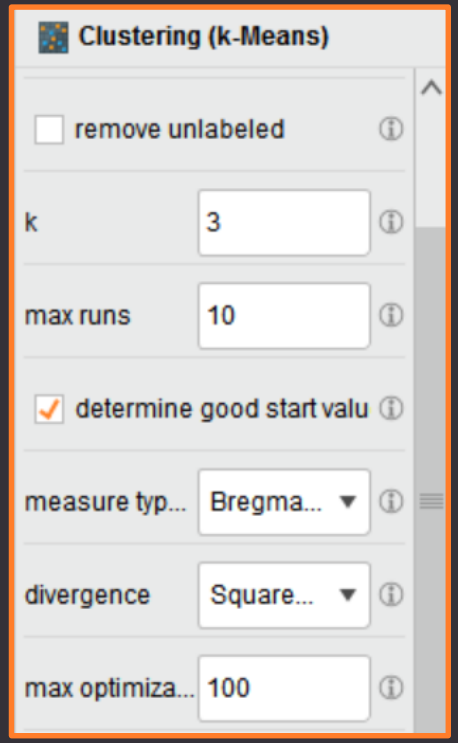
DS 13주차 수업 팀과제

군집화

팀 AOA

경영학과 2014109050 박한솔
경영학과 2015126062 신지수

01
K-mean operator 주요 파라미터



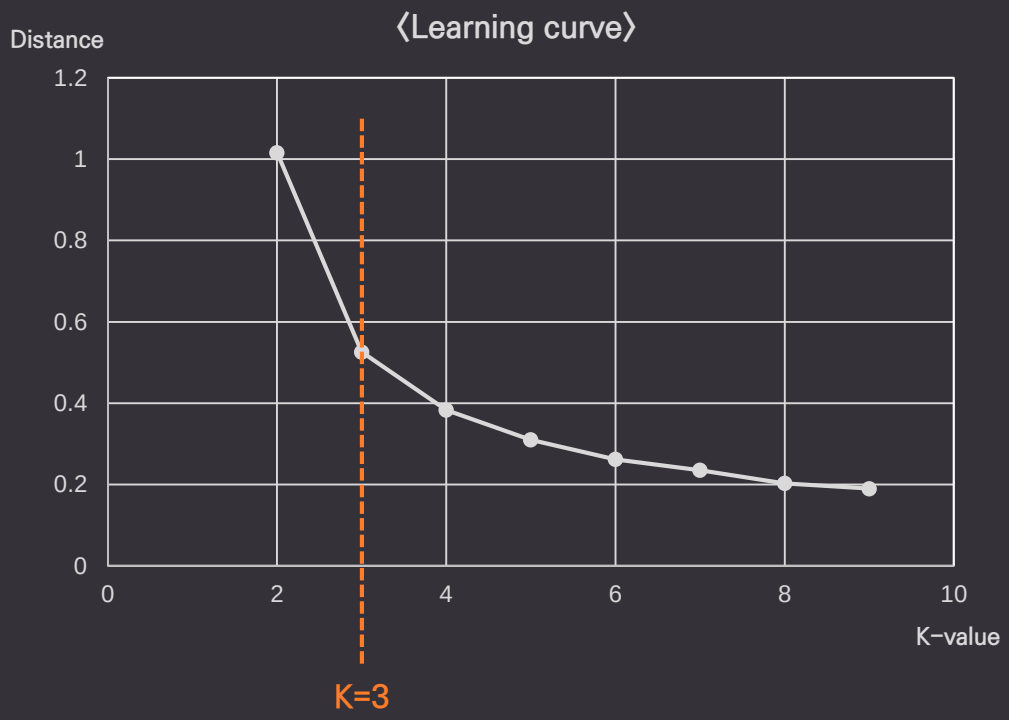
1. k: 클러스터 수를 지정한다.

2. Max runs: 수행되는 시작점을 임의로 초기화 하는 k-means의 최대 실행 수를 지정한다.

3. divergence: 분기 유형을 정의한다.

4. Max optimization steps: k-means의 한 실행에 대해서 수행된 최대 반복 횟수를 지정한다.

02
K-value



적절한 k값을 찾기 위하여 Learning curve 방식 사용

K값이 늘어남에 따라 감소하는 평균 군집 내 거리를 그래프로 표현해 보았다.

여기서 군집이 많아질수록 거리가 감소한다는 것을 알 수 있으며, 거리값의 감소 폭이 가장 큰 K=3구간이 모델의 적절한 K값이라 판단할 수 있다.

03
군집화 평가 척도

AVG. within centroid distance

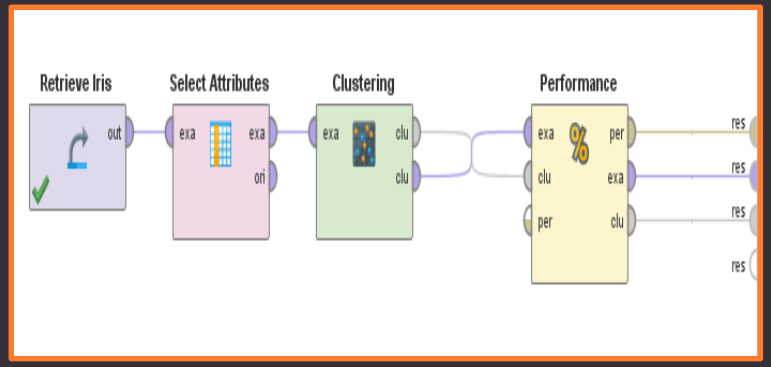
평균 군집 내 거리는 각 군집마다 중심에서 객체마다의 거리를 계산한 후 그 평균값을 사용한 척도이다. 좋은 모델일 수록 각 군집마다 거리가 낮고 모든 군집에 대한 전체 평균 거리 또한 낮게 된다. 평균 군집 내 거리는 **군집 안**에서 객체들이 얼마나 잘 뭉쳐있는지, 즉 **응집력**을 표현한 척도이다.

Davies-Bouldin index

데이비스-볼딘 지수는 군집들의 특성에 대한 척도로서, 군집 안의 응집도와 **군집 밖**의 **분리성**을 고려한 척도이다. 군집의 응집도는 평균 군집 내 거리를 구할 때처럼 군집 안 객체들과 중심과의 거리를 말하며, 군집의 분리성은 군집들 간 거리를 말한다. 데이비스-볼딘 지수는 군집의 응집도와 군집의 분리성의 비율로 표현하며, 데이비스-볼딘 지수가 낮을수록 더 좋은 군집화 결과를 뜻한다.

그러나 평균 군집 내 거리와 데이비스-볼딘 지수 모두 한계가 있으므로, 낮은 값이 나오더라도 좋은 결과를 보장할 수 없는 경우도 있다.

04
Process



〈irs 데이터셋 군집화 process〉

PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: -0.526
Avg. within centroid distance_cluster_0: -0.305
Avg. within centroid distance_cluster_1: -0.652
Avg. within centroid distance_cluster_2: -0.628
Davies Bouldin: -0.666
```

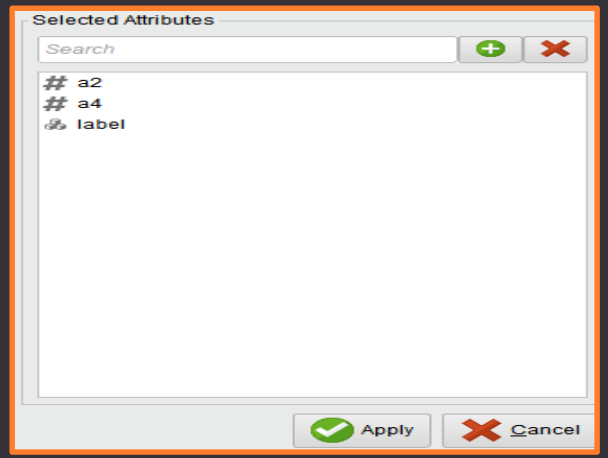
〈iris 데이터셋의 성능측정 결과〉

Iris 데이터셋으로 군집 모델을 생성해보았다.

군집 개수는 앞의 Learning curve에서 확인한 3으로 설정하였다.

성능벡터 창에서 확인한 결과 평균 군집 내 거리는 0.526으로 확인할 수 있다.

05
변수 선택



〈select attribute로 최적 변수 선택〉

PerformanceVector

PerformanceVector:
Avg. within centroid distance: -0.138
Avg. within centroid distance_cluster_0: -0.131
Avg. within centroid distance_cluster_1: -0.144
Avg. within centroid distance_cluster_2: -0.140
Davies Bouldin: -0.704

〈최적 변수 모델의 성능 측정 결과 〉

최적 변수 선택은 우선 각 변수 하나씩 선택하여 각각의 평균 군집 내 거리를 확인하였다.

그 결과 거리가 상대적으로 작은 변수인 a2, a4변수를 선택하고, 모델을 생성하였다.

성능 벡터에서 앞의 모델에 비교하여 평균 군집 내 거리가 감소한 것을 확인할 수 있다.

06
결과 평가

label	cluster
Iris-setosa	cluster_0
Iris-setosa	cluster_0
Iris-versicolor	cluster_2
Iris-virginica	cluster_1
Iris-virginica	cluster_1
Iris-virginica	cluster_1
Iris-virginica	cluster_1
Iris-virginica	cluster_1
Iris-virginica	cluster_1
Iris-virginica	cluster_2
Iris-virginica	cluster_1
Iris-virginica	cluster_2
Iris-setosa	cluster_0

앞의 최적변수 군집화 모델의 결과를 실제 Iris 데이터셋의 label과 비교하여 결과를 평가했다.

Result 탭의 example set에서 실제 결과를 비교할 수 있었으며 군집은 label 수와 동일하게

3개씩 존재한다.

Setosa= cluster 0, Virgincia= cluster 1, Versicolor= cluster 2로 주로 분류 되었으며,

간간히 오분류된 경우도 확인할 수 있었다.

결과, 150개의 데이터 중 11개의 경우만이 군집과 label이 일치하지 않았다.

반면 최적변수가 아닌 모든 변수의 모델은 150개 중 16개의 경우가 일치하지 않았다.

그러므로 최적변수 군집화 모델의 성능이 더 높다는 것을 확인할 수 있다.

07
고객 세분화를 위한 데이터 마이닝 방법 제안

[세분화 방법]

DBSCAN 방법 적용시 저밀도 영역에 속하게 되는 고객은 누락처리가 된다. 따라서 k-mean 방법으로 고객을 세분화한다.
먼저 고객의 가입 정보를 토대로 고객 세분화를 진행하고 저장한다. 이후, 구매 목록 데이터가 있는 고객에 한하여 새롭게 세분화를 진행한다.

[속성선택방법]

고객이 회원가입 시 입력하는 기본 정보(나이, 성별 등) 보다는 사는 지역과 신용도를 통해 지불 능력을 더 정확하게 평가할 수 있을 것이다. 또한 고객의 구매가 이루어지면, 구매 목록에 대한 데이터가 남으므로 이 데이터들을 토대로 하여 관심 분야, 자주 구입하는 목록을 새로운 하나의 속성으로 만들어낼 수 있다.

[세분화의 적절성 평가]

고객이 특정 페이지에 머물러 있던 시간이 길수록, 해당 페이지를 클릭한 빈도수가 많을수록 고객의 흥미가 크다는 것을 의미하며 이는 구매로 이어질 확률이 높다. 따라서 이러한 클릭 횟수, 머무른 시간 등을 통해 세분화의 적절성을 판단할 수 있다.

08
신입고객 처리 방법 제안

신입 고객의 경우 소비목록에서 데이터를 불러올 수 없다.

따라서 기본 정보만으로 세분화했던 데이터에서 비슷한 기본 정보를 가지는 그룹에 포함시켜 그 그룹의 특성을 토대로 신입 고객의 관심을 끌 만한 것들을 예측하여 보여줄 수 있다.

신입고객의 구매데이터가 충분히 쌓이기 전까지는 데이터가 추가시마다 처음에 설정한 그룹 안에서 같은 그룹 내 다른 구매 고객의 데이터에 따라 상품을 추천한다.

구매 데이터가 충분히 축적된 이후에는 구매목록까지 포함하여 세분화한 데이터에 추가한 후 구매 성향에 맞게 세분화하여 상품을 추천한다.