프로그램 북

DS School 수업 소개

실전 머신러닝반

수강료 : 594,000원

DS School (더넥스트스쿨)

강성희 대표이사

강연 진행

강연은 총 4주간 진행되며, 주 1회씩 총 4회를 진행합니다. 매 회차마다 실습이 포함되어 있으며, 이를 위해 회당 강연 시간은 다섯 시간으로 책정하였습니다. 그러므로 총 수업 시간은 4 회 20 시간입니다. 강연의 세부 내용은 다음과 같습니다.

실전 머신러닝반 1주차

첫 번째 수업에서는 머신러닝 알고리즘에 대한 기본적인 설명부터 시작합니다. 머신러닝의 기본 개념인 지도학습(Supervised Learning)과 비지도학습(Unsupervised Learning)에 대해 살펴보고, 그 차이점과 장단점에 대해 살펴봅니다. 또한 데이터셋마다 가장 효과를 낼 수 있는 알고리즘에 대해 살펴보며, 주로 구조화된 데이터셋(Structured Dataset)과 비구조화된 데이터셋(Unstructured Dataset)을 중심으로 살펴봅니다.

이후 첫 번째 머신러닝 알고리즘으로, 가장 기초적인 알고리즘 중 하나인 의사결정나무(Decision Tree)에 대해 살펴봅니다. 의사결정나무의 기본적인 동작 방식과, 의사결정나무 알고리즘의 핵심이 되는 지니 불순도(Gini Impurity)에 대해서 다룹니다. 이후 프로그래밍 언어 파이썬을 통해 의사결정나무를 직접 작성하고 실행해봅니다.

실전 머신러닝반 2주차

두 번째 수업에서는 여러 개의 머신러닝 모델을 섞어서 성능을 끌어올리는 앙상블(Ensemble) 알고리즘에 대해 살펴봅니다. 가장 유명한 앙상블 알고리즘인 배깅(Bagging)과 부스팅(Boosting), 그리고 부스팅의 업그레이드 버전인 그래디언트 부스팅(Gradient Boosting)에 대해서 살펴보고, 이 앙상블 알고리즘 간의 차이점과 장단점을 살펴봅니다.

이후 이 앙상블 알고리즘을 의사결정나무(Decision Tree) 알고리즘에 적용합니다. 먼저 의사결정나무에 배깅(Bagging) 알고리즘을 적용한 랜덤 포레스트(Random Forest)에 대해 살펴보고, 마찬가지로 그래디언트 부스팅(Gradient Boosting) 알고리즘을 적용한 그래디언트 부스팅 머신(Gradient Boosting Machine)을 살펴봅니다. 이후 프로그래밍 언어 파이썬을 통해 이 알고리즘들을 직접 작성하고 실행해봅니다.

실전 머신러닝반 3주차

세번 째 수업에서는 머신러닝을 실용적인 관점에서 접근합니다. 먼저 그래디언트 부스팅 머신(Gradient Boosting Machine)을 사용하기 쉽도록 구현한 세 가지 파이썬 패키지(XGBoost, LightGBM, CatBoost)를 살펴본 뒤 이 패키지들의 장단점에 대해 살펴봅니다. 또한 왜 기존의 그래디언트 부스팅 머신 패키지(ex: scikit-learn)에 비해 XGBoost, LightGBM, CatBoost이 더 성능이 좋은지도 살펴봅니다.

이후 그래디언트 부스팅 머신(Gradient Boosting Machine)의 성능을 튜닝할 수 있는 하이퍼패러미터(Hyperparameter)에 대해 다룹니다. 가장 중요한 하이퍼패러미터들(ex: 트리의 깊이, 갯수 등)과 이 역할, 각각의 튜닝 방식에 대해 살펴보고, 마지막으로 모든 하이퍼패러미터를 동시에 튜닝하는 방법(ex: Grid Search, Random Search)에 대해 살펴봅니다.

실전 머신러닝반 4주차

마지막 네 번째 수업에서는 그래디언트 부스팅 머신(Gradient Boosting Machine)을 실전에 적용해봅니다. 데이터 사이언티스트들이 참여하는 온라인 경진대회 캐글(Kaggle)에 도전하며, 주어진 정보를 활용해 전자상거래(ex: 쿠팡, 11번가) 서비스의 상품을 분류하는(Product Classification) Otto Group Product Classification Challenge에 참여합니다.

이 경진대회에서는 데이터를 분석하는 스킬도 중요하지만, 그보다 머신러닝 알고리즘에 대한 이해와 하이퍼패러미터 튜닝 방법을 숙지하는 것이 더 중요합니다. 이번 경진대회에서의 목표 등수는 상위 10% 입니다. 보통 캐글에서는 상위 10% 안에 든 참석자를 현장에서 당장 일할 수 있는 실력을 갖추었다고 평가하는데, 만일 스스로의 힘으로 상위 10% 안에 들 수 있다면 수업을 충분히 따라왔다고 볼 수 있고, 머신러닝에 관해서는 당장 즉시 전력으로 현장에 투입될 수 있는 실력을 갖췄다고 볼 수 있습니다.