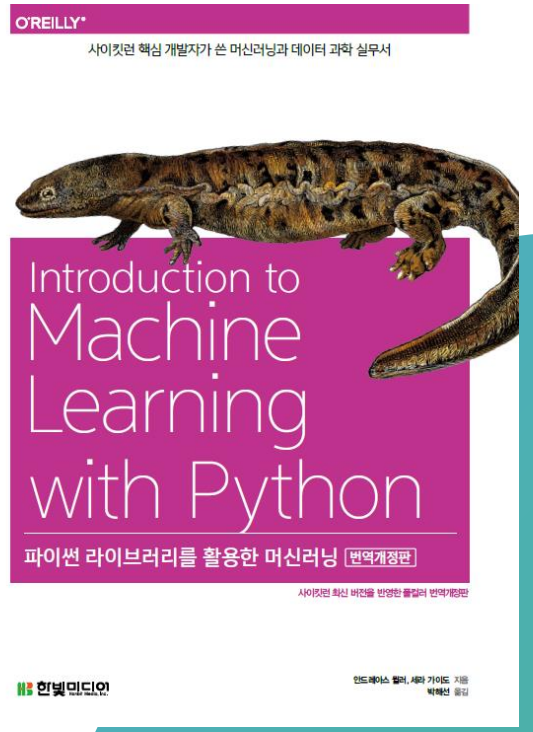
Chapter 03 : 비지도 학습과 데이터 전처리

파이썬라이브러리를 활용한머신러닝 (개정판)



시작하기전에

- **책에서 사용하는 소프트웨어 버전**
  - Python 버전: 3.7.2 (default, Dec 29 2018, 06:19:36)
  - [GCC 7.3.0]
  - pandas 버전: 0.23.4
  - matplotlib 버전: 3.0.2
  - NumPy 버전: 1.15.4
  - SciPy 버전: 1.1.0
  - IPython 버전: 7.2.0
  - scikit-learn 버전: 0.20.2
- **예제 다운로드 링크**
  - https://github.com/rickiepark/introduction_to_ml_with_python
  - https://nbviewer.jupyter.org/github/rickiepark/introduction_to_ml_with_python/tree/master/

이 책의 학습 목표

- 1장: 머신러닝과 머신러닝 애플리케이션의 기초 개념을 소개 및 사용 환경
- 2장: 지도 학습 알고리즘
- 3장: 비지도 학습 알고리즘
- 4장: 머신러닝에서 데이터를 표현하는 방법
- 5장: 모델 평가와 매개변수 튜닝을 위한 교차 검증과 그리드 서치
- 6장: 모델을 연결하고 워크플로를 캡슐화하는 파이프라인 개념
- 7장: 텍스트 데이터에 적용하는 방법과 텍스트에 특화된 처리 기법
- 8장: 개괄적인 정리와 어려운 주제에 대한 참고 자료 안내

Contents

CHAPTER 03 비지도 학습과 데이터 전처리

CHAPTER 03 비지도 학습과 데이터 전처리

머신러닝의 비지도 학습 알고리즘

SECTION 3.1 비지도 학습의 종류

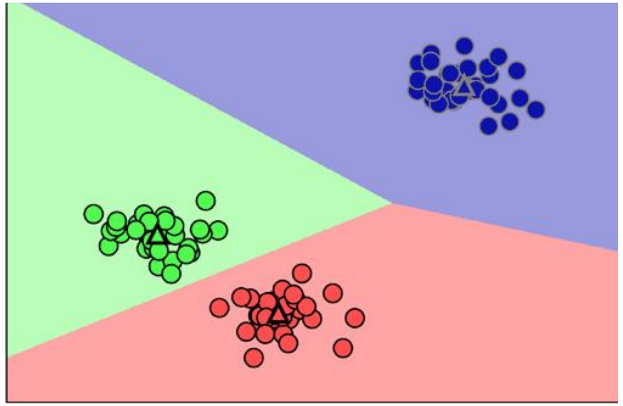
- 비지도 학습 unsupervised-learning이란?
  - 알고 있는 출력값이나 정보 없이 학습 알고리즘을 가르쳐야 하는 모든 종류의 머신러닝
    - 비지도 학습 알고리즘은 입력 데이터만으로 데이터에서 지식을 추출
  - 비지도 학습에는 비지도 변환(unsupervised transformation과 군집(clustering)이 있음
  - 비지도 변환: 데이터를 새롭게 표현하여 사람이나 다른 머신러닝 알고리즘이 원래 데이터보다 쉽게 해석할 수 있도록 만드는 알고리즘임
    - 많은 고차원 데이터를 특성의 수를 줄이면서 꼭 필요한 특징을 포함한 데이터로 표현하는 방법인 차원 축소(dimensionality reduction)의 대표적 예는 시각화를 위해 데이터셋을 2차원으로 변경하는 경우임
    - 비지도 변환으로 데이터를 구성하는 단위나 성분을 검색: 많은 텍스트 문서에서 주제를 추출
      - 소셜 미디어에서 선거, 총기 규제, 팝스타 같은 주제로 일어나는 토론을 추적할 때 사용 가능
  - 군집: 데이터를 비슷한 것끼리 그룹으로 묶는 작업
    - 소셜 미디어 사이트에 사진을 업로드하는 경우의 예
      - 업로드한 사진을 분류하려면 같은 사람이 찍힌 사진을 같은 그룹으로 묶을 수 있으나 사이트는 사진에 찍힌 사람이 누군지, 전체 사진 앨범에 얼마나 많은 사람이 있는지 알지 못함
      - 이때 가능한 방법은 사진에 나타난 모든 얼굴을 추출해서 비슷한 얼굴로 그룹 짓는 것임. 이 얼굴들이 같은 사람의 얼굴이라면 이미지들을 그룹으로 잘 묶은 결과임

SECTION 3.2 비지도 학습의 도전 과제

- 비지도 학습에서 가장 어려운 일은 알고리즘이 뭔가 유용한 것을 학습했는지 평가하는 것임
  - 비지도 학습은 보통 레이블이 없는 데이터에 적용하기 때문에 무엇이 올바른 출력인지 모름
  - 비지도 학습의 결과를 평가하기 위해서는 직접 확인하는 것이 유일한 방법일 때가 많음
  - 비지도 학습 알고리즘은 데이터 과학자가 데이터를 더 잘 이해하고 싶을 때 탐색적 분석 단계에서 많이 사용됨
  - 비지도 학습은 지도 학습의 전처리 단계에서도 사용됨. 비지도 학습의 결과로 새롭게 표현된 데이터를 사용해 학습하면 지도 학습의 정확도가 좋아지기도 하며 메모리와 시간을 절약할 수 있음
  - 전처리 메서드: 지도 학습 알고리즘에서 전처리와 스케일 조정을 자주 사용하지만, 스케일 조정 메서드는 지도 정보(supervised information)를 사용하지 않으므로 비지도 방식임

SECTION 3.5 군집

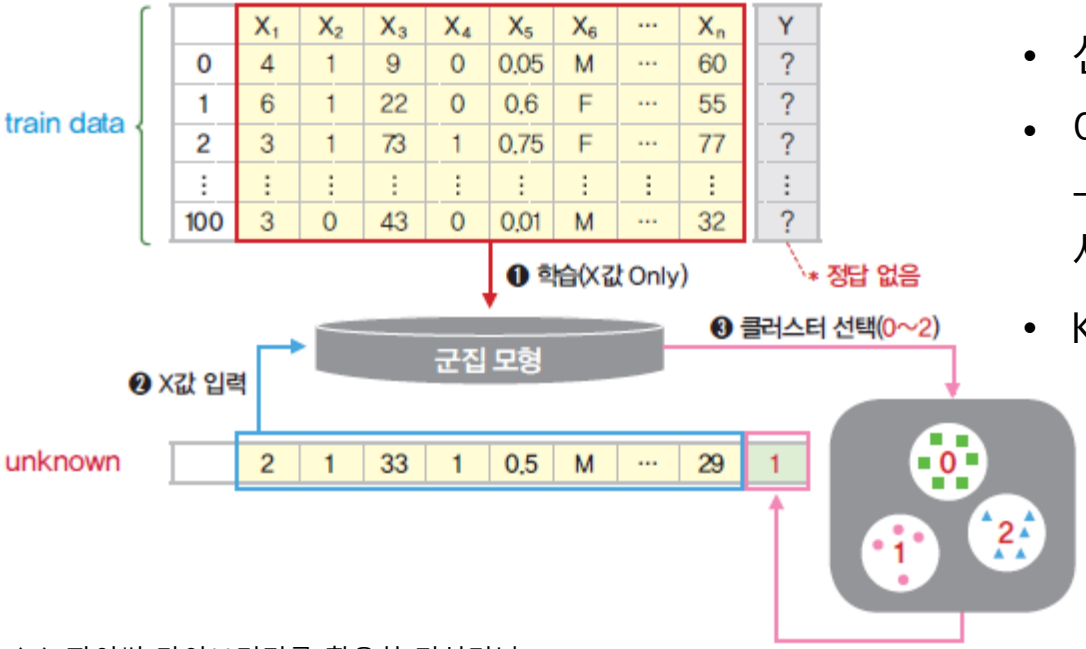
- 군집(clustering)은 데이터셋을 클러스터(cluster)라는 그룹으로 나누는 작업
  - k-평균 군집
    - k-평균 알고리즘이 실패하는 경우
    - 벡터 양자화 또는 분해 메서드로서의 k-평균
  - 병합 군집
    - 계층적 군집과 덴드로그램
  - DBSCAN
  - 군집 알고리즘의 비교와 평가
    - 타깃값으로 군집 평가하기
    - 타깃값 없이 군집 평가하기
    - 얼굴 데이터셋으로 군집 알고리즘 비교
  - 군집 알고리즘 요약



▲그림 3-24 k-평균 알고리즘으로 찾은 클러스터 중심과 클러스터 경계

SECTION 3.5 군집

- 군집(clustering)은 데이터셋을 클러스터(cluster)라는 그룹으로 나누는 작업
  - 군집 분석은 데이터셋 관측값이 갖고 있는 여러 속성을 분석하여 서로 비슷한 특징을 갖는 관측값끼리 같은 클러스터(집단)으로 묶는 알고리즘
    - 다른 클러스터 간에는 서로 완전하게 구분되는 특징을 갖기 때문에 어느 클러스터에도 속하지 못하는 관측값이 존재 할 수 있음
    - 관측값을 몇 개의 집단으로 나눈다는 점에서 분류 알고리즘과 비슷
      but 정답이 없는 상태에서 데이터 자체의 유사성만을 기준으로 판단하는 점



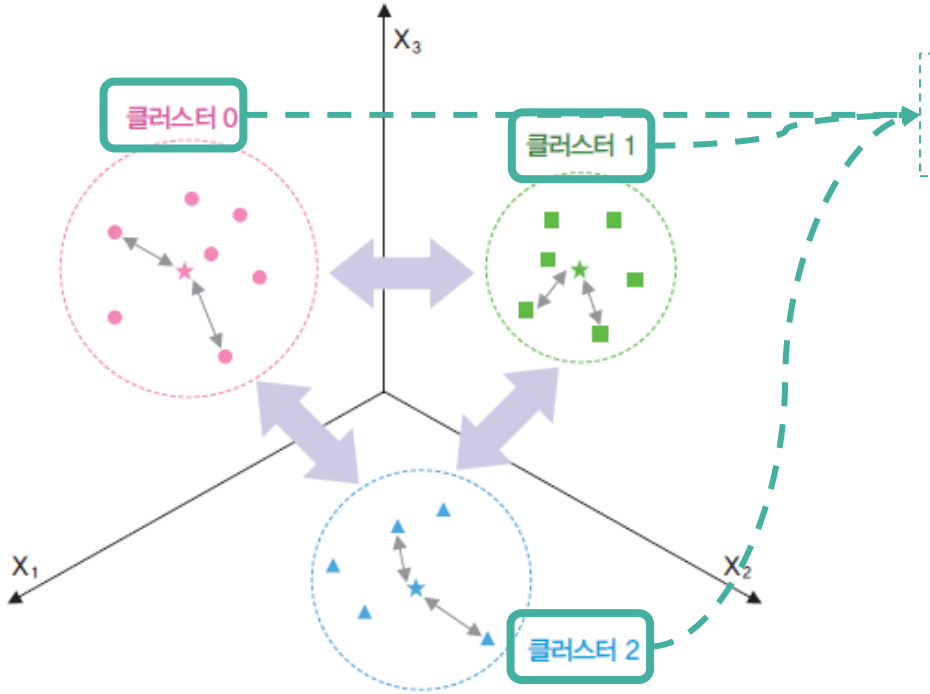
- 신용카드 부정 사용 탐지, 구매 패턴 분석 등 소비자 행동 특성 그룹화
- 어떤 소비자와 유사항 특성을 갖는 집단 구분
  → 갖은 집단 내의 다른 소비자를 통해
  새로운 소비자의 구매 패턴이나 행동 예측에 활용
- K-Means 알고리즘, DBSCAN 알고리즘

SECTION 3.5 군집 알고리즘

◦ k-평균 군집 (k-Means)

▪ k-평균 군집

- 데이터 간의 유사성을 측정하는 기준으로 각 클러스터의 중심까지의 거리를 이용
- 벡터 공간에 위치한 어떤 데이터에 대해서 k개의 클러스터가 주어졌을 때 클러스터의 중심까지 거리가 가장 가까운 클러스터로 해당 데이터를 할당
- 다른 클러스터 간에는 서로 완전하게 구분하기 위해 일정한 거리 이상 떨어져야 함



몇 개의 클러스터로 데이터를 구분할 것인지 생성하는 k 값에 따라 모형의 성능 달라짐

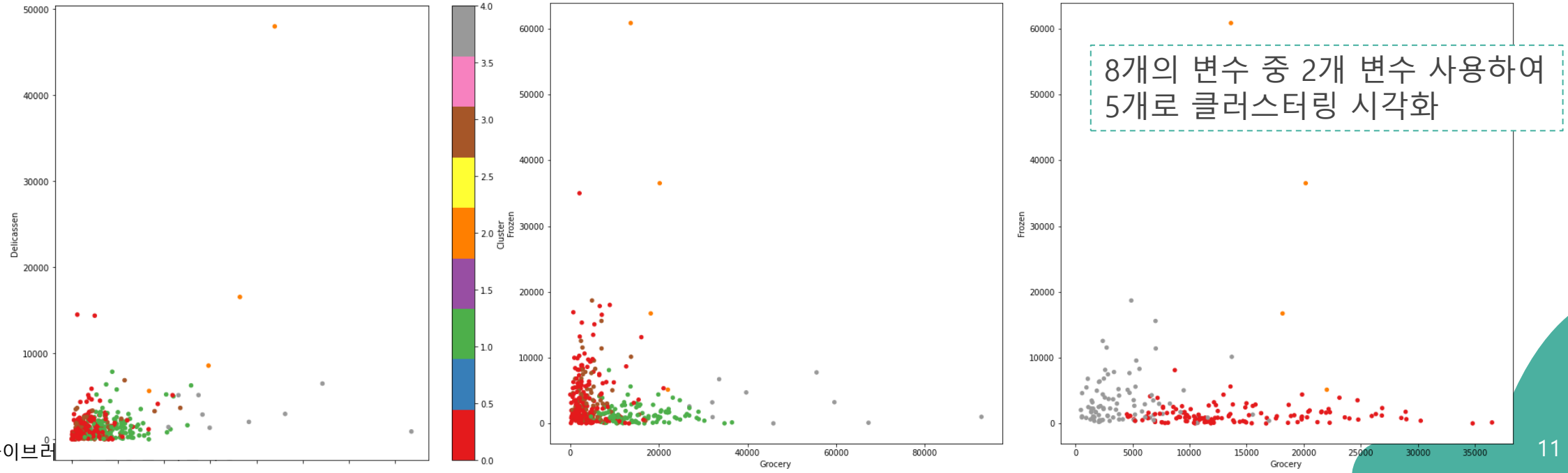
일반적으로 k가 클수록 모형의 정확도 개선
K 값이 너무 커지면 선택지가 너무 많아지므로 분석의 효과가 사라짐

SECTION 3.5 군집 알고리즘

◦ k-평균 군집 (k-Means) 예제

▪ Kmeans 도매업 고객 군집 분석

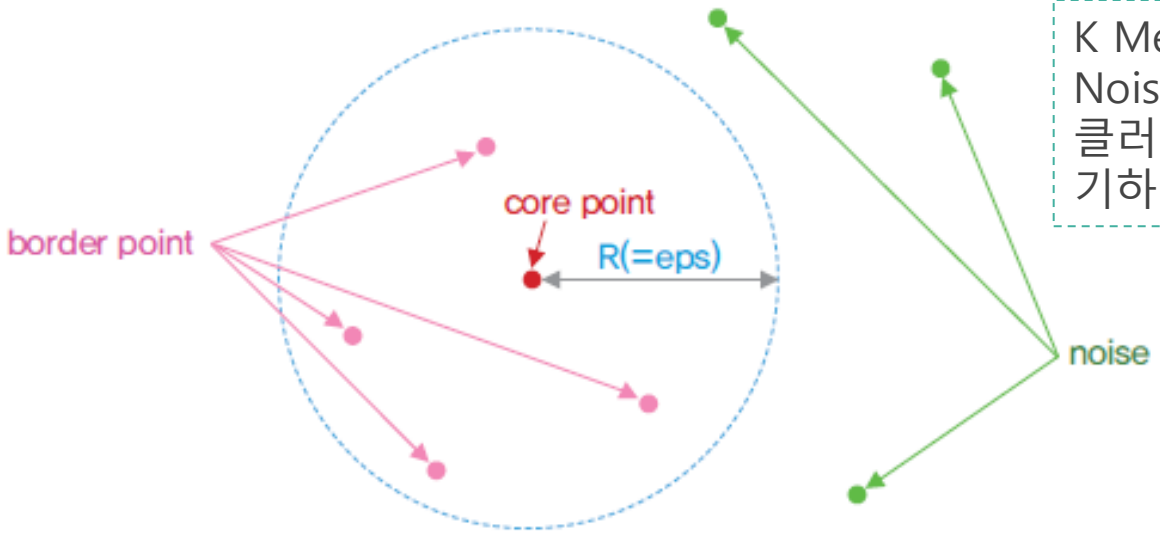
- UCI ML Repository의 도매업 고객(wholesale customers) 데이터 셋 사용
- https://archive.ics.uci.edu/ml/datasets/wholesale+customers
- 데이터 전처리 (StandardScaler() 데이터 정규화; 특정 범위 값으로 데이터 범위 축소)
- 데이터 셋 분리(훈련셋, 테스트셋)
  → Kmeans 메소드 n_clusters 5를 적용하여 모델 생성 → 클러스터 데이터 시각화



8개의 변수 중 2개 변수 사용하여 5개로 클러스터링 시각화

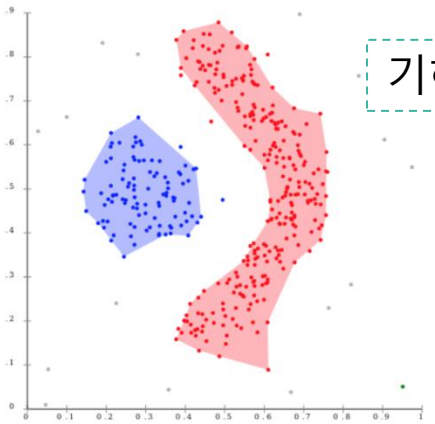
SECTION 3.5 군집 알고리즘

- DBSCAN (Density-based Spatial Clustering of Applications with Noise)
  - DBSCAN 밀도 기반 클러스터링
    - 데이터가 위치하고 있는 공간 밀집도를 기준으로 클러스터를 구분
    - 가지를 중심으로 반지름 R의 공간에 최소 M개의 포인트가 존재하는 점을 코어 포인트(core point)라고 함
    - 반지름 R 안에 다른 코어 포인트가 있는 경우 경계 포인트(border point) 라고 함
    - 코어 포인트, 경계 포인트도 속하지 않는 점을 Noise(or outlier)로 분류



* m(=min samples) : 반경 R 안에 들어오는 점의 최소 개수(5)

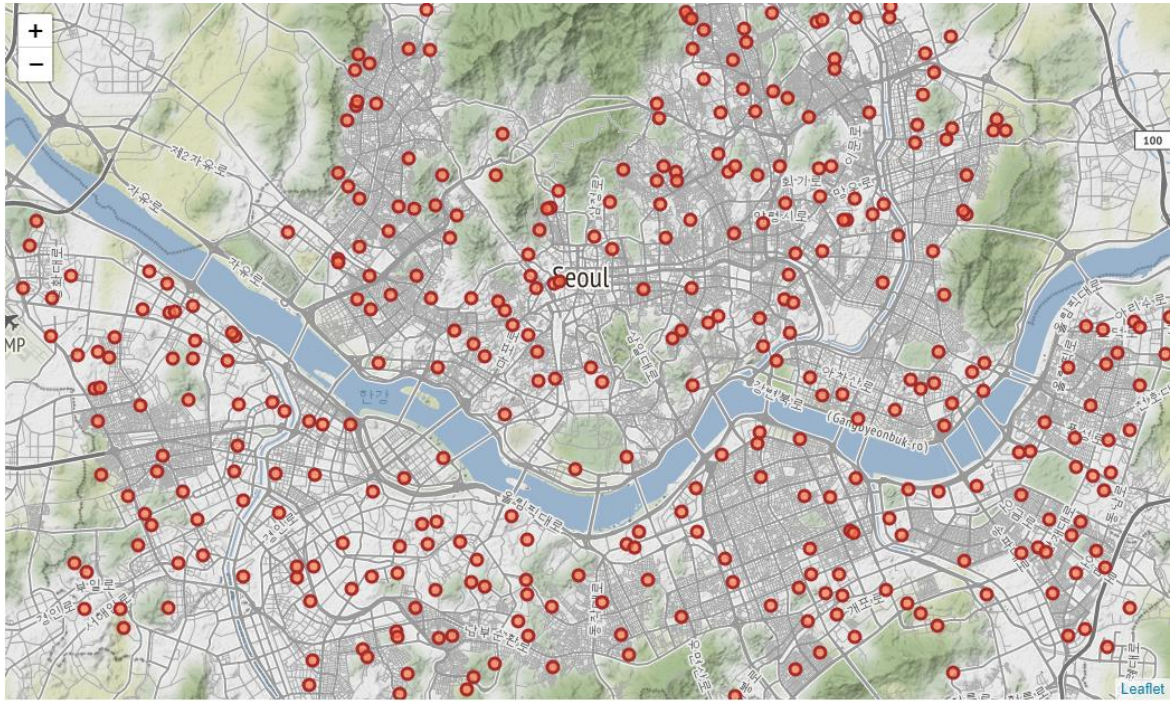
K Means와 같이 클러스터의 수를 정하지 않아도 됨
Noise point를 통하여, outlier 검출이 가능
클러스터의 밀도에 따라서 클러스터를 서로 연결하기 때문에 기하학적인 모양을 갖는 군집도 잘 찾을 수 있음

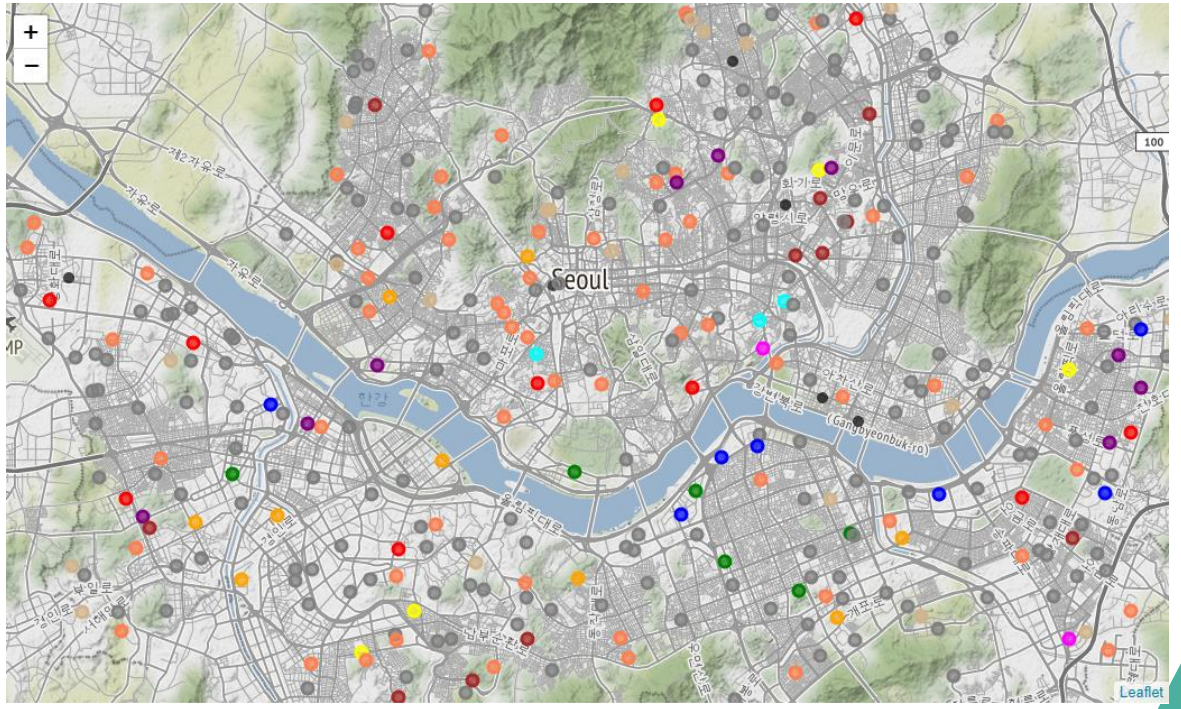


기하학적인 구조를 군집화 예

SECTION 3.5 군집 알고리즘

- DBSCAN (Density-based Spatial Clustering of Applications with Noise) 예제
  - DBSCAN 서울시 중학교 졸업생의 진로 형황 데이터셋을 사용한 밀도 기반 클러스터링
    - 학교알리미 공개용 데이터 중 서울시 중학교 졸업생의 진로 현황 데이터셋에서 고등학교 진학률 데이터를 활용하여 속성이 비슷한 중학교끼리 클러스터링
    - 클러스터링한 결과를 지도 시각화





SECTION 3.6 요약 및 정리

- 탐색적 데이터 분석과 데이터 전처리에 사용할 수 있는 여러 가지 비지도 학습 알고리즘 이해
- 데이터를 올바르게 표현하는 것은 지도 학습과 비지도 학습을 잘 적용하기 위해 필수적임
- 전처리와 분해 방법은 데이터 준비 단계에서 아주 중요한 부분임
- 분해, 매니폴드 학습, 군집은 주어진 데이터에 대한 이해를 높이기 위한 필수 도구이며, 레이블
- 정보가 없을 때 데이터를 분석할 수 있는 유일한 방법
- 지도 학습에서도 데이터탐색 도구는 데이터의 특성을 잘 이해하는 데 중요함
- 2차원 예제 데이터와 scikit-learn에 있는 실제 데이터셋인 digits, iris, cancer 데이터셋에 직접 군집과 분해 알고리즘을 적용하는 연습이 도움됨