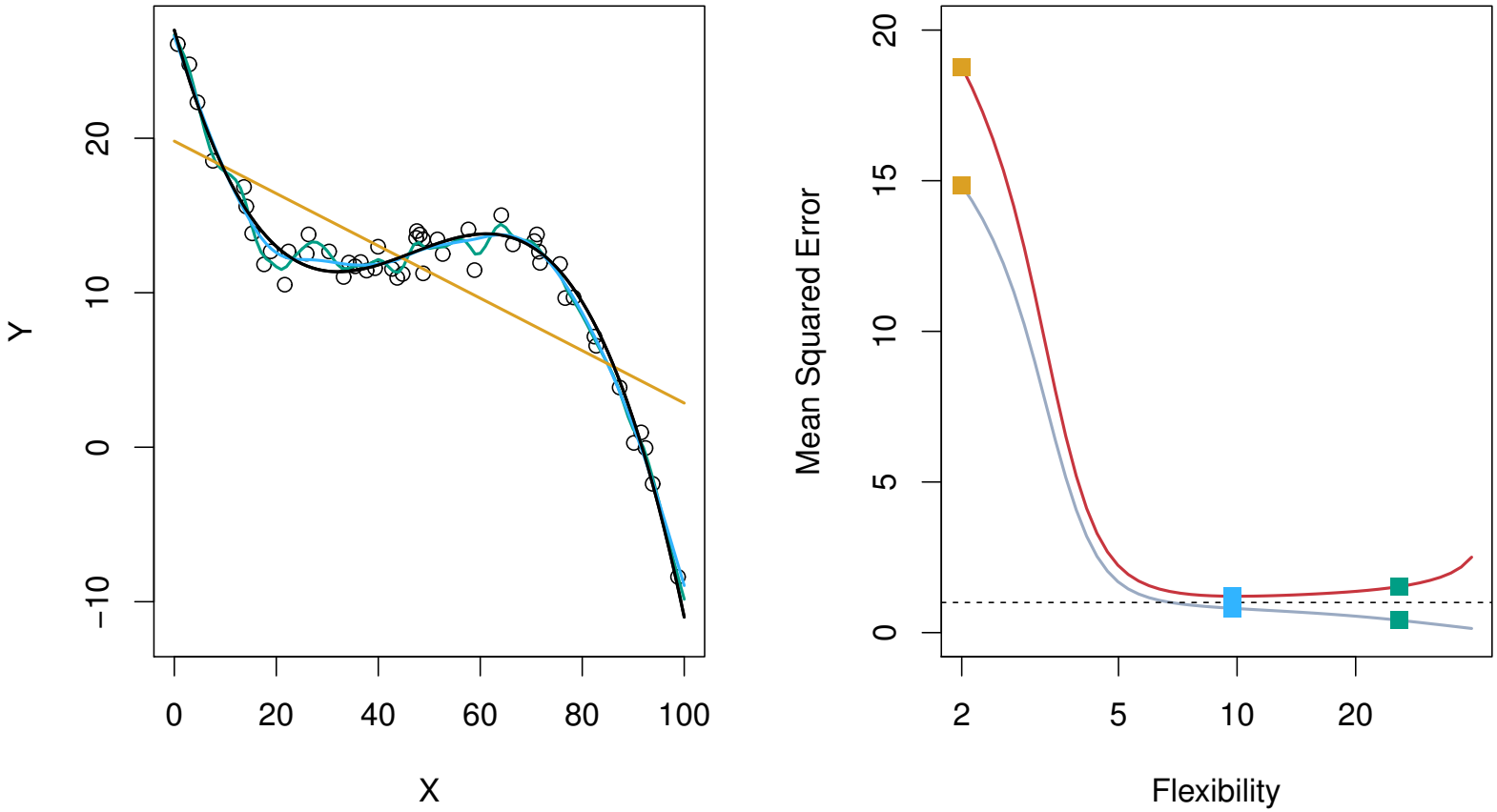

Figure 8: 적합모형과 MSE

Bias-Variance trade-off

- 평가 MSE의 기대값의 분해

$$E(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0)) + Bias(\hat{f}(x_0))^2 + Var(\epsilon)$$

- 모형의 분산 : 훈련자료의 변화에 $\hat{f}$가 얼마나 민감하게 변화하는가?

- 모형의 편의 : 실제자료를 모형으로 얼마나 가깝게 근사할 수 있는가?

- 주어진 데이터의 크기가 같을 때 더 유연한(복잡한) 모형은 더 큰 분산을 가지고, 더 단순한 모형일 수록 큰 편의를 가진다.

Bias-Variance trade-off

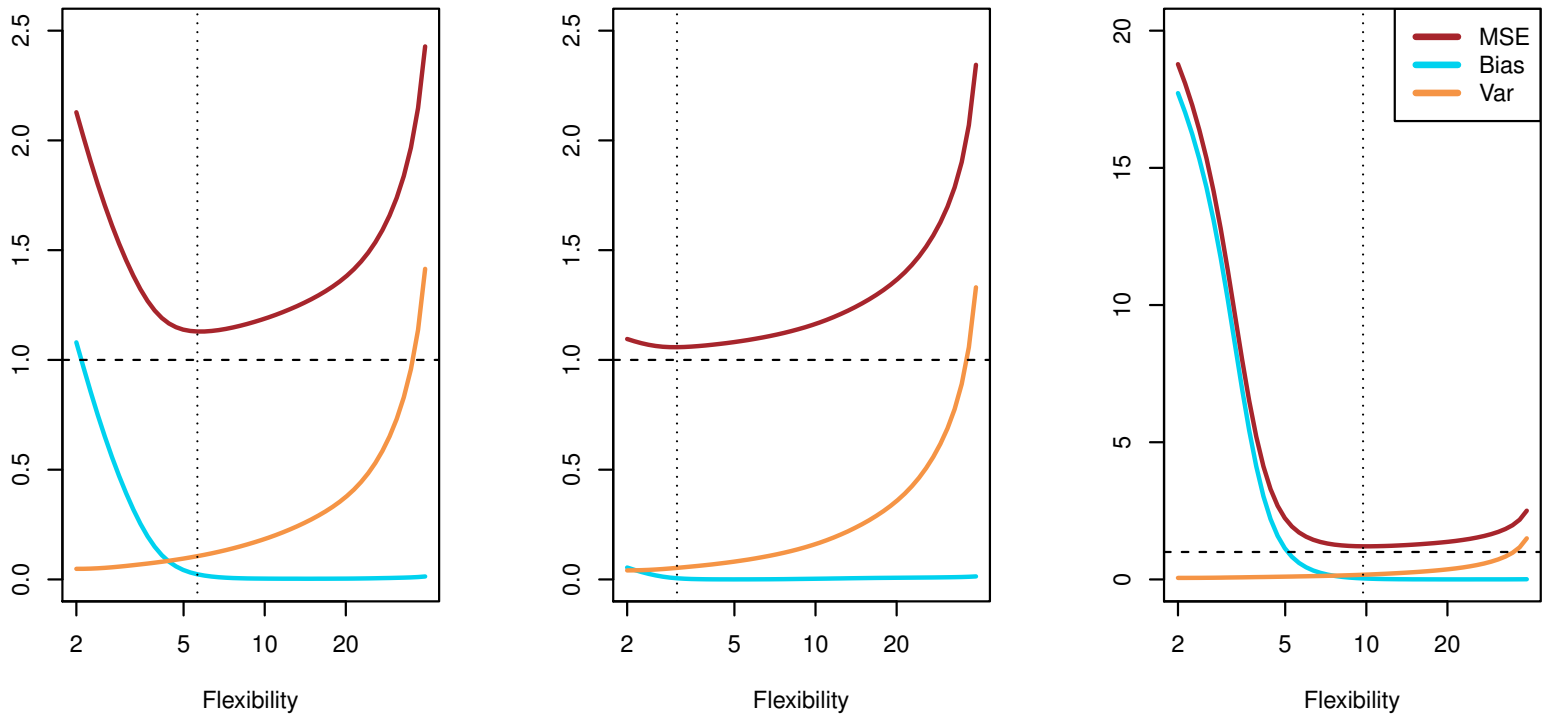

Figure 9: 여러 자료에서 모형의 편의(blue), 분산(green), MSE(red)

평가 측도 (Measuring the Quality of Fit) : 분류모형

- 오류율 (error rate)

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

- 평가자료에서의 오류율을 최소화하는 것이 목표가 된다.

- 다음을 최대화하는 방식으로 분류하는 것이 평가자료의 오류율을 최소화시킴이 알려져 있다.

$$Pr(Y = j|X = x_0)$$

베이즈 분류기 (Bayes Classifier)

- 앞 페이지의 분류기를 베이즈 분류기라 한다.

- 이진분류(1또는 2)의 경우

$$Pr(Y = 1|X = x_0) > 0.5$$

  를 만족할 때, 1로 분류하는 것을 의미한다.

- 베이즈 분류기를 실제로 얻는 것은 불가능하다 (모집단의 조건부 분포를 모르는 것이 일반적이기 때문).

- 많은 분류방법들이 위 조건부 확률의 추정에 기반하고 있다.

- 베이지 분류기를 얻는 것은 최대 확률의 값을 정확히 몰라도 가능하기 때문에, 확률의 적합성 문제가 발생한다.