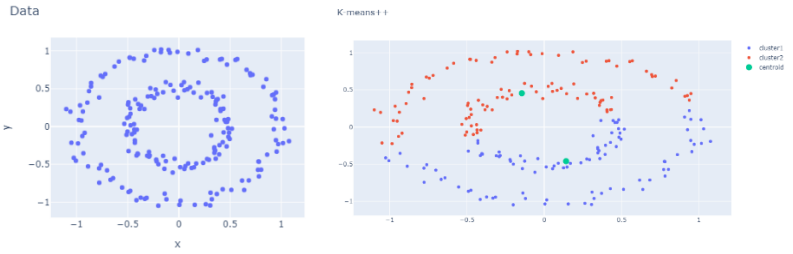
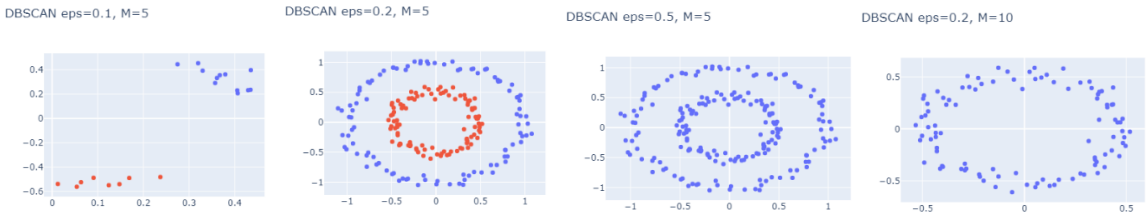
통계적 머신러닝 13장 과제

(1) K-means++



make_circles data 데이터를 생성하면 첫번째 plot과 같이 큰 원 군집, 작은 원 군집으로 나누는 것이 타당한 것으로 보인다. 군집 수를 2개로 하여 K-means++ 군집을 적용하면 두번째 plot과 같이 군집이 선형적으로 나누어져 바르게 군집이 형성되지 않는다.

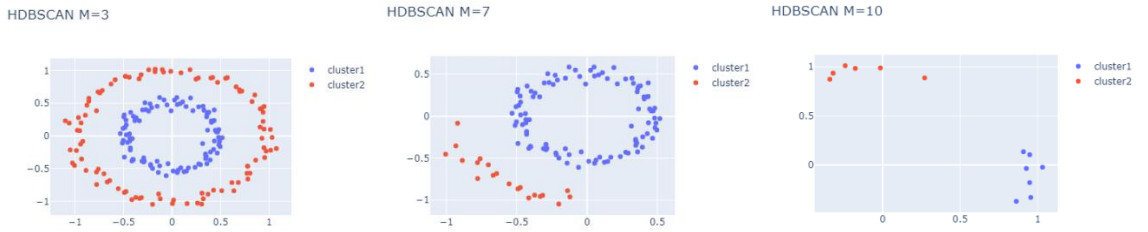
(2) DBSCAN



데이터가 비선형일 때 DBSCAN과 HDBSCAN이 더 잘 분류되고 이들은 초모수에 민감하다. DBSCAN을 적용할 때, eps는 2차원 공간에서 반지름이 eps인 원이다. 따라서 eps가 0.1로 작으면 원의 반지름이 작아서 많은 관측치가 잡음분류로 분류되고 eps가 0.5로 크면 원의 반지름이 커서 군집의 개수가 작아져 군집의 기능을 하지 못한다.

하나의 군집을 형성하기 위해 최소의 관측치의 개수인 M은 너무 크면 군집에 속하지 못한 관측치(잡음)이 많아져 4번째 plot(eps=0.2,M=10)처럼 데이터가 많이 잘린 모습을 보인다. Eps가 0.2, M이 5일 때 잘 분류되었다.

(3) HDBSCAN



M이 3일 때 잘 분류되었다. M이 7, 10으로 커질 때, 군집에 속하지 못한 관측치(잡음)가 많아져 잘려나간 관측치가 많아 군집이 잘 형성되지 않는다. M이 커지면 최소 관측치의 개수를 만족하는 것이 어려워져 군집의 수가 줄어들기 때문이다.