강의 특징

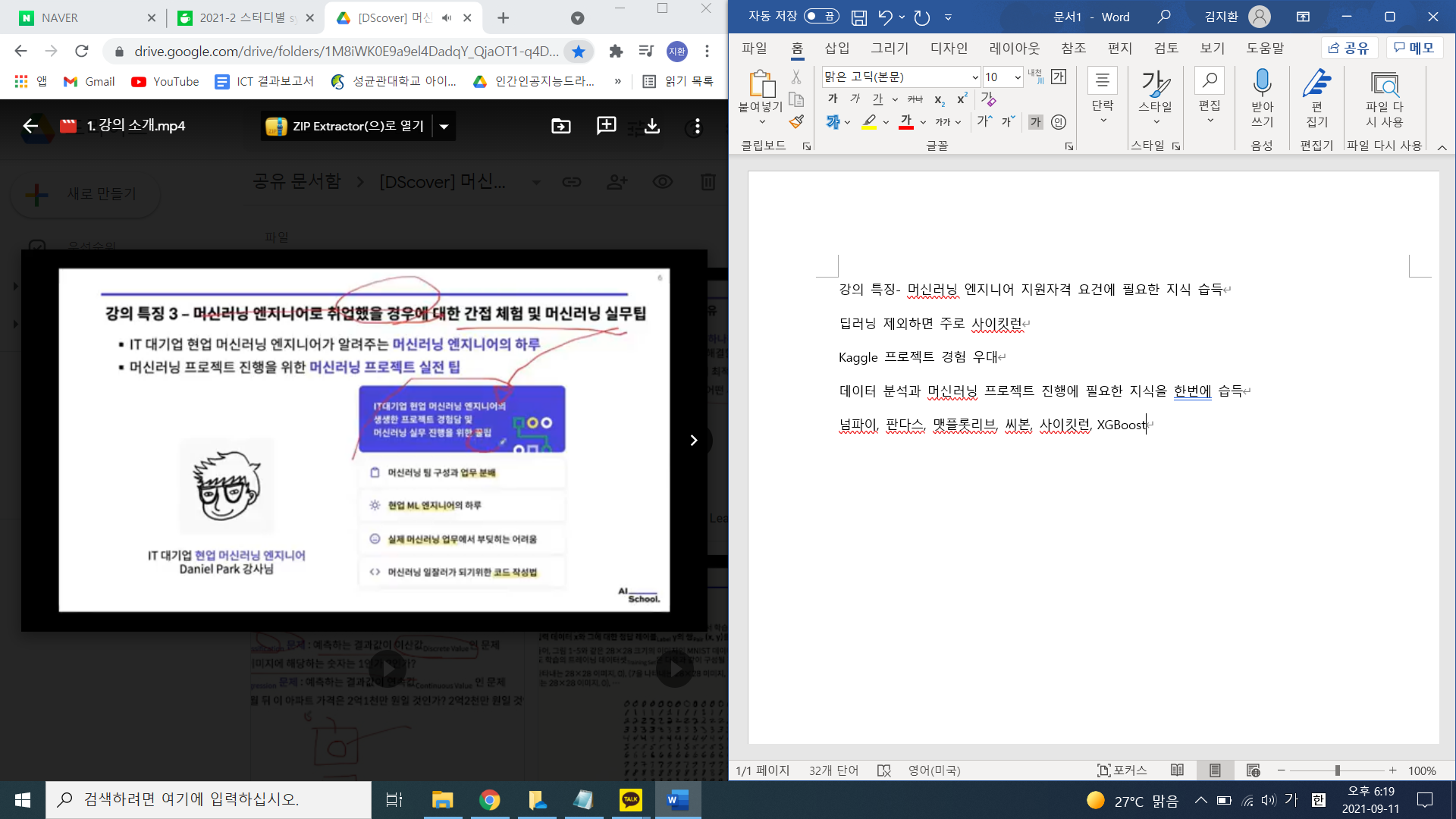
1. 머신러닝 엔지니어 지원자격 요건에 필요한 지식 습득

딥러닝 제외하면 주로 사이킷런

Kaggle 프로젝트 경험 우대

1. 데이터 분석과 머신러닝 프로젝트 진행에 필요한 지식을 한번에 습득

넘파이, 판다스, 맷플롯리브, 씨본, 사이킷런, XGBoost



1. 머신러닝 엔지니어로 취업했을 경우에 대한 간접 체험 및 머신러닝 실무팁
2. 기본 베이스라인 모델부터 점진적인 성능향상 기법 적용
3. Google Colab 이용
4. 머신러닝 처음 접하는 사람, 데이터 분석 기법, 머신러닝 엔지니어로 취업 원하는 사람, 취업 이후 업무 프로세스, 실무 꿀팁

머신러닝의 개념

1. 머신러닝이란?

명시적인 프로그래밍 없이 데이터를 이용해서 예측을 수행하는 알고리즘을 구현하는 기법

1. 데이터로부터 컴퓨터가 어떤 패턴을 학습하는 알고리즘

데이터 -> 머신러닝 모델을 통한 학습 -> 예측

1. 머신러닝이 필요한 이유

인간이 정확히 하나하나 로직을 지정해주기 어려운 복잡한 문제를 데이터에 기반한 학습을 통해서 해결할 수 있음

Ex. 최적의 광고 배분 전략, 사용내역 사기 여부 파악

1. 예측 모델의 필요성

정교한 예측 모델을 갖고 있으면 중요한 비즈니석 의사결정을 안정적이고 계획적으로 수행할 수 있음

1. 휴대폰 판매량 예측 -> 생산 계획 및 재고관리 전략 수립 가능

분류 vs. 회귀

1. 예측 모델 값의 형태에 따라 분류 혹은 회귀

분류: 예측하는 결과값이 이산값(discrete value) -> 하나의 카테고리 선택

회귀: 예측하는 결과값이 연속값(continuous value) -> 근사 숫자를 맞추는 것!

머신러닝 알고리즘의 3가지 분류

1. Supervised Learning(지도학습)

* 인풋 데이터 x와 그에 대한 정답 레이블 y의 쌍 (x,y)를 이용해서 학습하는 방법론
* 이산값 예측하는 경우 분류, 연속값 예측하는 경우 회귀
* 딥러닝: CNN, RNN

1. Unsupervised Learning

* 인풋 데이터 (x)만을 이용해서 데이터의 숨겨진 특징을 학습하는 방법론
* 클러스터링 알고리즘을 적용해 여러 분류 기사 분류 가능
* 단독을 사용하기보다 비지도 학습으로 파악한 데이터의 숨겨진 특징을 원본 데이터 대신 지도 학습의 인풋 데이터로 활용해서 지도 학습의 성능을 더욱 끌어올리는 용도
* 대표적인 방법론: 주성분 분석(PCA)
* 딥러닝에서 비지도 학습을 위해 많이 사용되는 구조: 오토인코더

1. Reinforcement Learning

* Reward를 제공하는 Environment와 Environmnet 내에서 행동을 수행하는 Agent의 상호작용을 통해 학습하는 방법론
* 정적인 상태의 데이터를 학습하는 지도학습과 비지도학습과는 달리, 강화학습은 에이전트가 주어진 환경에서 어떤 행동을 취하고 이에 대한 보상을 얻으면서 학습 진행
* 에이전트는 보상을 최대화하도록 학습 진행
* 에이전트가 시뮬레이션을 해볼 수 있는 동적인 상태(Dynamic Environment)에서 데이터를 수집하는 과정까지 학습 과정에 포함되어 있는 알고리즘
* 대표적 알고리즘: Q-learning,, 딥러닝과 결합: DQN 기법

지도학습

* 지도학습 방법론을 사용하기 위해서 트레이닝 데이터의 구성이 (인풋 데이터, 데이터에 대한 정답) 쌍으로 구성되어 있어야만 함
* 지도학습이란 정답을 보여주며 학습시키는 머신러닝 방법론
* 인풋 데이터: x, 데이터에 대한 정답: y ex> f(키) = 몸무게

1. Training data, Test data

* (x,y)로 구성된 학습용 데이터로 머신러닝 모델을 학습시킴
* 학습된 머신러닝 모델의 성능을 트레이닝 데이터에 포함되어 있지 않고 따로 빼놓은 테스트 데이터(Test Data)로 측정

1. 트레이닝 데이터, 테스트 데이터 나누기(split)

* 머신러닝 모델 학습시키기 위해 전체 데이터의 일부를 Training Data, 일부는 Test Data로 나눠서 사용
* 일반적으로 데이터의 80% 정도는 트레이닝 데이터, 20% 정도는 테스트 데이터로 나눠서 사용

1. Validation Data(검증용 데이터)

* 직관적으로 설명하면 검증용 데이터는 트레이닝 과정 중간에 사용하는 테스트 데이터
* 트레이닝 과정에서 학습에 사용하지는 않지만 중간중간 테스트하는데 사용해서 학습하고 있는 모델이 오버피팅에 빠지지 않았는지 체크하는데 사용됨
* 70%(트레이닝), 15%(검증), 15%(테슽)
  1. Overfitting(과적합)
  + Training Data에만 최적화된 형태로 모델이 바뀌는 경우를 말함
  + 트레이닝 에러는 작아지지만 검증 에러는 커지는 경우
* Early stopping: 트레이닝 에러가 작아진다고 학습을 계속 진행하는 것이 아니라 검증 에러가 커지는 지점에서 업데이트를 중지하면 최적의 파라미터(?)를 얻을 수 있음
* 오버피팅을 방지하기 위한 기법을 총칭하여 Regularization 기법!
  1. Underfitting
  + 사용하는 알고리즘이 부실하여 Training data 마저도 잘 잡아내지 못할 때
  + O/X 분류할 때 곡선으로 분류하는 것이 아니라 선형으로 분류하게 되는 경우 정밀성 떨어짐

Google Colab 소개

1. 코랩의 장점

* 구성이 필요하지 않음
* GPU 무료 액세스
* 간편한 공유
* 라이브러리 바로 사용 가능