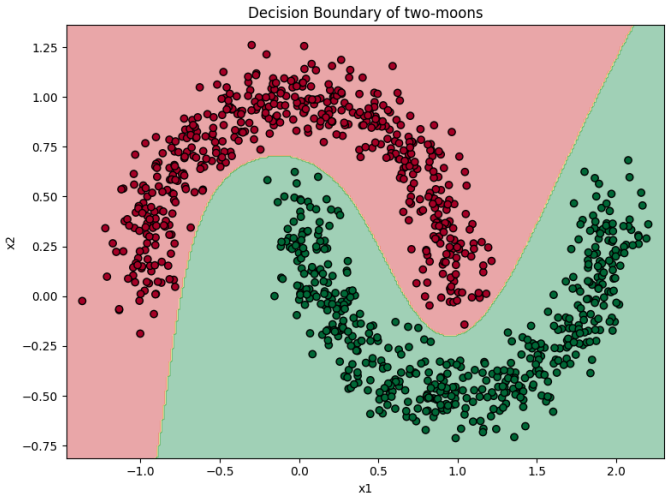
기계학습 과제 분석 리포트

학과: 인공지능공학과 학번: 12223550 이름: 서보성

1. Feature Mapping + Logistic Regression for “two-moons”

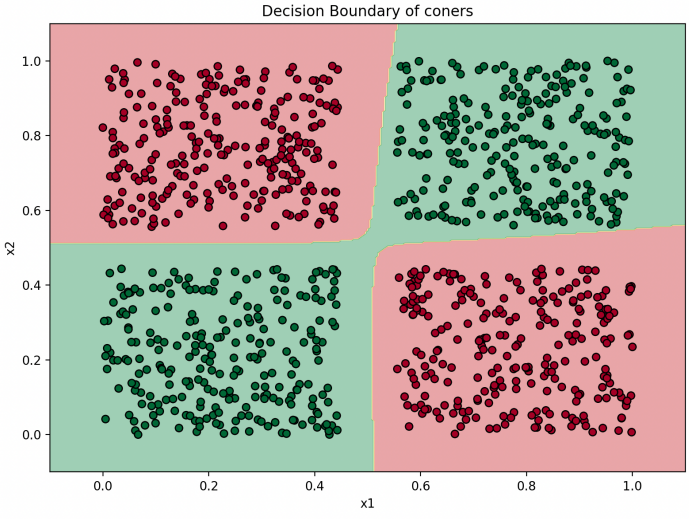


로지스틱 회귀를 사용해 “two-moons” 데이터셋을 분류한 결과 데이터를 잘 분류하는 결정경계를 생성할 수 있었다. 특히 feature mapping을 통해 비선형 구조의 데이터를 잘 분류할 수 있었다. 하이퍼파라미터는 learning rate와 epoch만을 사용하였으며, 이는 k-fold cross validation으로 조합을 결정하였다.

최종 loss는 0.01226이고, 학습된 모델 파라미터는 다음과 같다.

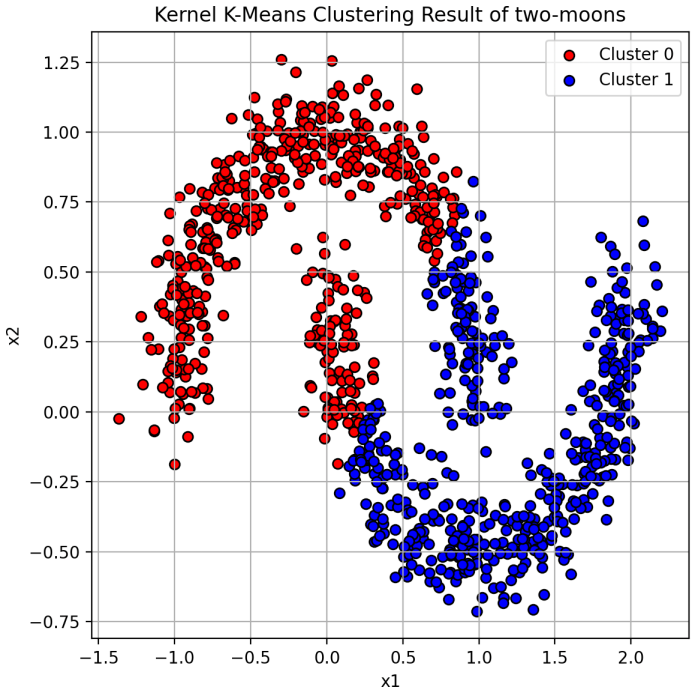
[5.01678149e+01 -4.58760049e+01 -1.03965954e+01 -1.41139508e+01 -2.55959636e+01
-8.38648583e+01 -1.93194339e+01 -8.84850017e-03 -4.46806945e+01 -3.44395614e+01
-1.45477639e+01 -4.17096759e+00 -1.19474912e+01 -2.38812702e+01 -2.71739421e+01
-1.41910388e+01-1.10178557e+00 -1.09103131e+01 -1.01475795e+00
-2.66736580e+017.41335564e+01]

2. Feature Mapping + Logistic Regression for “corners”



로지스틱 회귀를 통해 "corner" 데이터 셋을 분류한 결과 데이터를 잘 분류하는 결정경계를 생성할 수 있었다. 특히 feature mapping을 통해 비선형 구조의 데이터를 잘 분류할 수 있었다.
마찬가지로 k-fold cross validation을 통해 learning rate와 epoch를 결정하였다.
최종 loss는 1.738116이고, 학습된 모델 파라미터는 다음과 같다.
[48.09505151 -53.62580337 -51.81655184 -49.65334126 46.1856364 -52.34735961 -37.72409484 59.66814551 60.1514888 -40.0042387 -27.30981395 54.34631885 71.04151063 56.26651477 -28.4457305 -19.48273937 46.05532756 62.42674704 63.5884852 49.07208303 -19.8747927]

3. RBF Kernel + K-means for "two-moons"

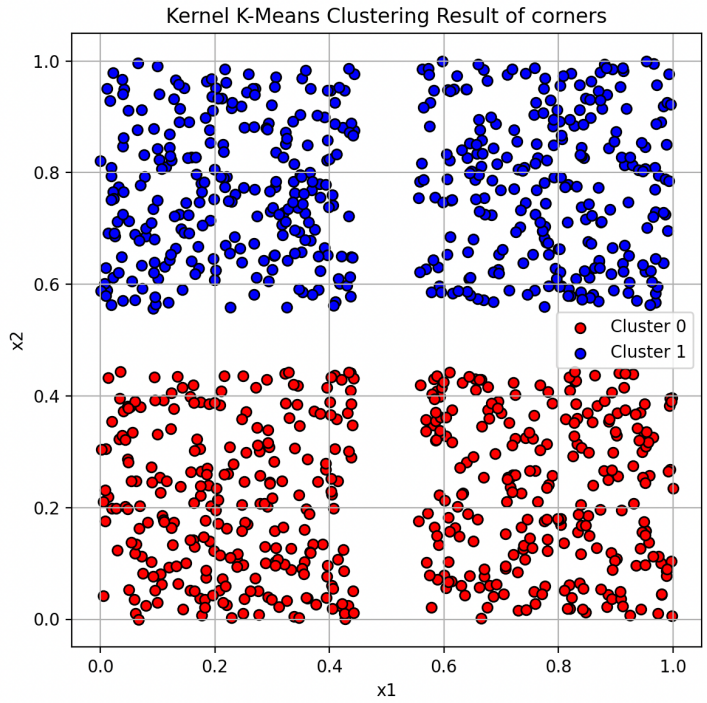


K-means를 통해 "two-moons" 데이터를 학습한 결과 대체로 곡선 구조를 따라 분류 해내지만, 일부 오차가 발생하였다.
거리 기반의 비지도 학습 알고리즘이기 때문에 가까운 데이터끼리 군집을 이루게된다.
이로인해 로지스틱 회귀에 비해 완벽히 데이터를 분류해내지 못했다.
RBF의 gamma값을 0.5, 1, 3 등을 시도해보았지만, 결과는 비슷하였다. 또한 최대 반복수를 늘렸을 때에도 이미 최대 반복수에 도달하지 못하고, 수렴하기 때문에 의미가 있지 않았다.

또한 처음에 초기화에 따라 결과가 달라질 수 있음을 기대하고, 여러번 시도하였으나, 큰 변화는 없었다. 오차가 있는 부분의 경계선에서 서로의 경계가 조금씩 달라지는 것 외에는 차이가 없었다.

결과적으로, K-means에서는 'two-moons' 데이터를 잘 분류해내지 못하였다.

4. RBF Kernel + K-means for “corners”



K-means를 통해 “corners” 데이터를 학습한 결과 데이터 셋의 절반을 잘못 예측하는 결과를 보였다. 클러스터 중심이 데이터 구조를 잘 반영하지 못하였다. 이는 거리 기반의 비지도 학습 알고리즘이기 때문에 서로 가까이 밀집해 있는 데이터를 같은 군집으로 매핑한다. 또한 gamma 값이나, 최대 반복수를 바꿔주는 것이 결과에 변화를 일으키지 않았다.
명확하게 데이터가 분류되어 존재하기 때문에 마찬가지로 초기화에 따라 결과가 달라지지 않았다. 결과적으로 K-means에서는 ‘corners’ 데이터를 잘 분류하지 못하였다.

Feature mapping + Logistic Regression으로는 ‘two-moons”와 “corner” 데이터 셋 모두 비교적 정확한 결정경계를 형성하였다. ‘two-moons’ 데이터 셋은 선형모델로는 분리가 불가능 하지만, 특성변환을 통해 곡선 경계를 학습하였고, 최종 loss가 약 0.01로 매우 낮게 나왔다.
‘corners’ 데이셋의 경우도 마찬가지로 특성변환을 통해 데이터셋을 잘 분류하였다.

하지만 RBF Kernel + K-Means 알고리즘에서는 RBF Kernel을 통해 ‘two-moons’ 데이터셋의 반달 구조를 어느정도 분리해냈지만, 완벽한 결정경계를 형성하지 못하였다. ‘corners’ 데이터 셋에서는 중심값이 상/하로 분할되었고, 이로인해 절반 가까운 샘플이 잘못 할당되었다. 이는 하이퍼파라미터를 변화시켜도 비슷한 결과가 도출되었다.

결론적으로 Logistic Regression은 특성변환을 통해 비선형 구조 데이터에 대해 높은 정확도를 가지는 결정경계를 형성하는 강력한 분류기임을 보였고, Kernel K-means는 데이터 구조에 따라 성능이 크게 달라지고, 정확한 클러스터링이 되지 않을 수 있다는 것을 보였다.