# 2017-06-21 (1주차)

1. 수학
    a. 선형대수 : 중급까지
    b. 미적분학 : 최적화를 위해
    c. 최적화 : 가장 좋은 무언가를 찾는 것
    d. 확률론 : 베이지안은 논리학의 연장으로 볼수 있음. 어떤 주장(모델)의 신뢰도를 나타내기 위하여

2. 데이터 분석의 목적
   데이터분석 이론의 대부분은 20~30년 전에 다 만들어진 내용, 최근에 만들어진 딥러닝의 일부분을 제외하고는 계산능력의 증가(소프크웨어, 하드웨어)로 가능해졌음
   도메인날리지 중요: 문제설정, 모델에 대한 성능판단
    a. 예측문제
       데이터간의 관계를 파악 후 하나의 부재의 경우에 예측할 수 있음
       forecasting(미래에 대한 예츨)이 아닌 prediction(데이터의 빈공간을 채우는)
       ex) 부동산 가격예측: 단답형
       붓꽃종 예측: 객관식,정적인 관계
       얼굴사진 사람이름 예측: 객관식, 정적인관계
       알파고: 동적인관계, 이런문제를 풀때는 강화학습을 이용

    b. 머신러닝과 과거 시스템간의 관계
        - 과거: rule-based, expert system (전문가가 만들어낸 규칙을 시스템에 입력)
          전문가 시스템(프롤로그)을 이용하여 현업을 대체하려 노력 -> 실패
          어느정도 성공한것: 번역(사전,룰북 + 문법) 하지만 복잡한 문제는 실패

        - 현재:  data-based, learning-based
          전문가가 규칙을 어떻게 만들어 냈는가에 대한 고찰 -> 오랜시간 데이터를 보면서
          전문가: 기계에 룰을 삽입 vs data-based: data를 통해 룰을 생생(학습)

          머신러닝시스템을 구매하고 끝이 아니라 학습을 시키는 사람이 지식을 가지고 있어야 사용가능

3. 지도학습
4. 예측문제의 수학적표현
   예측 문제는 y = f(x)에서 x와 y의 관계를 나타내는 f를 찾는 것
    a. 주가예측문제 : 이평선을 통하여 상승과 하락정보를 차트에 표시하고 뉴럴넷 학습 -> 정확히 이평선을 찾아냄(어떤 문제에 룰이 존재하면 결국엔 풀 수있다.)
       주가 예측문제는 애초에 룰이 존재하지 않는 문제라 생각

    b. 룰베이스로 만든 데이터를 이용하여 학습: 그 룰을 학습, 도메인 문제에 존재하는 룰을 찾지 못함(잘못된 생각)
    c. 이미지넷
       대학원생들이 노가다해서 단순히 레이블 뿐만아니라 온톨로지까지 레이블링한 데이터셋을 생성 -> 대회개최 -> 뉴럴넷 이용한 성공 -> 관심
       데이터에 레이블링 작업은 data-based시스템을 만들기 위해선 피할수 없는 과정

5. 인코딩
   데이터를 숫자로 바꾸는 작업
   기본적인 입력차원은 고정되나 최근엔 유연한 시스템도 존재

6. 아나콘다
   계정별로 가상환경별로 파이썬과 패키지를 설치가능
    - conda vs pip
      pip: 소스코드로 받아와서 빌드를 해야함
      conda: 아키텍처별로 서버가 나뉘어 있고 바이너리 패키지를 갖고 있음, 단점은 pip의 모든 패키지를 커버하지 못함.
    - 패키지
      sympy: 미적분 중에서 수식을 찾아내는 기능제공
      tensorflow: theano대체, 분산처리가 가능한 패키지
      keras: tensorflow, theano의 상위레벨 패키지, 쓰기 쉽게 랩핑하고 있음.

7. 질문
    - 선형대수에서 나오는 분해는 어디에 쓰나?
      -> 분해는 선형대수에서 매우중요한 부분
      -> decomposition은 approximation에서 쓴다.
      -> 행렬을 분해한후 감마값이 작은 분해된 결과를 제거 함으로써 데이터의 사이즈를 줄이고 값은 근사하도록 유지(큰행렬을 분해하여 감마값이 작은 축을 삭제하여도 근사함.)

    - 클래스를 구성하는 데이터수와 클래스 수간의 비율이 이상적이지 않을 때(imbalanced dataset, 현재 카탈로그매칭문제)
      -> sampling: 많은 데이터를 갖는 클래스에서는 샘플링을 통하여 데이터의 수를 줄이고
      -> population: 적은 수의 데이터를 갖는 클래스에서는 데이터를 증가시킨다.

    - 데이터구성을 조작하는 방법
    - 문제에 룰이 존재하지 않는지 어떻게 알수 있는가?
      -> 해당 도메인의 전문가가 알수 있음
      -> 100%의 룰이아니더라도 80~90%로의 의미있는 수준의 룰

    - 주가 예측문제에서 X 즉 그시점을 표현하는 feature가 부족해서 룰을 못찾는 것은 아닌가?
      -> 그러한 차원의 문제가 아님, 해당도메인의 지식이 필요한 말을 하심, 인간의 행동간의 관계

- 전체 시간에 대한 주가의 흐름은 법칙이 없을 수 있지만 피처로 표현한 그시점에 대한 사람들의 행위가 확률적으로 동등할것이라 생각, 하지만 추세라는 것을 봤을때 일정시간(단기)에서 사람들의 행동은 예측가능하지 않을까?
- superersolution 가능한것인가
- 머신러닝 파이프라인 구성법
- 케글에서 쓰이는 강력한 알고리즘으로 알려진 XGboost가 실제로 쓸만한가

- tensorflow자체를 집중적으로 배울 수 있는 서적이나 강의
  -> 1.2전까지 api가 너무 자주 바뀌어왔다. 서적이 존재하기 힘듬
  -> keras를 이용하는 이유

- 현업에서 100만건 이상 데이터 학습 및 데이터 예측시간을 줄이고 싶다.
  ->  Dask package 이용하여 작업(가상의 데이터 프레임) 뒷단의 스켈줄러를 띄워서 취합해서 적용(mapreduce작업)