Machine Learning 과제 Report

2023149028 정민지

1. Ensemble Method

1.1)

Ensemble 기법 중에 Decision Tree, Random Forest 두 가지 모델을 사용해 Breast Cancer Dataset에 대해 Binary Classification을 진행했습니다.

그림 1. Learning Curve for Decision Tree

첫 번째 그래프는 Decision Tree 모델의 학습곡선을 나타냅니다. 초기에는 훈련 데이터와 테스트 데이터 모두에서 예측 성능이 낮지만, 학습 데이터가 증가함에 따라 정확도가 점진적으로 향상됩니다. 훈련 정확도는 거의 1에 가까운 값을 유지하지만, 테스트 정확도는 상대적으로 낮게 나타나는데, 결정 트리가 학습 데이터에 overfitting되는 경향이 있음을 알 수 있고, 일반화 성능이 제한적일 수 있음을 보여줍니다.

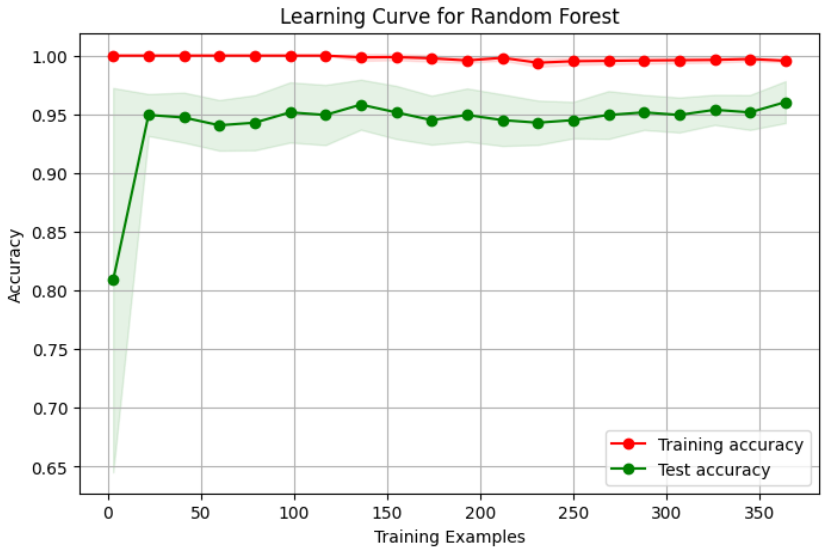


그림 2. Learning Curve for Random Forest

두 번째 그래프는 Random Forest 모델의 학습곡선입니다. 랜덤 포레스트는 다수의 개별 결정 트리를 결합하여 예측을 수행하는 앙상블 학습 기법으로, 단일 결정 트리보다 더 나은 일반화 성능을 보입니다. 훈련 정확도는 높은 수준을 유지하며, 테스트 정확도 또한 안정적으로 높은 값을 기록하고 있습니다. 특히, 결정 트리와 비교했을 때 테스트 정확도의 변동성이 적으며 보다 안정적인 학습을 수행합니다.

1.2)

결정 트리는 훈련 데이터에 대해 과적합되는 경향이 있으며, 테스트 데이터에서의 성능이 상대적으로 낮게 나타났습니다. 이는 단일 트리 구조에서 발생할 수 있는 한계로, 복잡한 데이터에서는 일반화 성능이 떨어질 가능성이 있습니다.

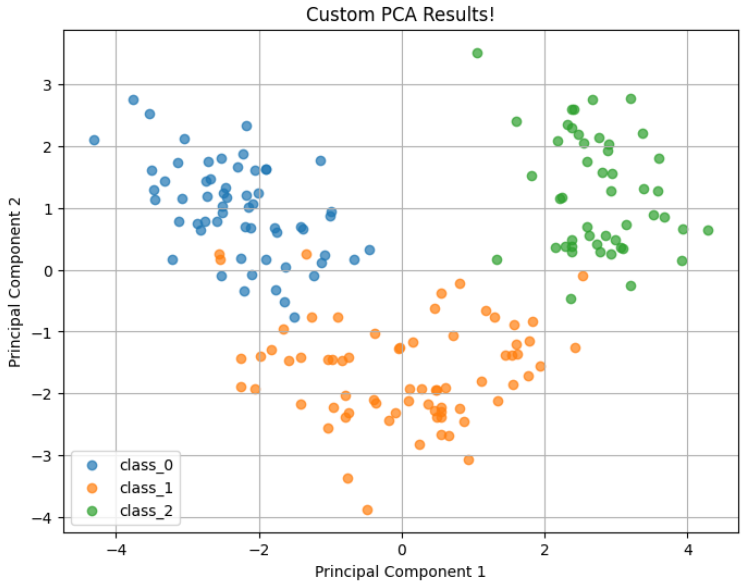
반면, 랜덤 포레스트 모델은 여러 개의 결정 트리를 조합하여 예측을 수행하기 때문에, 더 높은 정확도와 안정적인 성능을 보였습니다. 특히, 테스트 데이터에 대해서도 비교적 높은 성능을 유지하며, 결정 트리에 비해 과적합이 덜 발생하는 경향을 확인할 수 있습니다.

따라서, 랜덤 포레스트 모델이 결정 트리보다 더 나은 성능을 보였고, 해당 데이터셋에 더 적합한 모델이라고 판단할 수 있습니다.

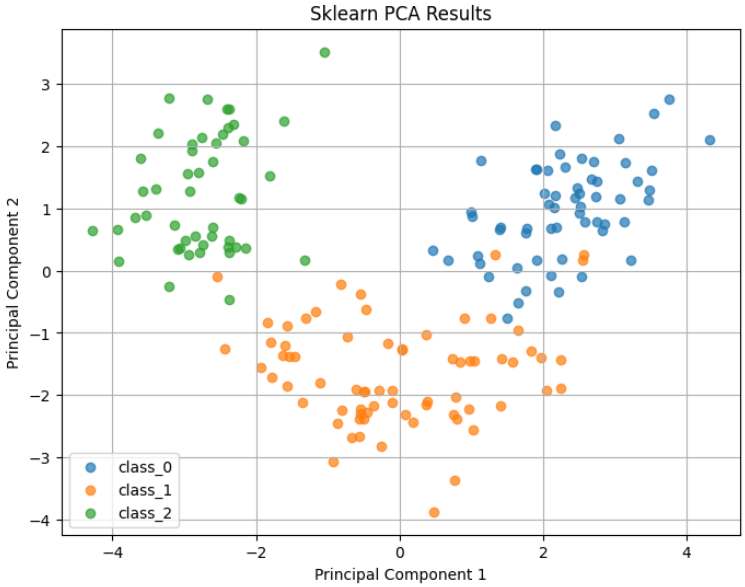
2. PCA

2.1) 주성분 분석(PCA, Principal Component Analysis)은 고차원 데이터를 저차원으로 축소하면서도 데이터의 분산을 최대한 보존하는 기법입니다. 이를 통해 차원 축소 후에도 데이터의 구조를 유지하며, 시각화와 분석을 용이하게 할 수 있습니다. 이번 실험에서는 Wine Dataset을 사용해서 직접 PCA 알고리즘을 구현하고, 이를 통해 데이터의 차원을 2차원으로 축소한 후 시각화했습니다.

먼저, Custom PCA를 구현하고 데이터를 2차원으로 축소한 결과를 그래프로 나타냈습니다. 차원 축소 후에도 데이터가 클래스별로 잘 구분되며, PCA가 데이터를 효과적으로 압축하면서도 중요한 정보는 유지하는 것을 확인할 수 있습니다.



다음으로, Sklearn PCA를 사용하여 동일한 데이터를 2차원으로 축소한 결과를 비교하였습니다. 두 방법을 비교한 결과, Custom PCA와 Sklearn PCA 모두 유사한 형태로 데이터를 변환하며, 차원 축소 이후에도 클래스 간 경계가 명확하게 유지됨을 확인할 수 있었습니다.



2.2)

PCA의 장점 중 하나는 차원의 저주(Dimension Curse) 완화입니다. 고차원 데이터는 계산량이 많아지고 모델의 복잡성이 증가할 수 있지만, PCA를 활용하면 차원을 줄여 연산 속도를 향상시키고 모델의 과적합을 방지할 수 있습니다.

또한, 데이터 시각화가 가능하다는 점도 큰 장점인데, 고차원 데이터를 2차원 또는 3차원으로 축소하면, 데이터의 패턴을 쉽게 파악할 수 있어 인사이트를 도출하는 데 유리합니다.

마지막으로, 노이즈 제거 효과를 기대할 수 있습니다. PCA는 데이터의 주요 성분을 남기고 상대적으로 중요하지 않은 차원을 제거하는 과정에서 불필요한 노이즈를 줄이는 역할을 하므로, 데이터의 품질을 향상시킬 수 있습니다.

반면, PCA에는 몇 가지 단점도 존재합니다.

가장 큰 단점 중 하나는 해석이 어렵다는 점입니다. PCA를 거친 데이터는 원래의 특징(feature)들과 직접적인 연관이 없기 때문에, 특정 주성분이 실제 데이터에서 어떤 의미를 가지는지 해석하기가 어려울 수 있습니다. 또한, PCA는 선형 변환(Linear Transformation)을 기반으로 하기 때문에 데이터의 선형적인 관계만 반영할 수 있습니다. 따라서, 복잡한 비선형 관계를 가진 데이터에서는 성능이 저하될 가능성이 있습니다. 마지막으로, 정보 손실 가능성도 고려해야 합니다.

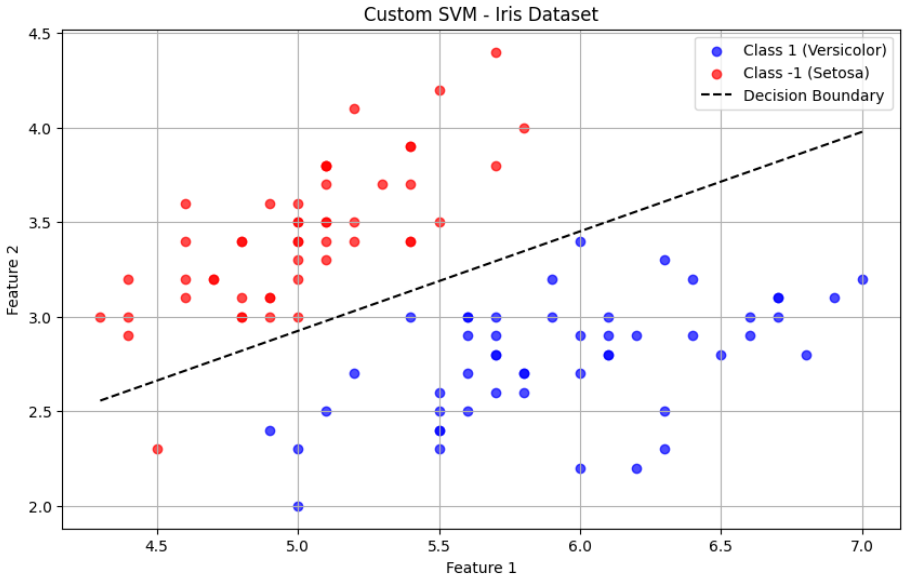
주성분을 선택하는 과정에서 일부 정보를 제거하게 되므로, 차원을 지나치게 축소할 경우 중요한 정보까지 손실될 위험이 있습니다.

PCA 이외의 차원 축소 기법으로 t-SNE (t-Distributed Stochastic Neighbor Embedding) 가 있습니다. PCA가 선형 변환을 기반으로 하는 반면, t-SNE는 비선형 차원 축소 기법으로 데이터의 지역적 구조를 보존하는 데 뛰어난 성능을 보입니다.

특히, 고차원 데이터의 시각화(2D, 3D)에 유용하며, 이미지 데이터나 NLP(자연어 처리)에서 자주 사용됩니다. t-SNE는 PCA와 달리 거리에 기반한 확률적 임베딩 방식을 사용하여 데이터의 클러스터링 구조를 명확하게 시각화할 수 있는 장점이 있지만, 단점으로는 계산 비용이 높고 새로운 데이터 포인트를 쉽게 추가하기 어렵다는 점이 있습니다.

3) SVM

서포트 벡터 머신(SVM)은 주어진 데이터에서 최적의 결정 경계를 찾아내는 모델입니다. 이번 과제에서는 이진 분류 문제를 다루기 위해 Setosa와 Versicolor 두 클래스를 선택하여 하드 마진 SVM을 직접 구현하였습니다. SVM 모델을 적용한 뒤 결정 경계를 시각화한 결과, 두 클래스가 뚜렷하게 구분되었으며, 모델이 각 클래스 간의 최적의 경계를 효과적으로 학습했음을 확인할 수 있었습니다.



각 모델의 학습곡선과 성능을 분석한 결과, 랜덤 포레스트 모델이 결정 트리보다 더 높은 정확도와 안정적인 성능을 보여주었습니다. 또한, PCA를 활용한 차원 축소 과정에서도 데이터의 구조가 잘 유지되었으며, 고차원 데이터를 효과적으로 시각화할 수 있었습니다. SVM의 경우, 두 클래스를 명확하게 구분하며 뛰어난 성능을 나타냈습니다. 이러한 결과를 종합해 볼 때, 랜덤 포레스트와 SVM이 주어진 문제에서 가장 효과적인 모델로 판단됩니다.