# Wine Quality Data Set 프로젝트

2018년 6월 5일

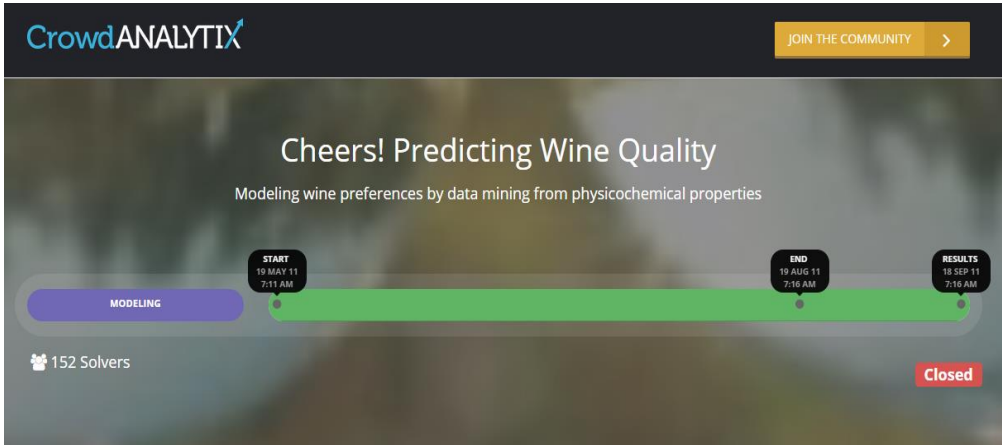
숭실대학교 컴퓨터학부
기계학습 연구실 유수정

# Wine Quality Data Set

- 포르투갈에서 생산되는 비뉴 베르드 (vinho verde) 와인
- 레드 와인과 화이트 와인 데이터
  - 레드와인 관측값 1599개
  - 화이트와인 관측값 4898개
  - 변수 12개
    - 설명변수 11개
      - 예) fixed acidity, residual sugar, sulphates, alcohol
    - 반응 변수 1개
      - Quality
        - » 와인의 품질을 0 (아주 나쁨)부터 10 (아주 좋음)까지 매긴 값.
        - » 와인 전문가들에 의해 최소 3번 이상 평가된 값의 중간 값으로 설정함.

# Wine Quality Data Set (cont'd)

- 2009년에 Paulo Cortez가 발표한 논문에서 사용된 데이터
  - CVRVV (Vinhos Verdes 지역의 포도 재배 위원회)에서 2004년부터 2007년까지 수집
- 2011년 CrowdAnalytix에서 주최한 Cheers! Predicting Wine Quality 대회




# Wine Quality Data Set (cont'd)

- 11개의 설명변수

| 설명변수 | 데이터 유형 | 설명 |
|---|---|---|
| Fixed acidity | Numeric | 와인의 $dm^3$당 타르타르산의 그램 수 |
| Volatile acidity | Numeric | 와인의 $dm^3$당 아세트산의 그램 수 |
| Citric acid | Numeric | 와인의 $dm^3$당 시트르산의 그램 수 |
| Residual sugar | Numeric | 발효과정이 끝난 뒤 남아있는 와인의 당량 $(g/dm^3)$ |
| Chlorides | Numeric | 와인의 $dm^3$당 염화나트륨 그램 수 |
| Free sulfur dioxide | Numeric | 와인의 $dm^3$당 무이산화황 그램 수 |

# Wine Quality Data Set (cont'd)

- 11개의 설명변수

| 설명변수 | 데이터 유형 | 설명 |
|---|---|---|
| Total sulfur dioxide | Numeric | 와인의 $dm^3$당 전체 이산화황의 그램 수 |
| Density | Numeric | 와인의 $cm^3$당 그램 수 |
| PH | Numeric | 와인의 PH 값 |
| Sulphates | Numeric | 와인의 $dm^3$당 황산칼륨의 그램 수 |
| Alcohol | Numeric | 알코올의 부피의 %값 |

# 문제 1

- 주어진 레드 와인 데이터와 화이트 와인 데이터에 결측치(NA)가 있다면 결측치를 평균값으로 대체하고, 데이터의 반응변수 분포를 그래프로 나타내시오.

# 문제 2

- 교재에 나와 있는 변수 선택 함수를 사용하여 forward, backward, both 방법을 문제 1에서 결측치를 제거한 두 와인의 데이터에 각각 적용하여, 세 가지 방법 중 어떤 방법이 각 데이터에서 가장 좋은 성능을 보이는지 비교하시오.

# 문제 3

- 레드 와인과 화이트 와인 데이터에서 각각의 설명변수 쌍에 대해 피어슨 상관계수를 계산해라. 계산 결과를 바탕으로 서로 상관관계가 제일 높은 변수 5쌍을 찾고 그 관계를 설명하라.

# 문제 4

- 레드 와인과 화이트 와인 데이터에 2차 interaction term을 추가하여라. 각각의 데이터에서 interaction term을 추가한 모델과 추가하지 않은 모델의 성능을 비교하라.

# 문제 5

- 문제 2에서 가장 좋은 성능을 보인 모델에 대하여 모델 평가 차트를 그리고, 교재를 참고하여 첫 번째와 두 번째 그래프의 결과를 설명하라. 또한 모델에 이상치가 있는지 확인하고, 있다면 이를 제거하고 제거하기 전과 성능에 변화가 있는지 비교하라.

# 문제 6

- 레드 와인과 화이트 와인 데이터의 설명변수 값을 랜덤으로 각각 1%, 5%, 10%씩 선택하여 결측치로 바꾼 뒤, 문제 1과 같이 결측치를 평균값으로 대체하고 선형 회귀 분석을 하시오. 이 작업을 각각 10번씩 반복하고 10번의 평균치와 표준편차를 계산한 뒤, 초기 데이터의 회귀 분석 결과와 비교하시오.

# 과제 제출

- 데이터는 myclass 게시판에서 다운로드
- 제출물
  - 보고서 (hardcopy)
    - 각 문제에서 사용한 함수와 답 기술
    - 수업 범위 내에서 사용한 함수 외에 별도로 패키지를 사용한 경우, 사용한 패키지와 함수를 명시.
  - R코드
- 제출 기한
  - 보고서: 6월 11일 18시까지 408호 과제 제출 박스
  - 코드: 6월 11일 18시까지 myclass로 업로드