회귀

목차

- 회귀 문제란?
- Boston Housing Dataset
- Diabetes Progression Dataset

회귀 문제란?

- 예측하고자 하는 값의 종류가 숫자인 경우

- 예) 온도에 따른 레모네이드의 판매량

날짜	요일	온도	판매량
2020.1.3	금	20	40
2020.1.4	토	21	42
2020.1.5	일	22	44
2020.1.6	월	23	46
2020.1.7	화	24	48
2020.1.8	수	25	

AIoT

Boston Housing Dataset

- 13가지의 지표로 집값 예측
- 회귀 문제를 다루는 데 대표적인 데이터
- Train data: 456 rows, Test data: 50 rows
- 데이터 출처: Carnegie Mellon University

Boston Housing Dataset

- Column 설명
- CRIM: 마을 별 1인당 범죄율
  ZN: 25,000 평방 피트를 초과하는 거주지역의 비율
  INDUS: 비소매상업지역이 점유하고 있는 토지의 비율
  CHAS: 찰스강에 대한 더미변수
  NOX: 10ppm 당 농축 일산화질소
  RM: 주택 1가구당 평균 방의 개수

Boston Housing Dataset

- AGE: 1940년 이전에 건축된 소유 주택의 비율
  DIS: 5개의 보스턴 직업센터까지의 접근성 지수
  RAD: 방사형 도로까지의 접근성 지수
  TAX: 10,000 달러 당 재산세율
  B: 1000(Bk-0.63)^2 (Bk=마을 별 흑인의 비율)
  LSTAT: 모집단의 하위계층의 비율

- MEDV: 본인 소유의 주택가격(중앙값) ($1,000 단위)

Boston Housing Dataset

- Baseline
  Model: Linear(13,16)->Sigmoid()->Linear(16,1)
  Loss: MSELoss()
  Optimizer: SGD(lr=0.001)
  Num epochs: 100
  Early stop: None

- Result: 25.4975 (MSELoss)

Boston Housing Dataset

- 성능을 향상시키기 위해

1. 모델의 구조 변경(layer 추가 또는 제거, 활성 함수 변경)
2. Learning rate 조절
3. 최적화 알고리즘 변경(SGD or Adam, …)
4. num_epochs 조절
5. 학습 데이터 변경

등을 적용하여 계속하여 학습, 성능이 향상된다면 원인 분석

Diabetes Progression Dataset

- 10가지의 지표로 당뇨병 진행도 예측
- Boston Housing Dataset과 더불어 회귀 문제에 자주 등장하는 데이터
- Train data: 398 rows, Test data: 44 rows
- 데이터 출처: North Carolina State University

Diabetes Progression Dataset

- Column 설명
  Age: 나이
  Sex: 성별
  Body mass index: BMI(=체중(kg)/키(m)^2)
  Average blood pressure: 평균 혈압
  S1~S6: 혈청에 관련된 6가지 지표(tc,ldl,hdl,tch,ltg,glu)

- Target: 당뇨병 진행도

Diabetes Progression Dataset

- 해당 데이터셋의 독립 변수는
  mean=0, sum(x^2)=1로 표준화 되어있음.

Diabetes Progression Dataset

- Baseline
  Model: Linear(10,16)->Sigmoid()->Linear(16,1)
  Loss: MSELoss()
  Optimizer: SGD(lr=0.001)
  Num_epochs: 100
  Early stop: None

- Result: 3886.121217 (MSELoss)

Diabetes Progression Dataset

- Boston Dataset과 마찬가지로 성능을 향상시키기 위해

1. 모델의 구조 변경(layer 추가 또는 제거, 활성 함수 변경)
2. Learning rate 조절
3. 최적화 알고리즘 변경(SGD or Adam, …)
4. num_epochs 조절
5. 학습 데이터 변경

등을 적용하여 계속하여 학습, 성능이 향상된다면 원인 분석

5장 Preview

- 주어진 데이터(X)가 어떤 부류에 속할 지 예측하는 Classification

- 다음 두가지 데이터셋에 대해 실습 진행
  - Titanic Dataset: 타이타닉 호에 탑승한 승객의 생존 여부 예측
  - MNIST Dataset: 0~9까지의 손글씨 숫자 데이터 예측

References

[1] https://opentutorials.org/module/4916/28942

[2] https://wikidocs.net/49966

[3] https://wikidocs.net/49981

수고하셨습니다!