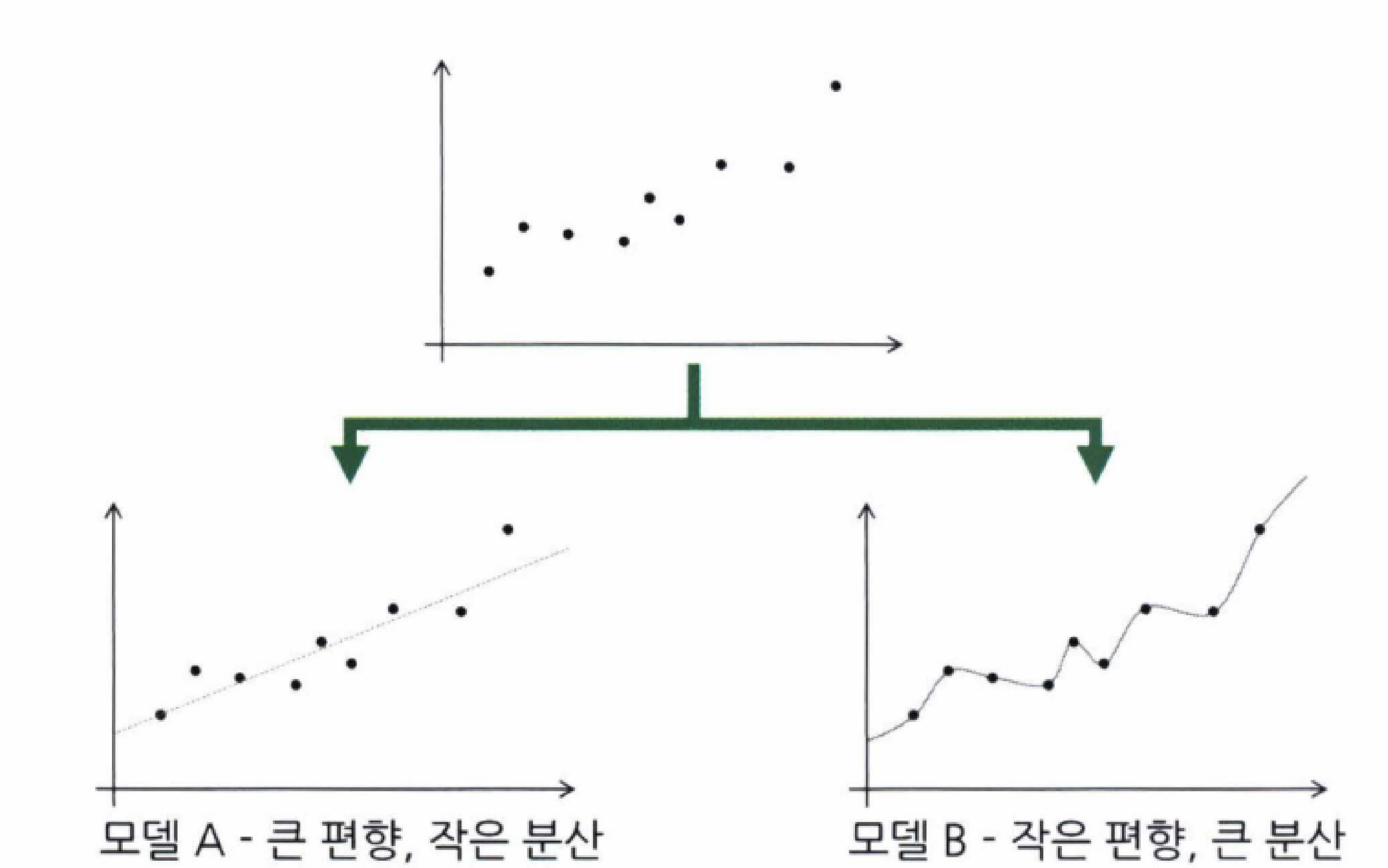
편향과 분산

- 편향은 예측값들이 정답과 일정하게 차이가 나는 정도를 의미
- 분산은 주어진 데이터 포인트(예를 들어 평균)에 대한 모델 예측의 가변성을 의미

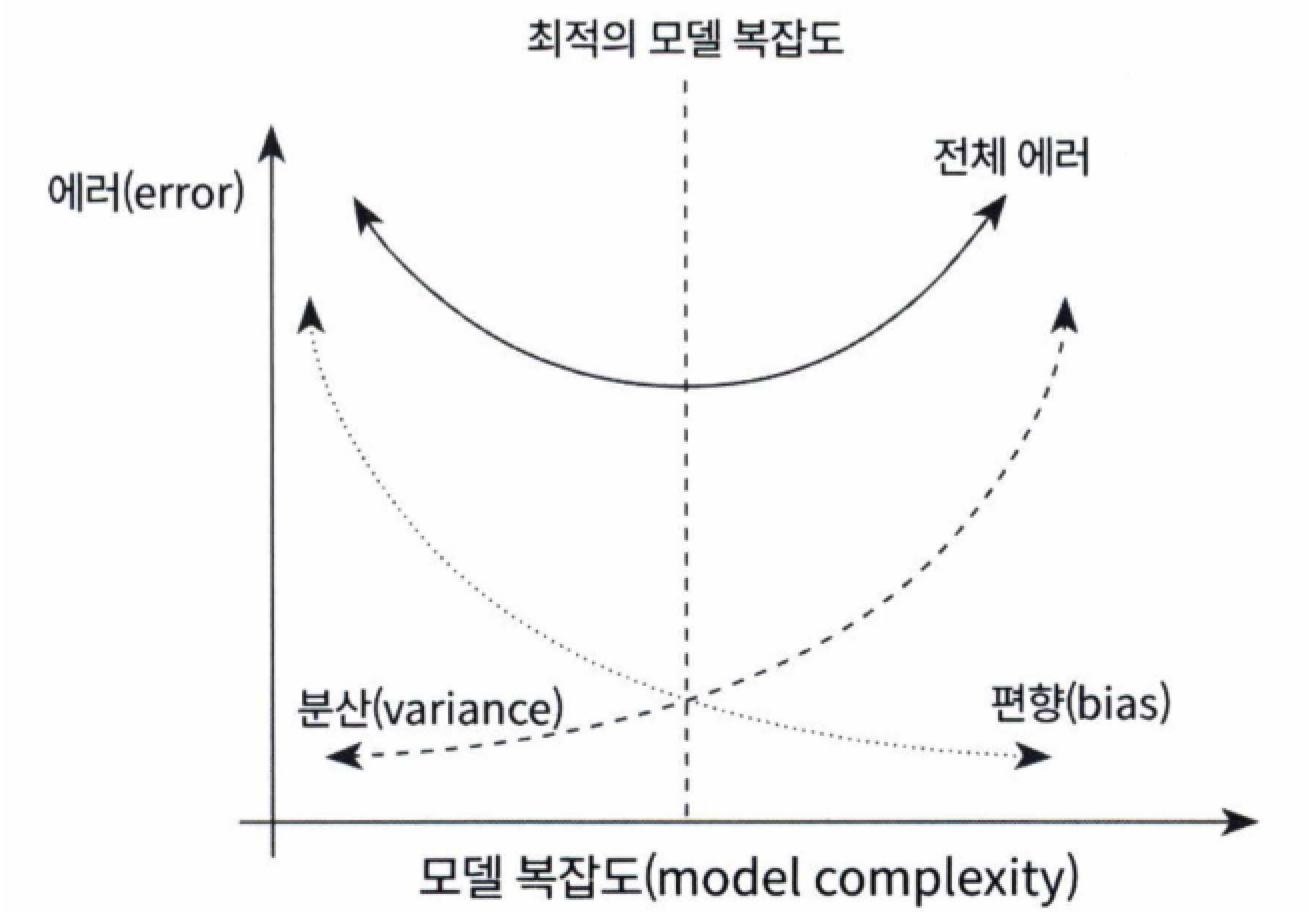


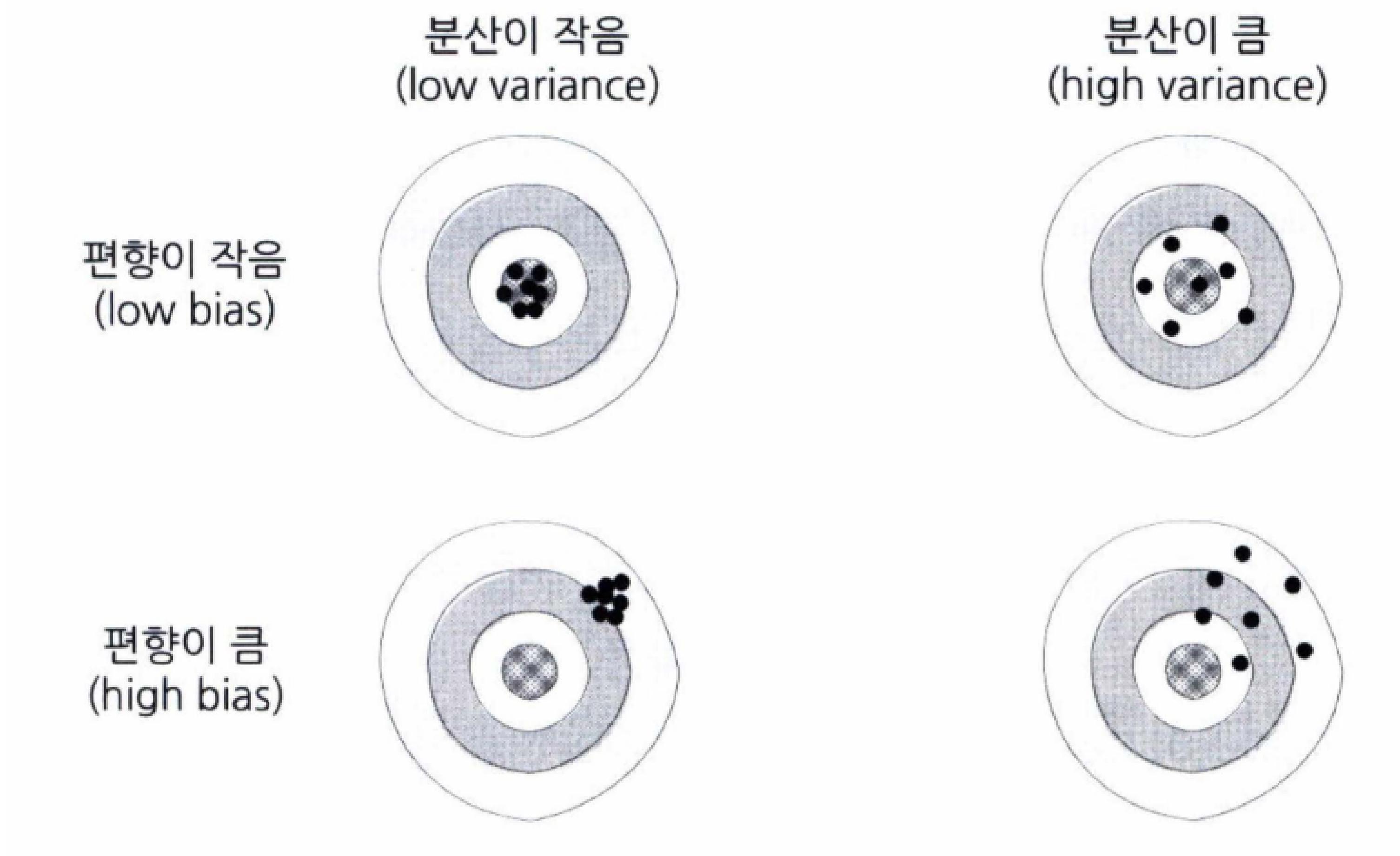
- 위의 그림처럼 모델 A, B 모델을 학습 시켰다고 하자
- A 모델
  - 예측값을 나타내는 선이 단순하게 직선으로 되어 있다.
  - 그렇기 때문에 정답과 예측값의 차이가 클수밖에 없다.
  - 그래서 이는 편향이 크다고 한다.
  - 일반화가 잘되어 있으므로 새로운 값에 대한 예측값이 일정한 패턴을 나타낸다.
  - 예측 결과를 보면 예측한 값들의 변동성이 적다
- B 모델
  - 예측값이 정답과 완벽하게 일치한다.
  - 따라서 모델 B는 편향이 매우 작다
  - 하지만 예측값이 매우 요동치고 있기 때문에 새로운 데이터를 예측하면 예측한 값과 정답의 차이가 모델 A보다 오히려 더 커질 수 있다.
  - 이를 분산이 크다라고 한다.

Trade-Off

- 편향과 분산은 트레이드 오프 관계

- 예측이나 분류 모델을 만들 때 주어진 학습 데이터에 잘 맞도록 모델을 만들수록 편향은 줄어들고 분산은 증가할수 밖에 없음
- 이 둘 간의 균형을 잘 맞춰 상황에 맞게 최적의 모델을 만드는 것이 데이터 과학자의 역할





- 이상적인 모델은 왼쪽 상단에 있는 그림이지만 현실적으로 힘듬
- 일반적인 머신러닝 모델은 오른쪽 위의 그림 형태를 갖고 있음
- 아래쪽 왼쪽은 과소적합된 모델이며, 아래쪽 오른쪽은 과적합된 모델이다.

과적합 해결방법

- 학습 데이터 양을 증가
- 배치정규화(Batch Normalization)
- 모델의 복잡도 줄이기
- 드롭아웃(Drop-out)
- 가중치 규제(Weight Regularization)

가중치 규제

- 모델의 손실 함수의 값이 너무 작아지지 않도록 특정한 값(함수)를 추가
- 가중치 값이 과도하게 커져서 일부 특징에 의존하는 현상을 방지하여 데이터의 일반적인 특징(일반화, Generalization)을 잘 반영하도록 함

L1 규제

L1 Norm

- 맨하탄 거리(Manhattan distance), 맨하탄 택시
- 택시가 도시의 블록 사이를 이동해 다른 지점으로 이동하는 것과 같이 표현
- 특정 방향으로만 움직일 수 있는 조건이 있는 경우, 두 벡터 간의 최단 거리를 찾는데 사용

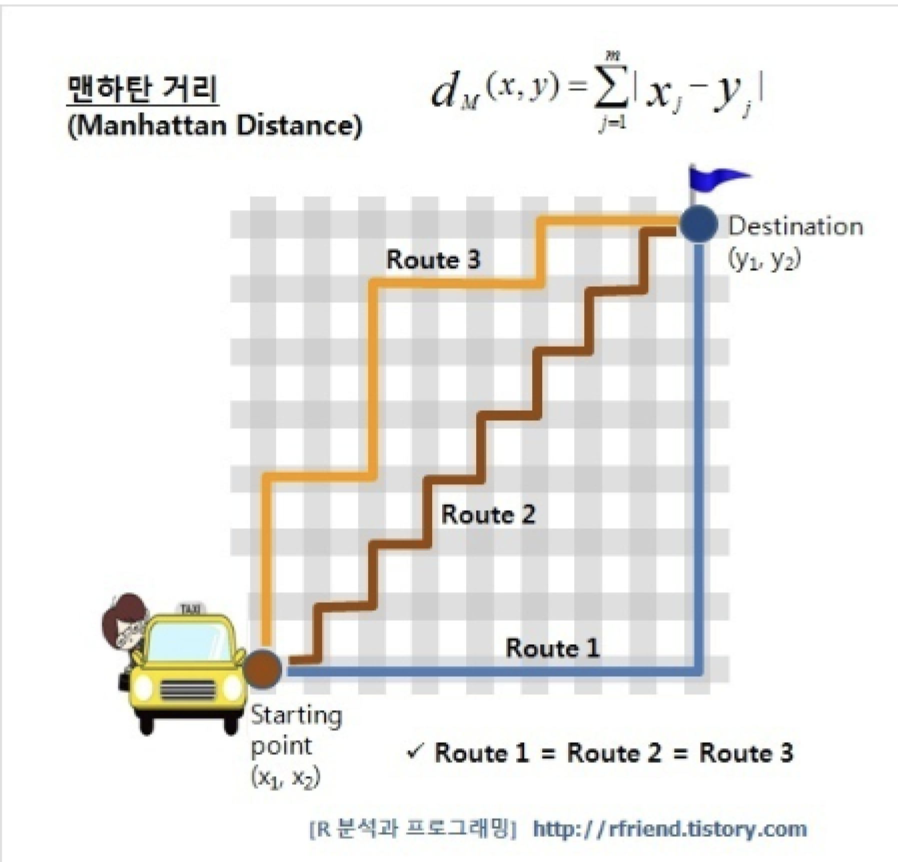
$$d = | a_1 - b_1 | + | a_2 - b_2 |$$

L1 norm (Manhattan distance)

$$L = \sum_{i=1}^{n} |y_i - f(x_i)|$$

L1 Loss Function

- L1 loss는 실제 값과 예측 값 오차들의 절대값에 대한 합을 나타냄



- 특징 :
  - 다른 점으로 이동하는데 다양한 방법이 존재
  - 절대값을 사용하기 때문에 L2에 비해 이상치에 강건함(robust)
  - 수식에 절대값이 있기 때문에 특정 특성을 0으로 처리하는 것이 가능하기 때문에 남아 있는 변수들을 선택해서 모델을 만들수 있음(변수 선택의 역할)
  - 특성이 많은 데이터셋에서 유용함
  - 0에서는 미분이 불가능하기 때문에 경사기반 알고리즘에서는 주의가 필요
  - 규제효과가 L2 대비 떨어짐

L1 Regularization

- 학습 데이터에 알고리즘에서 모델이 과적합이 되지 않도록 손실 함수에 특정한 규제 함수를 추가하여 손실 함수가 너무 작아지지 않도록 가중치에 페널티를 주는 기법

Lasso

L2 규제

L2 Norm

- 유클리디안 거리
- 두 점 사이의 최단 거리를 측정할 때 사용

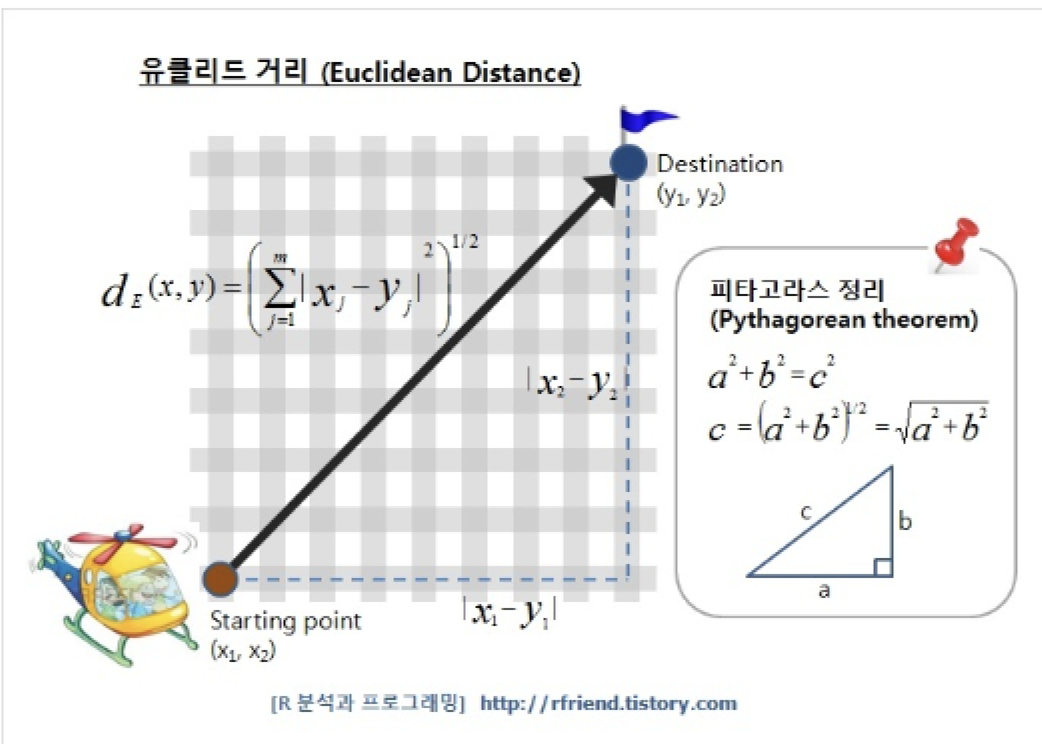
$$d = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$$

L2 norm (Manhattan distance)

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

L2 Loss Function

- L2 loss는 실제 값과 예측 값 오차들의 제곱의 합으로 나타냄



- 특징
  - 다른 점으로 이동하는데 단 한가지 방법만 존재
  - 오차의 제곱을 사용하기 때문에 L1보다 이상치에 더 큰 영향을 받음
  - L2는 가중치의 크기가 너무 커지는 것을 방지하는 가중치 감소가 가능
- L2 norm이 가중치에 대한 규제에 조금 더 효과적이기 때문에 학시 시 더 좋은 결과를 얻을 수 있어 L1 norm 보다 많이 사용

L2 Regularization

Ridge