Closing

Hyunjoong Kim

soy.lovit@gmail.com

github.com/lovit

Myths of data science

- Complex models are better than simple models
- Data Science requires massive computing power
- More data = more accuracy
  - Garbage in, garbage out
- They think that large organization already has enough data to use for analysis

https://www.datasciencecentral.com/profiles/blogs/10-myths-about-data-scientists
https://hub.packtpub.com/9-data-science-myths-debunked/
https://www.dataversity.net/ten-myths-about-data-science/

수업 목표

- 데이터 분석을 잘 수행하기 위해서는
  - 데이터 수집 계획을 세울 수 있어야 합니다.
  - 데이터를 정제하고 처리할 수 있는 수준의 프로그래밍을 수행해야 합니다.
  - 머신러닝 모델의 작동 원리를 이해하고, 제 기능에 맞게 활용해야 합니다.

수업 목표

- 데이터를 정제하고 처리할 수 있는 수준의 프로그래밍을 수행해야 합니다.
  - 데이터 분석의 90% 는 전처리, 10% 는 모델링이라는 농담도 있습니다.
  - 머신러닝 모델의 학습 부분보다 데이터 처리의 코드가 훨씬 깁니다.
  - 데이터 처리 → 모델 학습 → 학습 결과 탐색에 필요한 프로그래밍을 익혀봅시다.
    - Pandas, Bokeh, Seaborn, Numpy >> scikit-learn
  - 모든 준비를 마치고 분석을 시작할 수는 없습니다. 필요를 느끼지 못하면 기억도 잘 나지 않습니다. 실전에서 프로그래밍을 시작해봅시다.

수업 목표

- 알고리즘을 학습하는 것과 “잘 이용하는 것“은 다릅니다.
  - 알고리즘도 단점들이 존재합니다.
  - 각 알고리즘의 한계를 알고, 상황에 맞게 적절히 선택합니다.

- 복잡한 모델 이전에, 기본이 되는 모델부터 정확히 이해해야 합니다.
  - 복잡한 모델의 이해를 위한 기본기 입니다.

Topics

1. Introduction & Python basic
   - 머신러닝의 개념을 알아보고, 실습을 통하여 seaborn, bokeh, pandas, numpy 의 사용법을 익힙니다.
   - 실습 코드는 1일에 소화할 양이 아니며, 위 네 패키지의 튜토리얼입니다. 필요한 내용들은 각 일차별 내용에 포함되어 있으니, 자세한 튜토리얼을 보고 싶을 때 발췌하면 좋습니다.
2. Linear Regression
3. Logistic Regression
   - 지도학습 기법의 기본 모델인 선형회귀와 로지스틱 분류모델을 통하여 머신러닝의 기본 개념 및 scikit-learn 의 사용방법을 공부합니다.
   - Python, Seaborn, Numpy 의 기본적인 사용법을 공부합니다.
4. Feature extraction and preprocessing
   - Pandas 를 이용한 테이블 병합, 데이터 탐색을 공부합니다.
   - 텍스트 데이터의 벡터화 방법을 간단히 알아봅니다.

Topics

5. Feed forward neural network
   - 뉴럴 네트워크를 통하여 비선형 모델의 원리를 알아봅니다.
   - 이미지 데이터를 핸들링하는 방법도 살펴봅니다.
6. Support Vector Machine
7. Decision Tree
8. Tree based Ensembles
   - Kaggle 에서 좋은 성능을 보여주는 Random Forest, XGBoost 등의 앙상블 기반 모델들을 알아봅니다.
9. Nearest Neighbor methods
   - Collaborative Filtering 을 통한 추천 모델을 알아봅니다.
10. Clustering

- 복잡한 알고리즘들 (SVM, Random Forest, XGBoost) 등은 인공데이터를 이용하여 하이퍼 패러매터별 모델의 학습 경향과 실패 사례 등을 살펴보았습니다.
- 모델이 예상만큼 잘 작동할 때에는 반드시 "잘 작동하는 이유를 설명할 수 있어야" 하며, "모델이 제대로 작동하지 못하는 상황"을 알고 있어야 합니다.
  - 대부분의 경우에 예상과 다르게 모델이 작동할 것입니다.
  - 잘 작동하지 않을 경우들이 주로 "디버깅 포인트"가 됩니다.

- 모든 데이터에 우월한 모델은 없습니다. 각자가 잘 학습하는 데이터 패턴이 있습니다. 여러분이 분석해야 하는 데이터의 특징을 잘 반영할 수 있는 알고리즘을 선택하고, 때로는 데이터의 형태를 적절히 변경하세요.
  - 왜 이미지는 CNN 을 이용할 때 feed-forward neural network 보다 성능이 좋았을까요?
  - 왜 Random Forest / XGBoost 들은 Kaggle 에서 좋은 성능을 보였던 걸까요?

- 나의 데이터에 나의 문제를 해결하기 위한 정보가 정말로 존재하는지, 어떤 정보가 문제 해결의 힌트일지부터 고민하시기 바랍니다. 그 고민의 과정이 이뤄진다면 이미 적합한 프로세스와 모델을 선택하셨을 것입니다.