딥러닝 기반 핵심 산업별 빅데이터 분석

<머신러닝&딥러닝 파일럿 프로젝트>

주 제	Bike sharing PREDICT	링 크	https://www.kaggle.com/c/bike-sharing-demand
팀 명	듀란듀란	일 자	2018년 11월 29일
팀 장	민유진<ujmin0417@naver.com> 함윤선<dotstar801@gmail.com>	팀 원	유선우<ysw900524@gmail.com> 장정호<jangclod92@gmail.com>

1. 과제 개요

주제	샌프란시스코 범죄 예측
목표	Washington D.C의 자전거 대여 키오스크의 Bike Lental Data와 Weather Data를 활용하고, 이를 바탕으로 Test Data의 대여 횟수(Counts)를 예측.
과제수행효과	1. 실전데이터를 활용해, 데이터 전처리/시각화 및 데이터 분석 기법을 복습하는 기회로 삼을 수 있다 2. 주어진 데이터에 TensorFlow를 활용하여 문제해결능력을 키울 수 있다.
왜 이 주제인가?	1. 공유경제의 대표격인 공유자전거 시스템에 대한 팀 차원의 관심이 존재하였다. 2. 해당 주제를 통해 TensorFlow 실제 활용해 볼 수 있고, 이를 통해 머신러닝 기법을 통한 문제 해결 경험을 습득할 수 있을 것으로 기대. 3. 수치데이터로 구성되어 있어 데이터 전처리 시간을 줄일 수 있었다.

2. 데이터 설명

주어진 데이터 파일	
Train.csv	자전거 수요예측 훈련 데이터
Test.csv	자전거 수요예측 점검 데이터
데이터 필드 상세 사항	
Datetime	Hourly date + timestamp
Season	1 = spring, 2 = summer, 3 = fall, 4 = winter
Holiday	whether the day is considered a holiday
Workingday	whether the day is neither a weekend nor holiday
Weather	1: Clear, Few clouds, Partly cloudy, Partly cloudy 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

Temp	temperature in Celsius
Atemp	"feels like" temperature in Celsius
Humidity	relative humidity
Windspeed	wind speed
Casual	number of non-registered user rentals initiated
Registered	number of registered user rentals initiated
Count	number of total rentals

3. 과제 수행 내역

1) 목표

- 주어진 검증데이터의 날짜 및 기상데이터들을 바탕으로 최적의 자전거 대여 수요를 예측.

- 팀원 모두 데이터에 대한 이해와 더불어 Tensorflow를 활용한 머신러닝 분석기법 및 데이터 전처리 단계에서 기존에 학습한 핸들링 기법을 익숙하게 다룰 수 있도록 함.

2) 데이터 전처리

EDA 결과 특정 변수 값에 결측치로 보이는 0값이 많았으며, 각 변수들간의 최대값 및 최소값 차이가 크기 때문에 모델링 과정에서 데이터의 특성을 제대로 반영하지 못하는 경향을 보임. 이에 따라 데이터를 표준화할 필요성이 존재함.

상관계수 및 시각화를 통해 변수 간 상관성이 많았거나, test 데이터에 없는 변수들(Casual, registered, season, atemp 등)을 지우고 진행하였다.

또한 전처리 및 탐색과정에서 변수 시각화를 통해 연도별, 월별, 일별, 시간별 자전거 대여 수요의 변화를 파악할 수 있었고, 그 결과 변수와 수요량간의 특정한 경향이 있는 것을 확인하였다.

다음으로 regplot 함수에서 변수의 결측치를 파악할 수 있었다. Windspeed의 경우 결측치가 0으로 지정되있는 것을 확인하고, 다음 단계에서 결측치값을 다른 분포값으로 대치하고자 하였다.

마지막으로 수치 데이터의 크기 차이를 줄이기 위하여 Year 컬럼 값의 년도 데이터를 0, 1로 바꾸고자 하였다.

3) 모델링

처음 single layer인 단일 회귀식으로 진행하였으나, Cost값이 26000 이하로 떨어지지 않는 것을 확인. Layer층을 늘리는 방식으로 Cost값을 줄이고자 하였다.

하지만 Cost 값이 특정 수치 이하로 떨어지지 않았는데, feature 개수의 문제 혹은 layer 층을 거치면서 복잡도가 상승했거나, feature 전처리 단계에서의 문제가 있었던 것으로 파악된다.

4. 결과 보고

기존 주제였던 샌프란시스코 범죄 예측 주제는 데이터의 특성 상 정성적 데이터가 많았기 때문에 데이터 전처리 단계에서부터 필요이상의 시간을 허비하였고, 이와 더불어 전처리에서의 문제가 겹치면서 기계학습의 성능이 개선되지 않았다.

한편, 기존 주제에서 변경한 자전거 수요예측의 주제에서는 기존 주제와 달리 정량적 데이터가 주를 이루었고 그 결과 전처리에서 소비한 시간은 상대적으로 적었지만, 한정된 시간에서 데이터에 대한 심층적인 이해도가 낮았던 것으로 생각이 된다.

분석 주제 선정에서 보다 신중하게 주제를 선택해야 하는 점을 알 수 있었고, 주제 데이터 탐색단계에서 데이터 복잡도와 주제 난이도를 충분히 고려하여 다음 프로젝트에서는 보다 적절한 주제를 선정할 수 있도록 해야 함을 알 수 있었다.