데이터과학 절차

이영석

http://yslee.cs-cnu.org

할 일

- 문제정의
  - 목표, 범위, 성공/실패 기준, 비용/제약조건
- 데이터수집
- 데이터준비
- 탐험적데이터분석
  - 현상을 이해하기 위해서
- 통계적추론
  - 현상을 일반화하기 위해서
- 기계학습
  - 예측을 위해서

넷플릭스 프라이즈 사례

- **목표**
  - 넷플릭스 프라이즈의 목표는 넷플릭스 사용자들의 영화 선호도 데이터를 (별점:1~5) 바탕으로 각 사용자가 미래에 볼 영화의 선호도를 예측하는 것이다.
- **범위**
  - 참가자들은 넷플릭스에서 제공한 데이터를 가지고 예측 모델을 개발하며, 평가 역시 넷플릭스에서 미리 정해진 데이터셋을 기반으로 실시한다.
- **성공기준**
  - 참가자들의 예측 모델은 넷플릭스의 자체 모델보다 10%이상 예측 성능을 향상시켜야 하며, 평가를 위해서는 RMSE(Root Mean Squared Error)를 사용한다.

- **데이터**
  - 참가자들에게는 약 50만명 가량의 사용자가 17000여개의 영화를 평가한 백만
    데이터가 제공된다. 사용자들의 프라이버시를 보호하기 위해 모든 데이터는 익
    제공된다.
- **제약조건**
  - 참가자들은 하루에 최대 하나의 예측 결과를 업로드할 수 있다. 넷플릭스는 예
    평가하여 그 일부의 결과를 공개하고, 나머지는 최종 결과의 심사를 위해 사용한다.

탐험적 데이터 분석

- Exploratory data analysis (EDA)
- 방법
  - 원본 관찰
  - 통계
  - 시각화

기계학습(Machine Learning)

- 예
  - 스팸메일 인지 아닌지 ?
  - 문자인식 ?
- 기계학습 이란?
  - 훈련데이터를 통해 알려진 속성을 기반으로 예측
  - 참고: 데이터마이닝은 데이터의 속성 발견에 초점
- 접근방법
  - 결정 트리
  - 연관 규칙
  - 신경망
  - 유전계획법
  - Support Vector Machine
  - 클러스터링
  - 베이즈 네트워크
  - 강화학습법 ...

참고문헌

- http://www.hellodatascience.com/?p=252