1. 지도학습(supervised learning)

[Feature Mapping + Logistic Regression]
i) feature mapping을 통해 비선형 분포를 다항식 형태로 확장
ii) 확장된 특성 공간에서 logistic regression을 적용
iii) decision boundary는 sigmoid 함수 기반의 확률적 경계로 형성
iv) 학습은 gradient descent를 통해 theta parameter를 업데이터 하면서 이루어짐

<a. Two Moons dataset>

결과 : feature mapping degree를 1~10까지 실험해본 결과, 5~6 이상의 차원에서는 정확도 100%로 두 반달 모양의 클래스를 효과적으로 분리하였음. 반면 1~4 차원에서는 중간 부분의 데이터 포인터들이 제대로 분리되지 않는 것을 확인함.

정확도 : 적절한 차원(5~6)에서는 95~100% 수준의 높은 성능을 보임

분석 : 기본적인 logistic regression은 선형 분류기이지만, feature mapping을 통해 비선형 데이터도 분류가 가능함을 확인함. 다만 degree가 너무 낮은 경우 중앙부 데이터 포인터들을 제대로 분리하지 못하는 경향을 보였고, degree가 너무 높은 경우 overflow가 일어나거나 원하는 모양과 전혀 다른 형태로 분리되는 것을 확인하였다.

<b. Corners dataset>

결과 : feature mapping degree를 1~10까지 실험해본 결과, 5~6 이상의 차원에서는 네 개의 꼭짓점에 해당하는 분포를 2개의 클래스로 정확히 분리함. 다만 degree가 너무 낮은 경우 우측 상단과 좌측 하단의 데이터 포인터들을 정확한 경계로 나누지 못하고 대각선으로 횡단하는 결정경계가 생기는 것을 확인하였음.

정확도 : 적절한 차원(5~6)에서는 95~100% 수준의 높은 성능을 보임

분석 : feature mapping을 통해 중심 대각선 방향으로 나뉘는 형태의 분포를 효과적으로 분리함. 하지만 데이터가 비정형적일 경우, feature mapping으로 공간을 확장해도 모델이 표현할 수 있는 복잡도보다 데이터 구조가 복잡하면 decision boundary가 왜곡될 수 있음을 확인함.

2. 비지도 학습(unsupervised learning)

[Kernelized K-means Clustering]
i) K-means 방법 채택 : 데이터 포인트를 가장 가까운 centroid에 할당하는 과정 + 데이터 포인트를 기준으로 centroid를 업데이트 하는 과정
ii) centroid와 각 데이터 포인트 간의 거리를 계산하는 방식으로 기존의 유클리드 거리 대신, RBF 커널 기반의 비선형 유사도를 적용
iii) 커널 행렬을 이용해 각 샘플과 클러스터 간의 거리 계산을 수행
iv) 초기 라벨은 랜덤하게 지정하고 반복적으로 갱신

<a. Two Moons dataset>

결과 : 여러 gamma와 seed 조합을 실험해본 결과, 최적의 조합인 gamma(10), seed(42)에서 두 개의 반달형 군집이 가장 잘 분리됨. 각 군집의 centroid는 반달형 군집의 중심부에 위치하는 것을 확인함. B군집으로 할당되었으면 더 이상적이었을 2~3개의 일부 데이터 포인트들이 A 군집의 centroid와 매우 가까워 A 군집으로 할당되는 것을 확인함.

정확도 : 최적의 gamma, seed 조합에서 99.6% 정확도를 달성, 높은 수준의 성능을 보임

분석 : 비정형적 데이터에서 유클리드 거리를 사용하는 기존의 K-means는 이상적인 군집을 분리하는데 실패하지만, 커널 함수를 사용하면 비선형 구조도 잘 분리되는 것을 확인함. 핵심은 적절한 gamma와 seed를 설정하는 것이며 너무 작으면 선형에 가까워지고 너무 크면 노이즈에 민감해지는 것을 확인함.

<b. Corners dataset>

결과 : 여러 gamma와 seed 조합을 실험해본 결과, 최적의 조합인 gamma(3.24), seed(10)에서 네 개의 꼭짓점 중 두 대각선 방향을 하나의 군집으로 묶는 가장 이상적인 군집이 형성됨. 각 군집의 centroid는 각각 대각선에 위치한 같은 군집의 중간 부분에 위치하는 것을 확인함. 결과적으로 두 군집의 centroid가 매우 비슷한 위치에 형성됨.

정확도 : 최적의 gamma, seed 조합에서 100% 정확도를 달성, 높은 수준의 성능을 보임

분석 : 모양이 균일하지 않고 분산이 적기 때문에 적절하지 않은 gamma, seed 값에서는 클러스터 중심이 왜곡되는 것을 확인함. 마찬가지로 핵심은 적절한 gamma와 seed를 설정하는 것이기에, 최적의 조합 탐색을 위해 각 gamma, seed를 이중 반복문으로 돌리면서 최적의 조합을 탐색함. 결과적으로 gamma(3점대) seed(2)에서 가장 이상적인 군집을 분리할 수 있게 되었고, 난이도가 있는 군집 구조에서도 gamma & seed 조절로 충분히 성능 향상이 가능함을 확인함.

3. 고찰 및 회고

<ML 과제를 수행하며 느낀 수업과의 연관성>

ML 수업에서는 보통 수학적인 부분을 다루면서 “특정 모델, 알고리즘이 왜 이렇게 작동하는가”를 배웠는데 이번 과제를 통해 모델들을 직접 구현해보면서 해당 수식들이 실제 코드로 어떻게 구현되는지를 체감할 수 있었음. 단순히 공식을 외우는 데 그치지 않고, 이를 기반으로 모델이 학습하고 예측하는 전 과정을 코드로 재현해보며 수업에서의 이론이 실제로 어떻게 쓰이는지 명확히 연결할 수 있었음. 예를 들어 커널을 사용한 K-means에서 거리 계산 방식을 유클리드 방식 거리 계산이 아닌 커널을 활용한 거리 계산 방식을 채택하므로서 비선형석인 클러스터링을 잘 수행하는 것이 인상 깊었음.

<두가지 모델을 현실 과제에 적용했을 때 얻을 수 있는 효율성>

1. Feature Mapping + Logistic Regression

적용 분야 : 이메일 스팸 분류, 의료 진단

효율성 : 설명 가능성이 높고, 계산 비용이 낮은편. 학습이 빠르고 모델이 비교적 간단해 실제 운영 시스템에 바로 적용 가능할 것으로 예상됨. feature mapping을 통해 비선형 패턴도 반영 가능하므로 단순한 선형 분류 모델보다 다양한 데이터에 유연하게 대응이 가능할 것으로 보임

2. Kernelized K-means Clustering

적용 분야 : 이미지 분할, 고객 행동 기반 마케팅 세분화

효율성 : 커널을 사용하므로서 비선형 구조를 가진 데이터도 효과적으로 군집화 할 수 있어 실제 문제에서 훨씬 높은 분할력을 확보함. 사전 레이블 없이도 비슷한 속성을 가진 데이터들을 자동으로 분류해낼 수 있어 데이터 탐색이나 사전 분석 단계에서 유용할 것으로 보임.