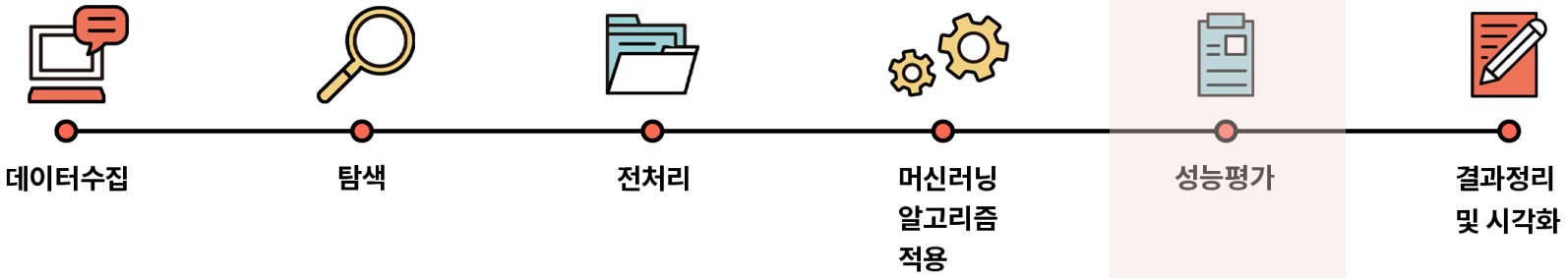
모형진단과 교차검증

Evaluation & Cross Validation

기계 학습을 이용한 문제해결 과정

- **모형 진단을 통한 모형 최적화 단계**



- 모형 진단(Evaluation)
- 교차 검증(Cross Validation)

모형진단

- **모형진단이란?**
  - 여러 후보모형 중 가장 좋은 모형을 선택하기 위한 성능비교 과정
  - 과대 적합(Overfitting) 혹은 과소 적합(Underfitting)을 피하여 모형 선택
- **모형진단 방법**
  - 모형 학습: 학습데이터를 통해 모형 학습
  - 모형 검증: 검증데이터를 통해 모형 최적의 하이퍼파라미터 선택
  - 모형 평가: 학습이나 검증에 이용한 적이 없는 시험데이터의 정밀도 사용
- **모형진단을 위한 자료의 분할 (고정 자료 분할법)**

일반 데이터 ➔ 70:15:15

대용량 데이터 ➔ 90:5:5

모형진단

- **초모수 (하이퍼파라미터) 결정법**
  - 학습데이터를 이용하여 모형 학습 후 검증데이터를 통한 초모수 조절
  - 검증데이터의 성능은 학습데이터의 성능과 같거나 근접하도록 결정
    - 성능: 손실 함수 값, 정밀도
  - 초모수 예시
    - SVM의 c, 커널 함수에 포함된 파라미터, 모형 규제를 위한 λ

- **고정 자료 분할의 단점**
  - 선택된 검증데이터에 의존하는 단점이 있음
  - 자료의 크기가 작을 경우 모형진단의 신뢰성이 떨어짐
  - 단점 극복을 위해 K-분할 교차검증(K-fold cross validation)이 제안됨

K-분할 교차검증

- **K분할 교차 검증이란?**
  - 학습데이터의 K분할을 통해 K번 모형진단을 통해 초모수를 결정하는 것
  - 가장 우수한 성능을 보이는 초모수로 모형의 초모수 결정
  - 일반적으로 K=10, 대용량자료일 경우 K=5

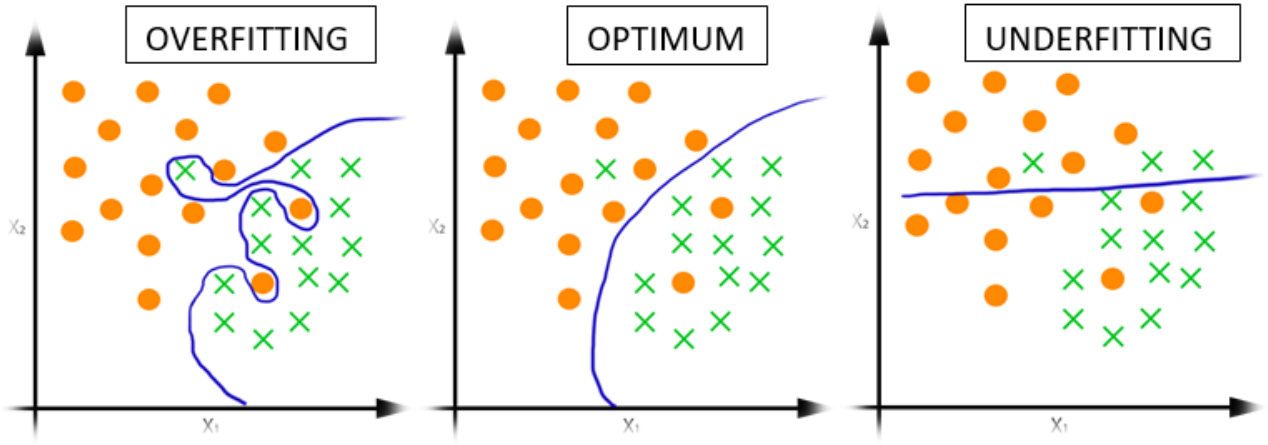
K=10인, 교차검증 예시

모델 최적화

- **과적합 문제**
  - **과대적합(overfitting)이란?**
    - 모델이 학습 데이터에 너무 잘 맞지만 **일반화(generalization)**가 떨어지는 상황

일반화(generalization)란?
테스트 데이터에 대한 높은 성능을 갖추는 것

  - **과소적합(underfitting)이란?**
    - 모델이 너무 단순하여 데이터에 내재된 구조를 학습하지 못하는 현상

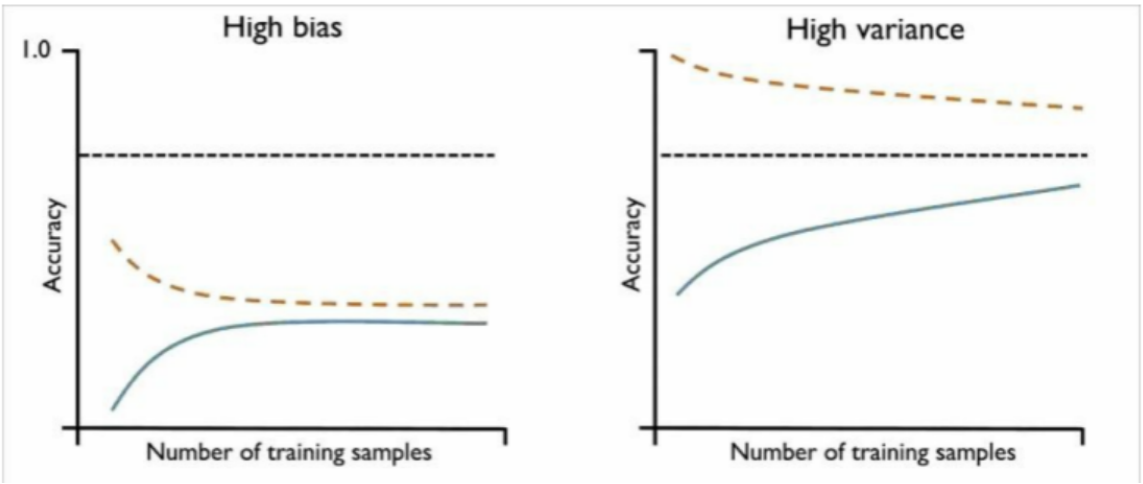


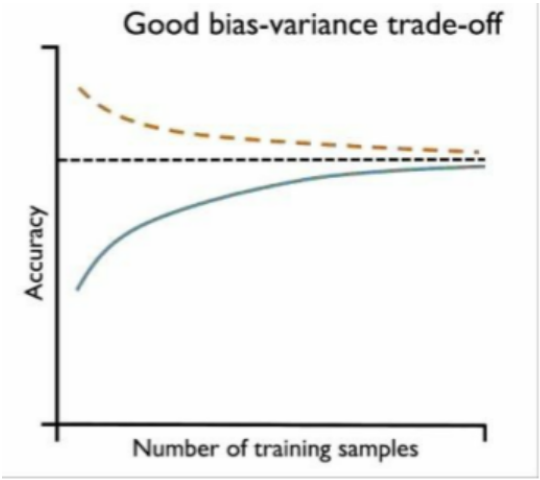
모델 최적화

- **과적합 문제 해결**
  - **과대적합 해결방법**
    - 학습 데이터 추가 수집
    - 모델 제약 늘리기: 규제 하이퍼파리마터 값 늘리기
    - 학습 데이터 잡음을 줄임 (오류 수정 및 이상치 제거)
  - **과소적합 해결방법**
    - 특성변수 늘리기
    - 모델의 제약 줄이기: 규제 하이퍼파리마터 값 줄이기
    - 과적합 이전까지 충분히 학습하기

모델 최적화

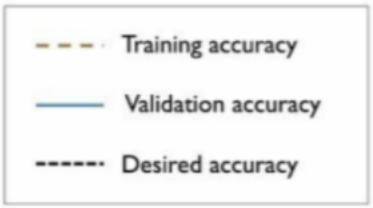
- **과대적합/과소적합 판단하기**
  - **학습 곡선(Learning Curve)의 편향과 분산 분석**
    - 샘플 데이터의 수에 따른 정확도 변화





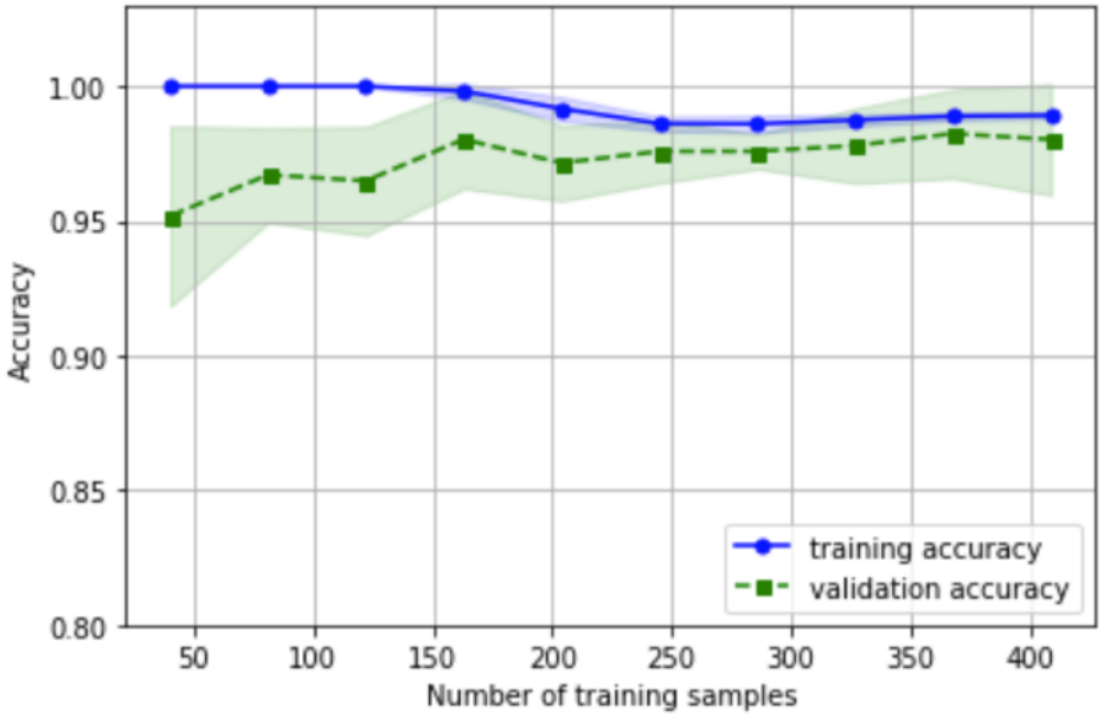
- Underfitting 사례
  : 모델 파라미터 수 늘리기
  : 규제 강도 줄이기

- Overfitting 사례
  : 학습데이터 보강하기
  : 규제 강도 높이기



모델 최적화

- **과대적합/과소적합 판단하기**
  - 샘플 데이터의 수에 따른 정확도 변화
    - 아래의 예) 250개 이상의 샘플을 사용할 때 모델이 잘 작동함



모델 최적화

- **과대적합/과소적합 판단하기**
  - 매개변수에 따른 정확도 변화
    - 로지스틱 회귀의 매개변수 C (규제 강도와 반비례)