내용 정리

박경규

빅데이터

데이터의 형태에 따른 구분 :

형태가 있으며 연산 가능하고 주로 관계형 데이터베이스에 저장되는 데이터는?

텍스트, 음성, 이미지, 동영상, 음성 데이터는?

스키마와 메타데이터와 같은 형태가 있는 데이터는?

머신러닝(Machine Learning)

회귀모델(Regression)의 손실 함수

Cross Entropy는 어떤 모델에 사용하는 손실함수인가?

정답이 있는 데이터를 활용해 데이터를 학습시키는 방법?

레이블(정답)의 값들이 이산적으로 나눠질 수 있는 문제에 사용하는 모델?

레이블(정답) 없이 진행되는 학습으로, 데이터 자체에서 패턴을 찾아내야 할 때 사용하는 학습방법?

딥러닝 학습방법

딥러닝 학습의 목표는 모델에 입력값을 넣었을 때의 출력값이 최대한 정답과 일치하게 하는 것이다.

딥러닝 학습은 손실(Loss, Error)를 최소화 하는 인공신경망의 가중치(weight)와 편향(bias)을 찾는 과정

러닝 학습은 순전파(Forward Propagation)와 오차역전파(Error Back Propagation)의 반복으로 진행

실제값과 모델 결과값에서 오차를 구해서, 가중치를 재업데이트 하는 과정은?

N-gram 모델

언어 모델(Language Model) 에서 사용되는 의미 기반의 벡터 표현방식인지?.

N-gram은 장기 의존성(Long-term dependency) 이 있는지?

N=1,2,3 인 경우의 명칭

N 을 크게 증가시킬수록 성능이 개선 되는지?

텍스트 데이터 특징 추출 방법

TF-IDF는 단어 빈도-역 문서 빈도로 중요한 단어와 중요하지 않은 단어를 구분할 수 있는 방법이다.

머신러닝 알고리즘에서 텍스트 데이터를 사용하기 위해선 수치적인 표현으로 변환해야 하며,
이 수치형 표현을 특징 벡터(feature vector)라고 부른다.

BoW(Bag of Words) 는 문서에서 각 단어들의 빈도수를 사용하는 방법으로,
문서를 숫자 벡터로 변환하는 가장 기본적인 방법이다.

TF-IDF는 개별 문서에서 자주 등장하는 단어에 높은 가중치를 주고,
모든 문서에서 자주 등장하는 단어에는 페널티를 주는 방식으로 값을 부여한다.

텍스트 분류

자연어 처리 기술을 활용해 문장의 정보를 추출해서 사람이 정한 범주(Class)로 분류하는 문제텍스트 분류

데이터 로드 -> EDA -> 데이터 전처리 -> 데이터 벡터화 -> 모델링 순으로 진행한다.

데이터 전처리에서는 HTML 태그 제거, 특수문자들을 공백으로 바꾸기, 불용어 제거 작업등을 한다.

텍스트 데이터를 토큰화 하여 바로 RNN, CNN 분류모델의 입력으로 사용할 수 있다? 없다?

언어 모델(Language Model)

언어 모델은 "자연어의 법칙을 컴퓨터로 모사한 모델"을 의미한다.

이전 state 정보가 다음 state 를 예측하는데 사용되어, 순차 데이터 처리에 특화된 언어모델은?

역전파 방법인 BPTT은 어떤 언어 모델에서 사용하는가?

다음의 등장할 단어를 잘 예측하는 언어모델은 그 언어의 특성이 잘 반영된 모델이다.

LSTM 언어 모델(Language Model)

RNN은 재귀를 통한 정보전이 및 전파가 하나의 레이어로 제어되는 반면,

LSTM은 Forget gate, Input gate, Output gate를 통한 정보전이 및 전파를 제어한다.

Encoder Layer에서는 Context Vector를 만들고 Decoder Layer에서는 Context Vector 를

입력으로 출력을 예측하는 언어모델은?

장기적인 종속성을 학습할 수 있는 특수한 종류의 RNN은?

입력과 출력사이 신경망이 재귀하는 구조를 갖고 있는 언어모델은?

BERT

BERT의 입력은 토큰 임베딩, 세그먼트 임베딩, 포지션 임베딩 3개의 합으로 구성

BERT는 wiki, book data와 같은 대용량 데이터로 모델을 사전 훈련(pre-train) 시킨 후,

특정 태스크를 가지고 있는 labeled data로 전이학습(transfer learning)을 하는 모델

입력 문장(sentence)의 첫번째 토큰은?

프리트레인을 마친 임베딩은 말뭉치의 의미적, 문법적 정보를 충분히 담고 있으므로

그 자체로 다운스트림 태스크에 활용할 수 있는지?

Thank you

kgpark88@gmail.com