zero-base /

Life Changing Education

머신러닝 프로젝트 가이드

- 만 74일. 여러분들이 이렇게 온라인으로 데이터사이언스 배우기 시작한지 딱 만 74일이 되었습니다.
- 그리고 놀랍게도 머신러닝을 시작한지는 딱 18일이 되었습니다.
- 이렇게 짧은 기간에 여러분들은 꽤 난이도가 높은 과제들을 다수 이수를 했습니다.
- 이제 우리 과정에서 딥러닝 프로젝트 다음으로 중요한 머신러닝 프로젝트를 진행하려고 합니다.
- 그런 여러분들이 정말 대단하다고 생각합니다.

- 이런 프로젝트를 수행할 때는 먼저 조심해야할 것이 있습니다.
- 수업은 그저 보조적 자료였을 뿐입니다.
- 여러분들은 수업 내에서 답을 찾으려 하지 말고 구할 수 있는 가능한 한 많은 자료를 참조해서 보시기 바랍니다.
- 한가지 조언을 한다면 다른 사람들의 발표나 자료를 많이 탐독해 보시기 바랍니다.
- 그리고 또한 파격적일 정도의 참신한 주제를 찾으려고 하지 마세요.
- 딱~ 지금까지 수행한 과제 정도의 난이도여도 상관없습니다.
- 주제의 참신성까지 있으면 좋지만, 그것까지 목표로 하다가 중도에 포기하거나, 정작 지켜야할 중요한 기초 절차나 목표까지도 놓치는 경우를 많이 봤기 때문입니다.

- 이제 여러분 스스로 진행하는 첫 프로젝트에서 몇 가지 주의해야할 내용을 정리하려고 합니다.
- 이 후 몇 가지 주의해야할 내용을 빼면 매우 자유롭게 데이터를 다뤄보시길 바랍니다.

- 주의1) 시간을 독립변수로 사용하려 하지 마세요.
- 예를 들어 내일의 주가예측, 한달간의 주가예측과 같은 주제를 선정하지 마세요.
- 여러분들이 배운 머신러닝은 시간을 특성(독립변수)으로 사용하지 않습니다.
- 머신러닝과 현시점 대다수의 딥러닝들은 시불변(Time Invariant)한 상황이라고 가정합니다.
- 즉, 평일과 주말의 구분, 요일별 분석, 점심과 저녁 등의 특성은 가능합니다. 그러나 이것은 이번주 일요일과 일년전 일요일은 같은 일요일이라고 봅니다. 통계에서도 시간을 변수로 사용하려면 시계열 분석과 같은 어려운 분야로 따로 들어야가 합니다. 딥러닝에서는 RNN 계열까지 가야 합니다.
- 어려우면 주식 가격, 유가, 금가격, 비트코인 시세 예측 등은 머신러닝 프로젝트의 범위가 아니라고 생각하시면 됩니다.

- 주의2) 데이터를 이해하려는 노력을 많이 하세요
- 프로젝트를 진행하고 발표하는 친구들에게 “근데 데이터가 몇 개에요?”라는 질문에 대답을 잘 하지 못하고 자료를 뒤적거리는 분들이 종종 있습니다.
- 데이터의 현황, 각 특성별 특징, 편향성, 이상치의 상황, 결측치의 상황 등을 파악하고 극복하기 위해 고민하는 것은 당연한 절차입니다.

- 주의3) 테스트 데이터를 오염시키지 마세요
- 테스트 데이터는 모델의 성능을 평가하는 최종적 장치입니다.
- 테스트 데이터를 훈련용 데이터 다루듯이 함부로 다루면 평가 결과를 신뢰할 수 없게 됩니다.

- 권고1) 팀원의 역할 분담을 직능별로 하지 마세요
- 어쩔 수 없거나 팀원의 합의에 따르는 문제라서 여러분의 몫이지만, 예를 들어 누군가는 크롤링, 누군가는 전처리, 누군가는 머신러닝 학습... 이렇게 나누지 마세요.
- 여러분 모두가 크롤링부터 머신러닝, 딥러닝을 공부하기 위해 이 과정에 합류했습니다.
- 이런 프로젝트는 정말 중요한 과정입니다. 그런데 이 중요한 과정을 이렇게 직능별로 역할을 나눠버리면 여러분들은 전 과정을 직접 경험할 기회를 놓치는 것입니다.
- 권고하는 것은 직능별이 아닌, 스토리별로 역할을 나눠보길 바랍니다.

- 권고2) 여러분들이 생각하는 결과에 과정을 맞추려 하지 마세요
- 데이터와 모델의 결과를 가지고 원하는 방향으로 진행하려고 억지로 과정을 진행하지 마세요
- 여러분이 혹시 잘못 기대했거나 과정이나 데이터가 뭔가 잘못되었을 수 있습니다.
- 전체 과정을 항상 다시 고민해보길 바랍니다.

- 권고3) 만약 여러분이 크롤링 데이터를 사용한다면 중간 목표를 시간적으로 설정해 두세요.
- 크롤링은 이런 프로젝트에서 시간을 많이 소모하게 됩니다.
- 프로젝트 마감일까지 크롤링만 집중하다가 정작 데이터를 들여다보고 관찰하고 고민하는 소중한 시간적 기회를 놓칠 수 있습니다.
- 크롤링 데이터를 사용한다면 크롤링에 대한 시간 계획을 잘 세워두세요

- 여러분들이 작성할 기획서에는 아래 내용들이 꼭 포함되도록 해주세요

- 주제
- 데이터 소개
- 데이터를 얻을 방법
- 머신러닝을 통해 얻고자 하는 결과
- 개괄적이지만 현실적인 시간 계획
- 팀원으로서의 각자 목표한 역할

- 언제나 여러분들을 응원합니다. 감사합니다.