[머신러닝 기반 데이터 분석] 04. 자율학습 모델 적용하기

학습 목표
가. 자율학습 머신러닝 알고리즘 기법을 이해해 본다.
나. K-means 의 동작 원리에 대해 알아본다.
다. K-means 의 K를 설정시의 주의사항에 대해 알아본다.

04. 자율학습 모델 적용하기
4-1 클러스터링(군집)[Clustering] 분석

(1) 자율학습 (Unsupervised Learning, 비지도학습)은 무엇인가?

- 데이터 세트에 목적변수(혹은 반응변수)(Y)가 없이, 주어진 **X1, X2, X3..만 주어진 경우의 머신러닝 기법**
- 무엇을 **예측한다기보다는** 주어진 데이터에서 **특정한 패턴**이나 알려지지 않은 **지식을 발견하고자 하는 것이 목표**
- 예측 대상을 '지도' 할 수 없다. 그래서 머신러닝 결과가 만족스러운지 점검이 곤란하다.

(2) 그렇다면 비지도학습(자율학습)은 어떤 머신러닝 기법이 있을까?

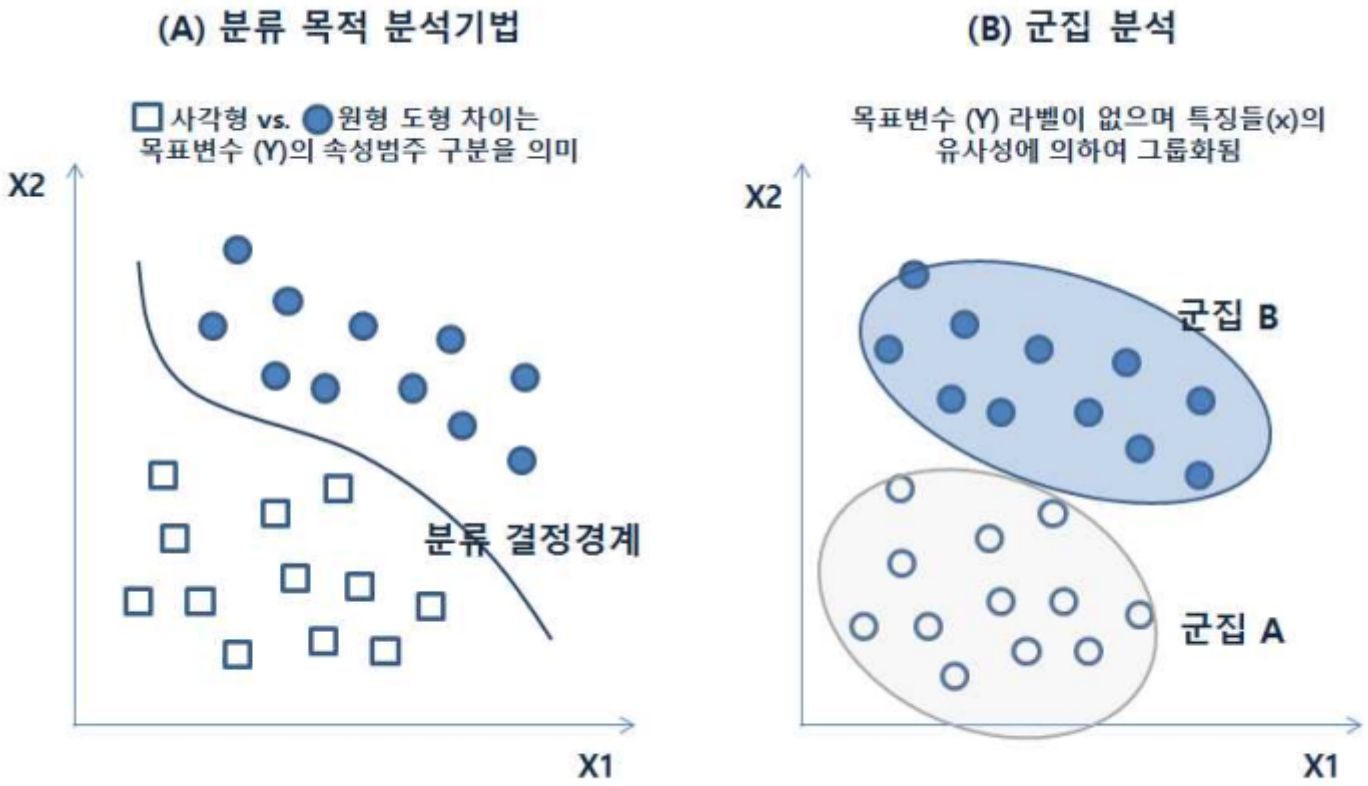
- Clustering(군집화) : 유사 개체나 사람들을 **그룹**짓는다.
- 연관성 분석(Association) : 대상들 간의 발생 **관련성을 파악**한다.
- 차원 축소(Dimension Reduction) : 주어진 변수 세트를 효과적으로 **설명이 가능한 더 적은 수의 대표 변수로 요약**

(3) 클러스터링(군집) 분석

왜? 분석가가 찾고 있는 것이 무엇인지 모른다. 분석가가 컴퓨터 프로그램에게 무엇을 지도할 수 없다.
이런 의미로 군집 분석이 자율학습(비지도 학습)으로 불리고 있다.

예측하기 위해 실행하기 보다 **지식 발견 그 자체를 위한 목적으**로 주로 활용된다.

클러스터링(군집) 분석의 원리
어떤 형태로 그룹을 형성하는가가 분석의 핵심 목적.
-> 일반적으로 **각 데이터 간의 유사성을 기준**으로 그룹화를 짓게 된다.



[그림 4-1] 분류목적의 머신러닝 기법과 클러스터링(군집) 분석의 개념 비교

분류와 군집의 차이점 :
(A) 분류(Classification) : 목적 변수(Y)의 라벨이 주어진다.
(B) 군집(Clustering) : 목적 변수(Y)의 라벨이 없다.

(4) 클러스터링(군집)은 주로 어떤 분야에 활용되는가?

(1) 마케팅 등 분야에서의 **고객 세분화(Segmentation)**
(2) **질병 및 환자** 특성에 따른 유사 **그룹화**
(3) 개체 유사성에 근거한 **문서 분류**
(4) 디지털 이미지 인식 통한 **사물 및 안면 인식**
(5) 금융 분야에서의 알려진 군집 이외의 사용 패턴 식별
(**신용카드 사기**, **보험료 과다 청구** 등)
(6) **공학 분야에서의 이상치 탐색**
(제조 과정에서의 불량 제품 자동 탐지, 통화 음질 개선을 위한 노이즈 구별 등)
(7) 컴퓨터 네트워크에 **비인가 된 침입 등의 비정상적 행위 탐지**

(5) 클러스터링(군집) 분석의 주요 종류

(가) 계층적 군집(Hierarchical Clustering)

- **병합적 군집화(혹은 상향식 군집화)**
- **분할적 군집화(혹은 하향식 군집화)**

--> 실제 빅데이터 환경에서 컴퓨팅(계산처리)에 상당히 많은 자원이 소요되는 경향이 있어, 잘 사용되지 않는다.

(나) 비 계층적 군집(Non-Hierarchical Clustering)

- **K-평균(K-Means)**
- **K-medoids**
- **DBSCAN**
- **퍼지 군집**

--> K-means의 여러가지 한계나 문제점을 극복하기 위해 다른 기법들이 개발되었다.
--> 비지니스 실무 환경에서 많이 사용되는 분석 기법

(다) 분할 기반의 군집(Partition-based Clustering)

참고 내용 (NCS 모듈 교재 참조)

<표 4-1> 대표적인 클러스터링(군집) 분석 주요 기법

구분	기 법	주요 내용
비계층 군집 (분할 기반 군집)	K-평균(K-Means) 클러스터링	주어진 군집 수 k 에 대해서 군집 내 거리 제곱 합의 합을 최소화하는 형태로 데이터 내의 개체들을 서로 다른 군집으로 그룹화하는 기법
	K-Medoids 클러스터링 혹은 (PAM : Partitioning Around Method)	K-평균 클러스터링의 보완한 기법으로서, 모든 형태의 유사성(비유사성) 측도를 사용하며, 좌표평면상 임의의 점이 아닌 실제 데이터 세트 내의 값을 사용하여 클러스터 중심을 정하므로 노이즈나 이상치 처리에 강건한 군집화 기법
	DBSCAN (Density Based Spatial Clustering of Application with Noise)	K-평균 기법이 K개의 평균과 각 데이터 점들 간의 거리를 계산하여 그룹화를 하는 반면, DBSCAN은 밀도개념을 도입하여 일정한 밀도로 연결된 데이터집합은 동일한 그룹으로 판정하여 노이즈 및 이상치 식별에 강한 군집화 기법

	자기 조직화 지도 (Self Organizing Map)	자율학습 목적의 머신러닝에 속하는 인공 신경망의 한 기법으로서 벡터 수량화 네트워크를 이용한 군집화 기법
	Fuzzy 군집	K-평균 기법이 하나의 데이터 개체는 하나의 군집에만 배타적으로 속하는 독점적 군집인데 반해, 퍼지군집은 하나의 데이터 개체가 여러 개의 군집에 중복해서 속할 수 있도록 하는 중복 군집화 기법
계층적 군집	병합적(Agglomerative) 혹은 상향식(Bottom-up) 군집화	모든 데이터 객체를 별개의 그룹으로 구성한 뒤, 단 하나의 그룹화가 될 때까지 각 그룹을 단계적으로 합쳐가는 계층적 군집기법
	분할식(Divisive) 혹은 하향식(Top-down) 군집화	모든 데이터 객체를 하나의 그룹으로 구성한 뒤, 각 데이터 점이 하나의 그룹으로 될 때까지 단계적으로 분할에 가는 계층적 군집기법
확률 기반 군집	가우스 혼합 모형	EM (Expectation Maximization) 알고리즘, 혹은 MCMC (Markov Chain Monte Carlo) 등의 알고리즘을 사용하여 모수를 추정하는 확률 기반의 군집분석

(6) K-Means Clustering(K 평균 클러스터링)

- K-평균 클러스터링은 주어진 군집 수 k에 대해, 군집 내 거리 제곱 합의 합을 최소화하는 것을 목적으로 한다.
- 계산량이 다른 군집 분석에 비해 적은편이다.
- 계산량이 적어 빅데이터 환경에서 실행속도가 빠른 편이다.
- 실무에서 가장 많이 활용되는 군집분석 기법
- 다른 많은 비 계층적 군집 분석 기법들이 K-평균 군집의 응용 또는 변형이다.

(7) K-Means Clustering(K 평균 클러스터링) 에서의 데이터 간 거리 측정

- 비 계층적 군집에서는 유사도 측정을 위해 아래와 같은 여러가지 거리 측정 방법을 사용한다.

A. 유클리디안 거리

$$Euclidean\ D = d(x,y) = (\sum_{i=1}^{p}(x_i - y_i)^2)^{1/2} \qquad (4.1)$$

Euclidean D = 각 데이터 점들간의 x, y 거리의 합

B. 그외 다양한 거리 측정 방법

$$Minkowski\ D = d(x,y) = (\sum_{i=1}^{p}(x_i - y_i)^m)^{1/m}$$

Minkowski Distance(민코브스키 거리) : 유클리디안 거리의 p차원 일반화 형태

$$Manhattan\ D = d(x,y) = \sum_{i=1}^{p}|x_i - y_i|$$

Manhatten Distance(맨해튼 거리) : 각 데이터점 간의 차이들의 절대값의 합을 이용.

$$Standardized\ D = d(x,y) = (\sum_{i=1}^{p}\frac{(x_i - y_i)^2}{s^{2_i}})^{1/2}$$

s : 표준편차, X, Y 각각의 점

Standardized Distance : 유클리디안 거리를 데이터의 분산을 사용하여 표준화한다.

$$Mahalanobis\ D = d(x,y) = (X- Y)^T\Sigma^{-1}(X- Y)$$

Mahalanobis Distance(마할라노비스 거리) : 표준화 거리를 분산-공분산 행렬로 일반화.

$$Chebychev\ D = d(x,y) = \max_{i=1,\ldots,p}|x_i - y_i|$$

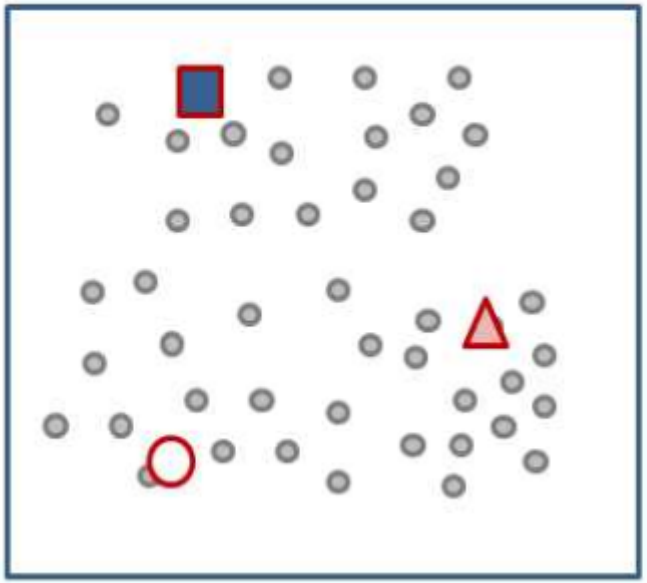
Chebychev Distance(체비세프 거리) : 데이터점 간의 차이의 절대값 중 최대값을 이용.

범주형 자료의 경우, 자료가 얼마나 불일치 하는 가의 비율 계산
Jaccard Distance(자카드 거리)

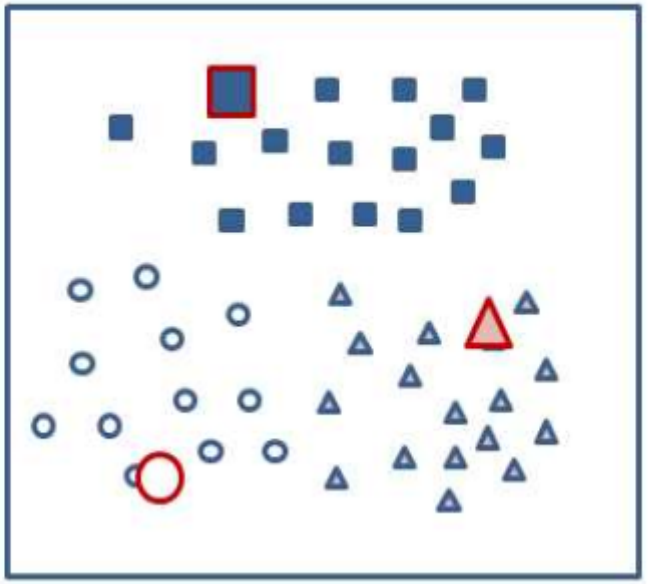
4-2 클러스터링(군집) - K-mean 분석 기법의 동작원리

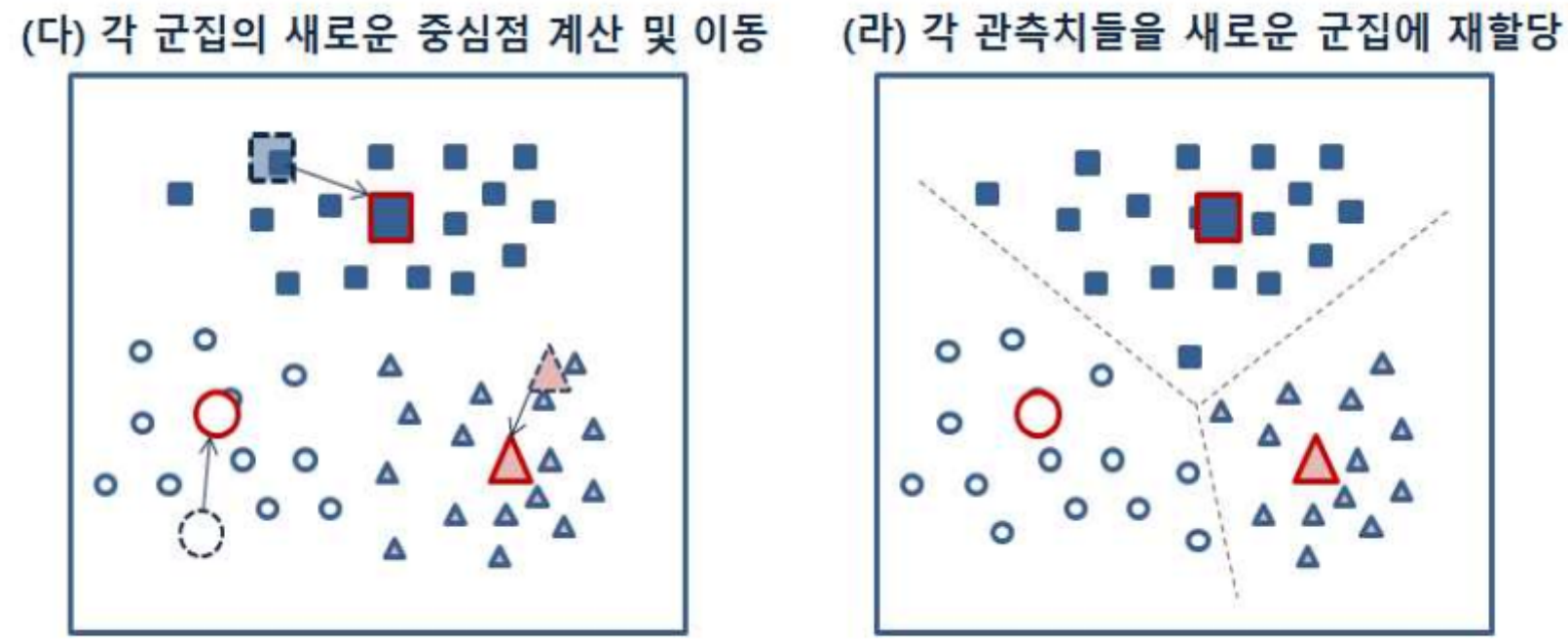
(1) 어떤 원리로 동작할까?

(가) 임의의 초기 군집 중심 설정



(나) 중심점과의 거리 계산 및 군집 할당





[그림 4-2] K-평균 클러스터링 분석의 절차 예시

(2) K-평균 클러스터링의 기법의 장단점은 무엇일까?

<표 4-3> K-평균 클러스터링 기법의 장단점

장 점	단 점
• 개념에 대한 이해가 쉽고 직관적이다. • 사전 모형 설정 및 모수 추정이 필요 없다. • 개체들 간의 거리측정과 군집 수 및 초기 중심점만 주어지면 바로 분석을 적용할 수 있다. • 빅데이터 상황에서 다른 군집분석 기법보다 계산시간이 빠르다. • 기법의 역사가 길어서 다양한 프로그래밍 언어에서 사용될 수 있는 많은 구현물이 있다.	• 무작위 초기점(중심점) 할당으로 인해 최적의 군집을 찾지 못할 수도 있다 • 군집 수 k에 대한 분석가의 임의적 판단이 필요함 • 데이터 점들 간의 중복(겹침)을 허용하지 않는다. • 데이터의 성격상 계층적 구조로 되어 있는 경우에는 사용하기 어렵다. • 연속형 변수의 거리 측도만 다룬다. • 노이즈나 이상치로 인해 군집분석 결과가 영향을 많이 받는다.

(3) K-평균 클러스터링은 K개수를 어떻게 정할까?

A. 군집 개수 **k=1부터 임의의 k가지를 지정**한 뒤, 군집 내 **동질성 및 이질성을 측정**한다.
B. **군집수를 늘려가면서** 동질성의 증가와 이질성의 감소 기울기의 절감 지점인 **엘보우(elbow)값**을 찾는 방법.
C. 여러 가지 군집 개수 K를 적용해 보고, 해당 분야의 비즈니스적 이해와 경험을 활용하여 가장 **결과 해석이 용이한 K를 선택**한다.

--> 여기서 K-평균 클러스터링 수행 시 직면하게 되는 이슈는~
K-평균 클러스터링은 초기 중심점을 임의로 선택한다. 그러한 초기 중심점 선택에 결과가 많이 영향을 받는다.

[해결]
중심점을 바꿔가며 반복적으로 **K-평균 클러스터링을 실행**하고 그 결과 중 **거리 제곱 합이 가장 작은 결과**를 선택하는 방법을 사용한다.

4-3 클러스터링(군집) - K-mean 분석 기법 실습

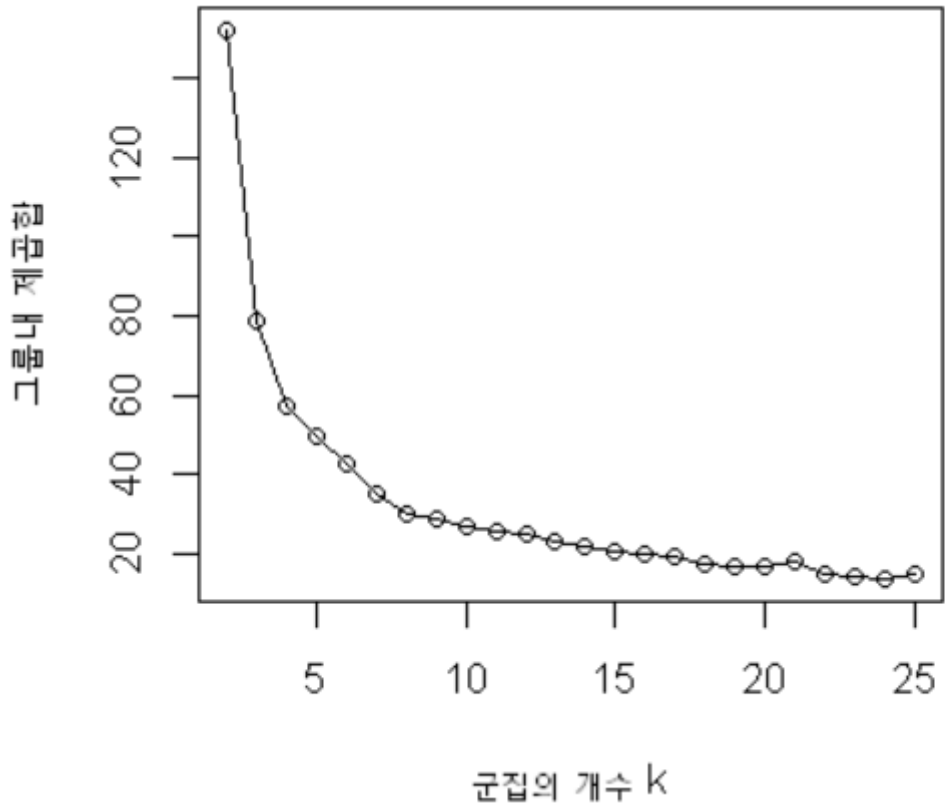
소스 코드	설명
iris2<-iris[,1:4]	
	#목표변수 (Species) 제외
k <- c() km.out.withness<-c() km.out.between<-c() for (i in 2:25){ set.seed(1)	
km.out<-kmeans(iris2, centers=i)	# K-means 기법 적용
k[i-1] <- i	#군집 내 제곱합 저장
km.out.withness[i-1]<-km.out$tot.withinss	#군집 간 제곱합 저장
km.out.between[i-1]<-km.out$betweenss } data.frame(k, km.out.withness, km.out.between)	

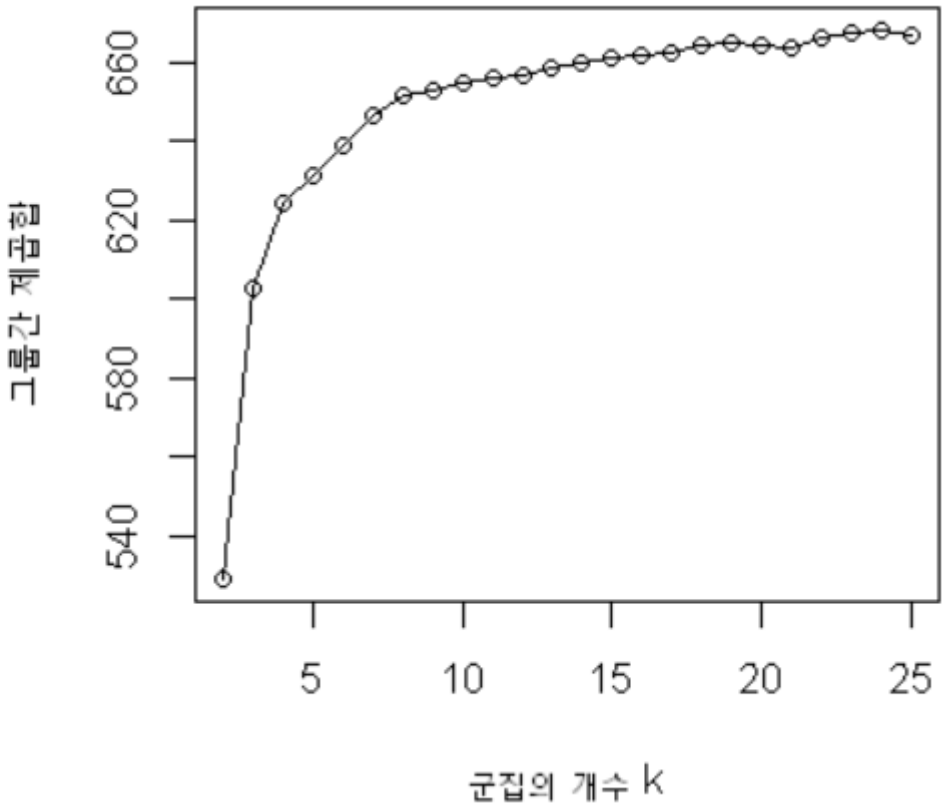
k	km.out.withness	km.out.between
2	152.34795	529.0226
3	78.85144	602.5192
4	57.22847	624.1421
5	49.82228	631.5483
6	42.45606	638.9145
7	34.75675	646.6139
8	29.98894	651.3817
9	28.71578	652.6548

...

14	21.79146	659.5791
15	20.49144	660.8792
16	19.44148	661.9291
17	18.97767	662.3929
18	17.16533	664.2053
19	16.54186	664.8287
20	16.86666	664.5039
21	17.60204	663.7686

(설명)
군집내 제곱합(km.out.withness)과 군집 간 제곱합(km.out.between)을
군집 개수 k=2부터 k=25까지 변화시켜 가면서 비교해 봄.





(1) 군집 개수 k가 증가함에 따라 **군집 내 제곱합은 감소**하고, **군집 간 제곱합은 증가**함.
(2) k=21이 되면서 군집내 제곱합이 증가하고, 군집간 제곱합이 감소. 이는 군집 개수가 지나치게 많이 설정됨.
(3) 군집내 제곱합의 경우 감소 폭의 기울기가 급감하는 엘보우 점이 K=3일 때이다.
군집 간 제곱합의 경우도 증가 폭의 기울기가 급감하는 엘보우 점이 K=3이다.
즉 K-평균 클러스터링에서는 군집 수 K=3이 적당하다.