출석수업 과제물(평가결과물) 표지(온라인제출용)

**교과목명 : 머신러닝**

**학 번 : 202234-153799**

**성 명 : 한승환**

**강 의 실 : 울산지역대학 (비대면)호**

**연 락 처 : 010-2862-0200**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**머신러닝 KNN 분류기 과제수행 보고서**

목차

1. K값의 변화에 따른 결과값 비교 및 정확도 분석

2. 거리측정 방식에 따른 결과 분석

3. 최종 성능 분석 결론

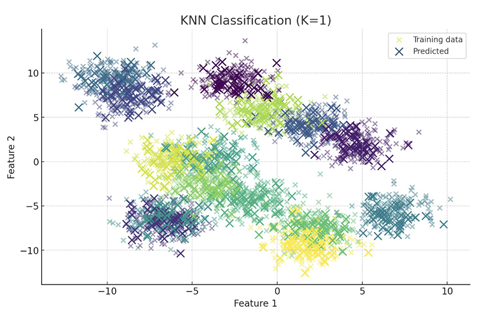
1. **K값의 변화에 따른 결과값 비교 및 정확도 분석**

사용된 거리 측정 방법: 유클리드 거리측정

훈련 데이터 비율: 80%

테스트 데이터 비율: 20%

**K=1일 때 결과값**



정확도: 73.3%

**K=5일 때 결과값**

A diagram of a number of colored crosses

Description automatically generated with medium confidence

정확도: 79.8%

**K=10일 때 결과값**

A diagram of a number of colored dots

Description automatically generated with medium confidence

정확도: 78.5%

**K=15일 때 결과값**

A graph showing a number of colored crosses

Description automatically generated

정확도: 78.6%

**K=20일 때 결과값**

A diagram of a number of colored crosses

Description automatically generated with medium confidence

정확도: 80.0%

**K=30일 때 결과값**

A graph showing a number of colored crosses

Description automatically generated with medium confidence

정확도: 80.3%

**결과요약**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| K값 | 1 | 5 | 10 | 15 | 20 | 30 |
| 정확도 | 73.3% | 79.8% | 78.5% | 78.7% | 80.0% | 80.3% |

K 값이 작을수록(1~5) 정확도가 높게 나올 수도 있지만, 오버피팅의 문제로 인하여 결과값의 변동성이 커질 수 있다는 것을 확인하였습니다.

10~20 사이의 K 값은 비교할 이웃의 수가 많아지므로 노이즈가 덜 발생하였고, 이로 인해 결과값이 더 안정적인 것을 관찰할 수 있습니다.

20 이상의 K 값은 결정경계를 더욱 섬세하게 해 더욱 높은 정확도를 만들어 냈습니다.

선험확률에 의존하여 분류를 수행하지 아니하도록 너무 큰 K 값을 사용하는 것을 지양했습니다.

실험결과 20에서 30 사이의 K 값이 가장 큰 정확도를 산출해 내었으며, 값의 변동범위가 가장 안정적이었습니다.

1. **거리 측정 방식에 따른 결과 분석**

K값은 30으로 동일하게 설정하였음

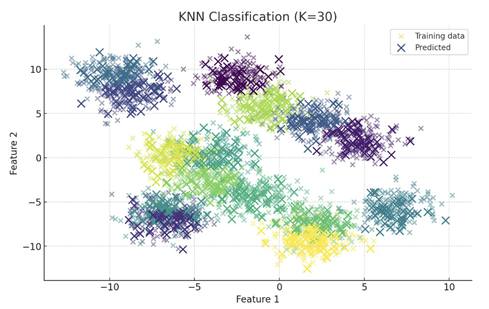
시도한 다른 거리측정방식의 리스트

- 마할라노비스 거리

- 코사인 거리

- 내적 거리

**마할라노비스 거리 측정 테스트 결과**



정확도: 80.0%

**코사인 거리 측정 테스트 결과**

A diagram of a number of colored crosses

Description automatically generated with medium confidence

정확도: 63.3%

**내적 거리 측정 테스트 결과**

A diagram of a number of colored crosses

Description automatically generated with medium confidence

정확도:30.5%

**유클리드 거리 측정 방식과 비교**

**마할라노비스 거리**: 유클리드 거리와 유사한 정확도를 달성했으며, 이는 이 경우 특성 상관관계가 분류에 큰 영향을 미치지 않을 수 있음을 나타냅니다. 그러나 마할라노비스 거리는 더 복잡한 특성 관계를 가진 데이터 집합에서 더 유용할 수 있습니다.

**코사인 거리**: 더 낮은 정확도는 코사인 유사성이 이 특정 데이터 집합에서 유클리드 거리나 마할라노비스 거리만큼 포인트 간의 공간적 관계를 효과적으로 포착하지 못할 수 있음을 나타냅니다.

**내적 거리**: 상당한 정확도 감소는 내적 유사성이 이 데이터 집합에 대한 거리 관계를 효과적으로 포착하지 못함을 시사합니다.

1. **최종 성능 분석 결론**

이 데이터 집합에 대하여 유클리드 거리와 마할라노비스 거리를 이용하여 데이터 사이의 거리를 측정하고, K의 값이 클수록(20~30) 정확도와 안정성이 충족되는 데이터 분류 결과를 산출할 수 있었습니다.

적절한 K값 선정: 훈련 데이터에 대한 교차 검증을 통해 주어진 데이터셋에 대한 최적의 K값을 확인하는 데 도움이 될 수 있습니다.

적절한 거리 측정 방법 선정: 데이터 사이의 거리 측정 방식의 적합성은 데이터셋의 특성에 맞아야 합니다. 공간적이거나 저차원 데이터의 경우 유클리드 거리와 마할라노비스 거리가 선호되며, 고차원 또는 방향 중심의 맥락에서는 코사인 유사성이나 내적이 유용할 수 있습니다.

이 실험을 통해, KNN 분류에서 최적의 성능을 달성하기 위해 K값과 거리 측정 방식을 모두 조정하는 것의 중요성을 알 수 있습니다.