**R – Py 컴퓨팅: Homework 2**

**Due: 2019-11-27, 블랙보드 제출, 파일 형식은 pdf로, 코드는 부록으로 첨부**

**Q1. 와인 클래스에 대한 kNN 알고리즘 적용**

**와인 데이터 구성**

The data is the results of a chemical analysis of wines grown in the same region in Italy by three different cultivators. There are thirteen different measurements taken for different constituents found in the three types of wine.

이 와인의 화학적 분석 결과는 다음과 같이 정리된다.

Attribute Information:

- Alcohol

- Malic acid

- Ash

- Alcalinity of ash

- Magnesium

- Total phenols

- Flavanoids

- Nonflavanoid phenols

- Proanthocyanins

- Color intensity

- Hue

- OD280/OD315 of diluted wines

- Proline

와인 종류는 class 변수로 구분되어 있으며, 1, 2, 3을 각 코드값으로 지니고 있다.

이 와인데이터 (wine\_data.csv) 파일에 대해서 다음과 같은 분석을 행하라.

1. Pandas를 사용해 wine\_data.csv파일을 wine 데이터프레임을 만든다.
2. Wine 데이터에 describe 메소드를 사용하여 요약통계량을 구한다.
3. 주어진 데이터에 첫 변수, class를 우리가 예측하는 타겟 변수, y, 나머지 모든 변수를 예측 변수 X로 활용하고, Train/Test 표본을 70% 와 30%가 되도록 데이터를 구성한다. 각각의 명칭은 X\_train/X\_test, y\_train/y\_test로 칭한다.
4. Scikit-learn의 KNeighborsClassifier를 사용하여 70%인 X\_train과 y\_train을 바탕으로 모형을 트레이닝 시킨다.
5. 이 Training 된 모형을 바탕으로 Training Set에서 예측해보고 원래의 라벨을 얼마나 정확하게 예측하는지 percentage 단위로 출력하고, k=5로 결정하고 다수결 및 유클리디안 거리를 사용한다. .
6. 이 Training 된 모형을 바탕으로 Test Set에서 원래의 라벨을 얼마나 정확하게 예측하는지 percentage 단위로 출력한다. .
7. 5-6번을 k=3일 경우에 대해서도 시행한다.
8. 3-6번에서 X 데이터를 Alchol, Malic acid, Ash, Alcalinity of ash 네 변수만 활용하여 다시 수행한다.

**Q2. 새 알고리즘을 적용해야 할 상황**

*문제 상황: A라는 학생 본인은 인턴쉽 도중 와인데이터를 활용한 결과 Q1에 대해서 B라는 대리에게 보고했다. B라는 대리는 support vector machine이라는 기법이란 것이 있으며, 자기가 모형 파라미터 값만 줄 테니 조금 다른 모형을 예측만 해서 결과값만 알려달라고 하였다. B대리는 다음과 같은 분석 조건에 내걸었다.*

1. 모형” from sklearn.svm import SVC
2. SVC 파라미터는 (kernel='linear', C=1.0, gamma='auto')
3. 주어진 데이터에 첫 변수, class를 우리가 예측하는 타겟 변수, y, 나머지 모든 변수를 예측 변수 X로 활용하고, Train/Test 표본을 70% 와 30%가 되도록 데이터를 구성한다. 각각의 명칭은 X\_train/X\_test, y\_train/y\_test로 칭한다.
4. SVC로 Training 된 모형을 바탕으로 Training Set에서 예측해보고 원래의 라벨을 얼마나 정확하게 예측하는지 percentage 단위로 출력한다.
5. 이 Training 된 모형을 바탕으로 Test Set에서 원래의 라벨을 얼마나 정확하게 예측하는지 percentage 단위로 출력한다

**Q3. K-means Clustering 알고리즘**

1. Wine 주어진 데이터에 첫 변수, class를 제외한 나머지 데이터를 X로 결정한다.
2. Cluster를 세개로 구성하여 K-means Clustering 알고리즘을 적용한다.
3. X 데이터를 바탕으로 라벨을 예측해보고, pd. Crosstab 명령어를 사용하여, 실제 라벨값과 예측된 라벨값을 비교한다.
4. k=1 부터 9까지 변형시키면서 각 모형의 inertia를 구해본다.