Preclass 02:
Introduction to Machine Learning II

[SCS4049] Machine Learning and Data Science

Seongsik Park (s.park@dgu.edu)

Department of Artificial Intelligence, Dongguk University

ML의 문제점: 불충분한 학습 데이터

- 복잡한 문제인 경우 알고리즘 보다는 데이터가 더 관건
  → 데이터는 다다익선
- 대부분의 데이터셋은 소규모/중규모 → 알고리즘이 중요

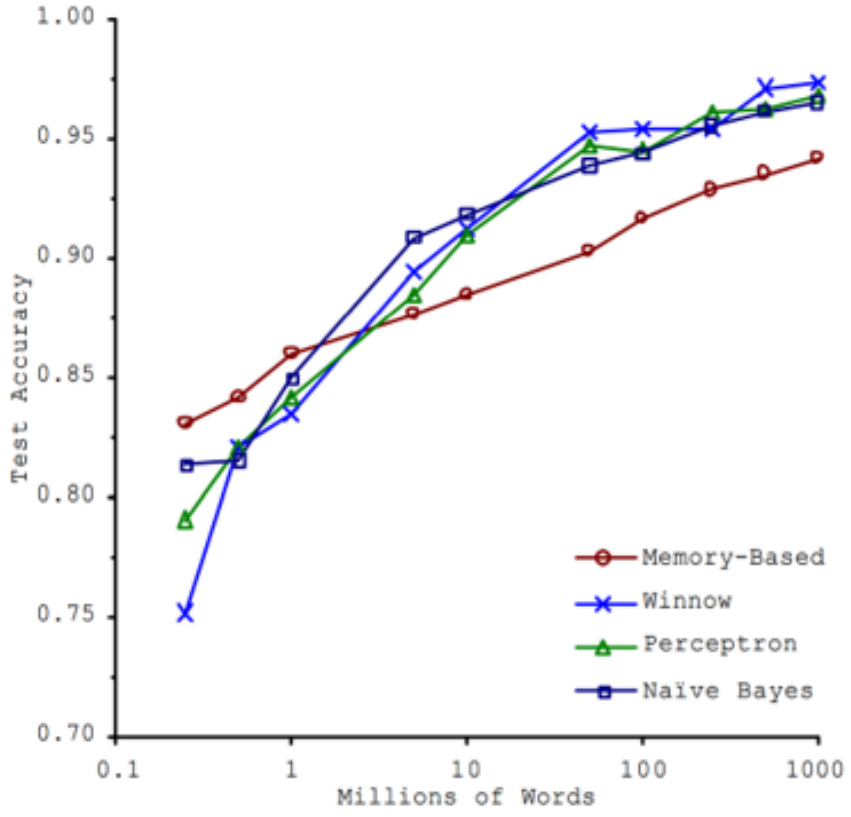


Figure 1: Dataset vs. algorithm

ML의 문제점: 학습 데이터의 비대표성

- 일반화시키려고 하는 모집단을 잘 대표하는 학습 데이터셋을 사용하는 것이 중요
- 행복도 예제에서 누락된 국가들을 포함시키면 회귀 직선이 극단적으로 변화 → 대표성 부족
- 데이터셋이 너무 작으면 → Sampling Noise 의 영향이 큼
- 데이터셋이 크더라도 샘플링 방법이 잘못되었으면 → Sampling Bias 가 발생

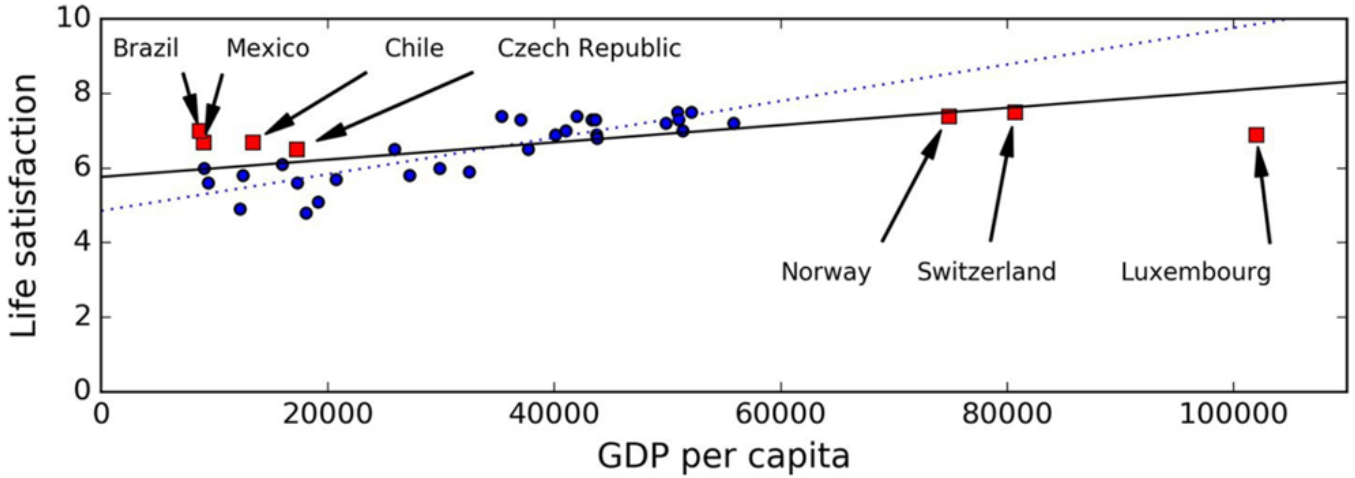
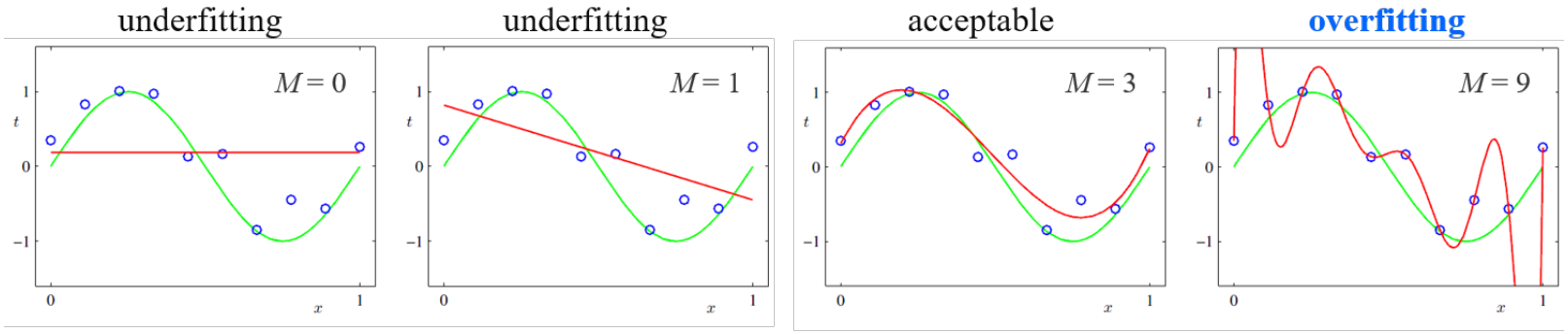


Figure 2: A more representative training sample

ML의 문제점: 과적합과 미적합

- 학습 데이터에 대한 과적합 overfitting
  - 모델이 학습 데이터에 대해서는 좋은 성능을 보이지만, 처음 보는 새로운 데이터에 대해서 잘 일반화하지 못함
  - 학습 데이터의 양과 노이즈 정도에 비해서 모델이 너무 복잡할 때 (too high capacity)
  - 모델은 경향(trend)로부터 일반화하는 것을 학습하기 보다는 학습 데이터를 기억하기 시작함
- 학습 데이터에 대한 미적합 underfitting
  - 모델이 너무 단순 → 학습 데이터에 대해서조차 부정확
  - 과적합의 경우보다는 발견하고 조치하기가 용이



[Source: Christopher M. Bishop, "PRML", 2006]

ML의 문제점: 과적합과 미적합

- 과적합에 대한 대책
  - 더 많은 데이터를 수집
  - 더 단순한 모델을 사용
  - 데이터의 속성을 줄임
  - 정규화 regularization
- 미적합에 대한 대책
  - 파라미터가 더 많은 더 강력한 모델을 선택
  - 특징 공학을 통해 더 우수한 특징을 사용
  - 모델의 제약 조건을 완화
  - 정규화 파라미터 값을 완화

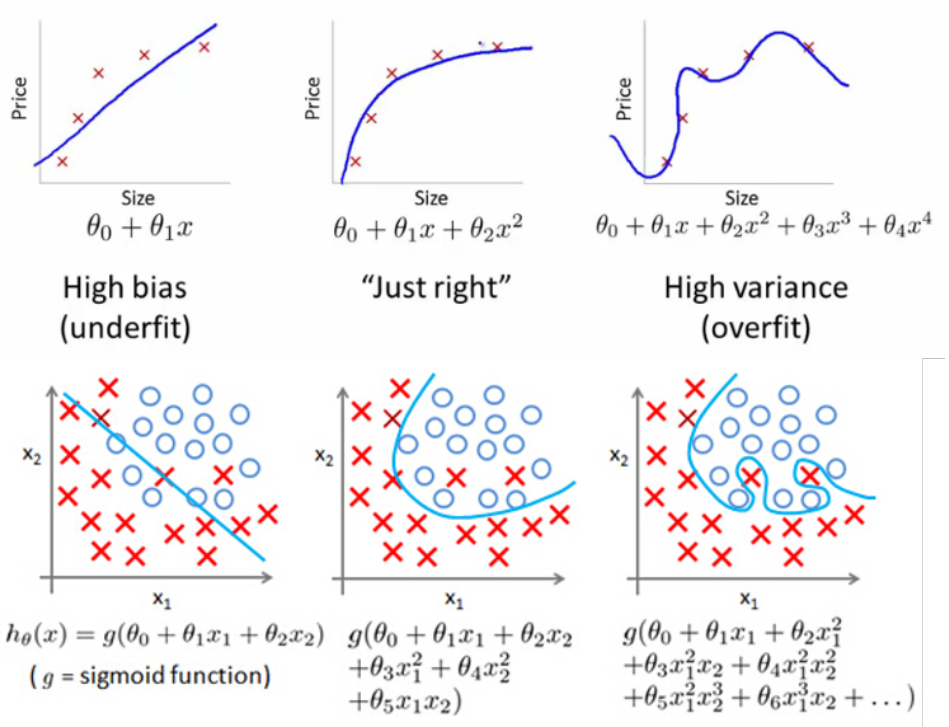


Figure 3: Overfitting the training data

테스트와 검증

- 테스트의 목적: 일반화 오차(Generalization error)가 관건
- 테스트 (Test)
  - 실무 적용 직전에 모델 성능을 평가하는 것
  - 테스트 데이터셋: 전체 데이터셋에서 학습 전에 별도로 떼어 놓은 데이터셋 (전체 데이터의 약 20%)
  - 훈련(학습) 도중에 절대 절대로 답(Test dataset)을 들여다 보면 안됨
- 검증 (Validation)
  - 모델 선택 및 미세 조정을 위해 실시
  - 검증 데이터셋 (hold-one-out set) : 나머지 학습 데이터셋의 약 20%
  - 한번 검증으로는 보통 불충분할 수도 있음
  - 그리고, 아까운 데이터가 낭비로 볼 수 있음
  - Cross validation (교차 검증) 필요

테스트와 검증

- 교차검증 (Cross-Validation)
- 테스트 (Test)
  - 훈련 데이터셋을 분리시켜 교대로 검증에 사용하는 방법
  - 평균 성능과 성능 분산을 통해 모델을 검증
  - 모델 선택 및 미세 조정

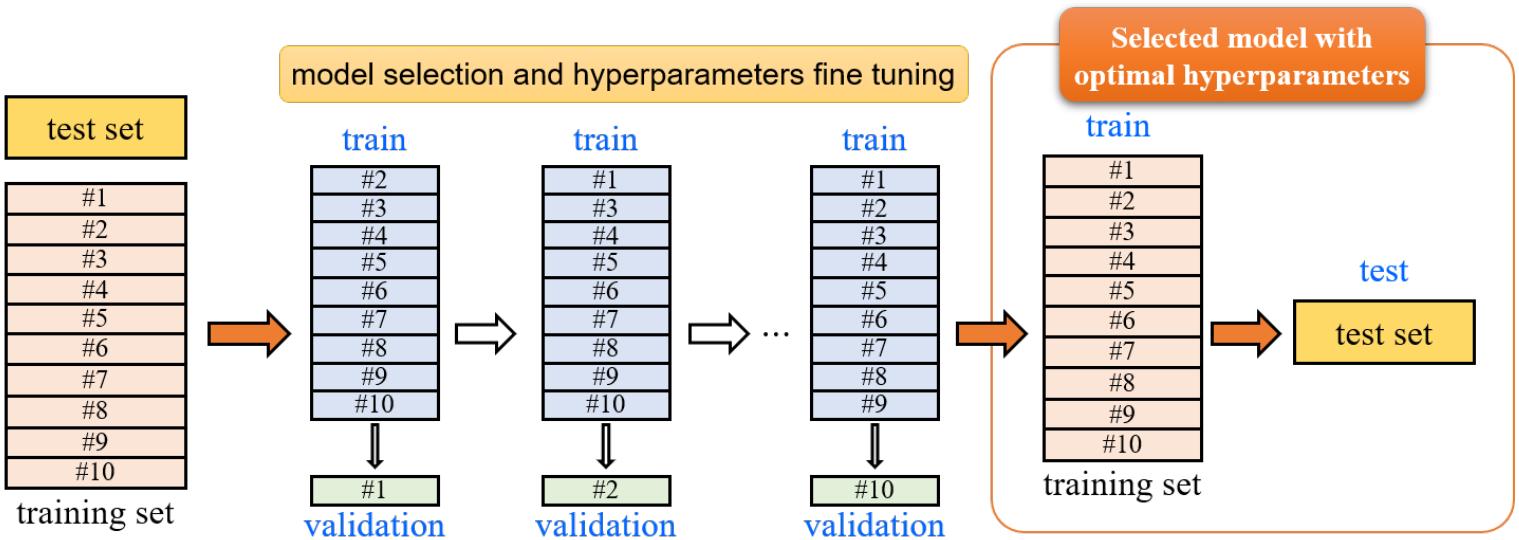


Figure 4: Cross-validation procedure

No free lunch theorem

- 모델은 관측의 단순화된 버전
  - 무엇을 고려하고 무엇은 무시할지 결정 → 하나 또는 이상의 가정
- 세상에 공짜는 없다 (NFL Theorem)
  - 데이터에 대해서 전혀 가정을 하지 않는다면 어떤 모델을 다른 모델보다 선호할 근거 없음
  - 데이터에 따라 선형 모델이 좋은 경우도 있고 신경망이 좋은 경우도 있으나, 미리(a priori) 알 수 있는 방법은 없음
  - 따라서 유일한 방법은 모든 모델에 대해서 평가해보는 것 → 그러나 이것은 현실적으로 불가능
  - 그러므로, 실제로는 데이터에 대해서 합리적인 몇가지 가정을 하고 몇 개의 모델을 평가할 뿐
  - 특정한 문제에 최적화된 알고리즘이 다른 문제들에서는 그렇지 않다는 것을 수학적으로 증명한 정리

Appendix

Reference and further reading

- “Chap 1” of A. Geron, Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow
- “Chap 1” of C. Bishop, Pattern Recognition and Machine Learning