정형데이터 마이닝 관련 ADsP 특강

데이터의 분류, 머신러닝의 분류

데이터 특성에 따른 분류

범주형 categorical

- 고유한 값이나 범주 수가 제한된 데이터 (문자, 숫자 등으로 표현)
- 순서가 없는 명목형, 순서가 존재하는 순위형이 있음

명목형 nominal

- 단순히 대상의 특성을 분류하는 것, 숫자로 바꾸어도 값의 크고 작음 없음
- 성별(F, M), 혈액형(A, B, O, AB), 부서(기획, 마케팅,...) 등

순위형 ordinal

- 항목들 간에 서열이나 순위가 존재하는 데이터 (양적 비교 불가)
- 학점(A, B, C, D, F), 메달(금, 은, 동), 만족도(1, 2, 3, 4, 5) 등

수치형 numeric

- 숫자를 사용하여 표현된 데이터
- 정수로 표현되는 이산형과 실수로 표현되는 연속형이 있음

이산형 discrete

- 셀 수 있고 특정 값과 값 사이에 다른 값이 존재하지 못하는 정수 데이터
- 주사위 눈의 수, 자녀 수, 사고 횟수, 제품의 개수 등

연속형 continuous

- 데이터 값과 값 사이에 무수히 많은 다른 값들이 존재 하는 실수 데이터
- 키, 몸무게, 기온 등 (등간, 비율 척도 포함)

기계학습(Machine Learning)의 분류

지도학습 Supervised

- X를 사용해 Y를 예측할 때, 학습 데이터에 X, Y 데이터가 모두 존재하는 학습
- X를 독립변수, Y를 종속변수라고 하며, Y에는 실제 값, 예측 값이 존재함
- 회귀(Regression), 분류(Classification) 모델이 있음

회귀 Regression

- 예측 값이 실제 값보다 크거나 작거나 사이 값일 수 있음 (연속형 결과)
- 부모 키를 사용해 딸의 키 예측, 판매량 예측, 집값 예측

분류 Classification

- 예측 값이 실제 값에서 주어진 데이터 범주(종류)로 제한됨 (범주형 결과)
- 화물의 정시 도착 여부 예측, 생존 여부 예측, 품종 예측, 이미지 숫자 예측

비지도학습 Unsupervised

- 학습 데이터에 X에 대한 데이터만 존재하는 학습
- 군집(Clustering), 연관(Association) 모델이 있음

군집 Clustering

- 데이터를 특성에 따라 구분되는 몇 개의 그룹으로 나누는 학습
- 고객을 3개 그룹으로 나눔 (그룹내 서로 유사한 특성, 범주형 결과)

연관 Association

- 항목들 간의 '조건-결과' 식으로 표현되는 유용한 패턴을 발견하는 것
- 삼겹살→상추, 빵→우유 (지지도, 신뢰도, 향상도 등으로 연속형 결과)

기계학습(Machine Learning)에서의 데이터

- 예측 또는 그룹나누기, 패턴 규칙을 발견하는 방법을 수식화 하여 모델 또는 알고리즘이라고 부름
- 회귀, 분류, 군집, 연관분석 각각에 여러 가지 모델들이 존재함
- 모델에 데이터를 넣어 학습(=분석)을 진행하는 것을 모델링이라고 함
- 모델링 과정에서 데이터들은 모델의 종류에 따라 다양한 수학적 연산을 통해 결과를 도출함
- 따라서, Machine Learning에 사용되는 모든 데이터는 수치형(=연산가능한) 이어야 함

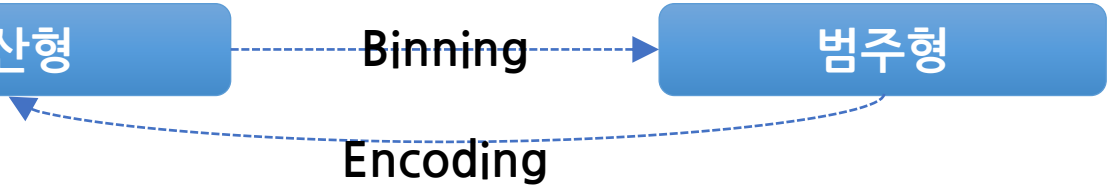
인코딩 Encoding	▪ 사용자가 입력한 문자나 기호들을 컴퓨터가 이용할 수 있는 것으로 바꾸는 것 ▪ 범주형 데이터가 문자/문자열로 표현된 경우, 이산형 수치로 바꿔 사용
구간화 Binning	▪ 데이터의 종류가 많은 경우, 구간으로 나누어 묶어 사용 (학습에 도움이 될 수 있음) ▪ 예) 날짜시간 ➔ 월별, 일별, 요일별, 시간별, 나이 ➔ 연령별 등 ▪ 임의의 구간을 만들 수 있음 ➔ 성수기/비수기

데이터 분류 어떻게 하지?

- 한 가지 데이터를 하나의 형으로 한정하지 않기!

나이는 어떤 데이터로 분류하면 될까요?

나이는 매일 매시간 늘어가고 있음	1, 2, 3, 4... 이렇게 태어난 년도를 1살로 년도가 바뀔 때마다 1씩 증가함	'10대', '20대', '30대' .. 와 같은 연령대
연속형	이산형	범주형



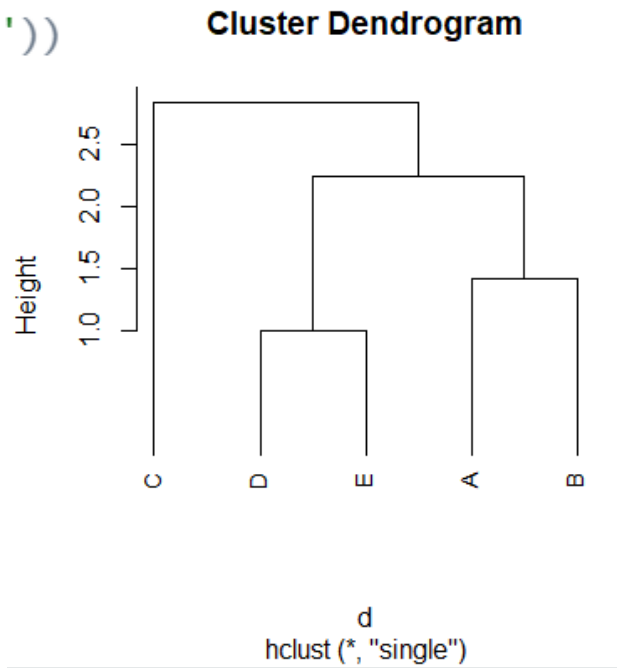
이산형과 범주형을 같은 개념으로 이야기 하는 경우도 있음

계층적 군집의 예

- 아래는 학생들의 키와 몸무게를 정규화 한 데이터이다. 최단연결법을 통해 학생들을 3개의 군집으로 나누면 어떻게 나누어 지는가? (Euclidean 거리 사용)

학생	(키, 몸무게)
A	(1, 5)
B	(2, 4)
C	(4, 6)
D	(4, 3)
E	(5, 3)

```
height <- c(1,2,4,4,5)
weight <- c(5,4,6,3,3)
student <- data.frame(height, weight,
                      row.names=c('A', 'B', 'C', 'D', 'E'))
d <- dist(student)
# single : 최단거리법, complete : 최장거리법
# average : 평균기준법
m <- hclust(d, method='single')
plot(m, hang=-1, cex=0.9)
```



계층적 군집의 예

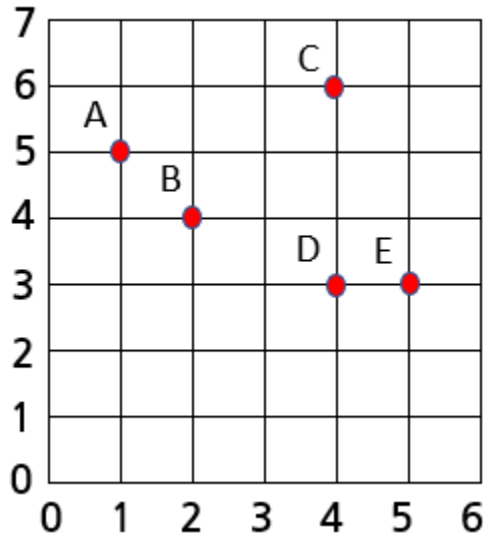
- 아래는 학생들의 키와 몸무게를 정규화 한 데이터이다. **최단연결법**을 통해 학생들을 3개의 군집으로 나누면 어떻게 나누어 지는가? (Euclidean 거리 사용)

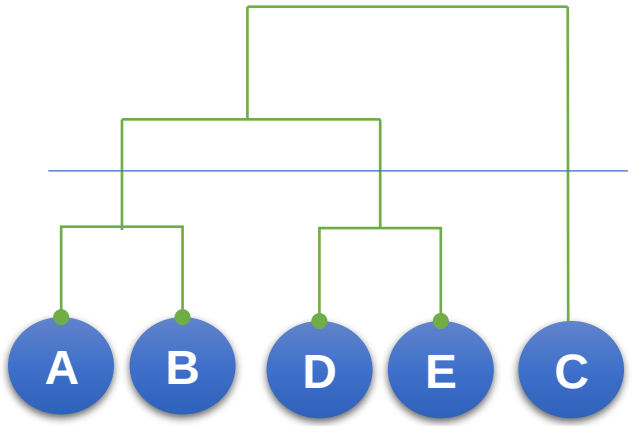
	A	B	C	D
B	2			
C	10	8		
D	13	5	9	
E	18	10	10	1

	A	B	C
B	2		
C	10	8	
DE	13	5	9

	AB	C
C	8	
DE	5	9

	C
ABDE	8





1. 각 학생 사이의 거리를 Euclidean 거리의 제곱으로 표시한 거리표를 작성한다 (표기, 연산의 간략화)
2. 가장 작은 숫자를 찾아 가장 먼저 군집을 형성하는 것을 찾고, 최단거리표를 작성한다 (최단연결법)
3. 그 다음 작은 값들을 찾아가며 계속 군집을 만들고 최단거리표를 다시 작성한다.

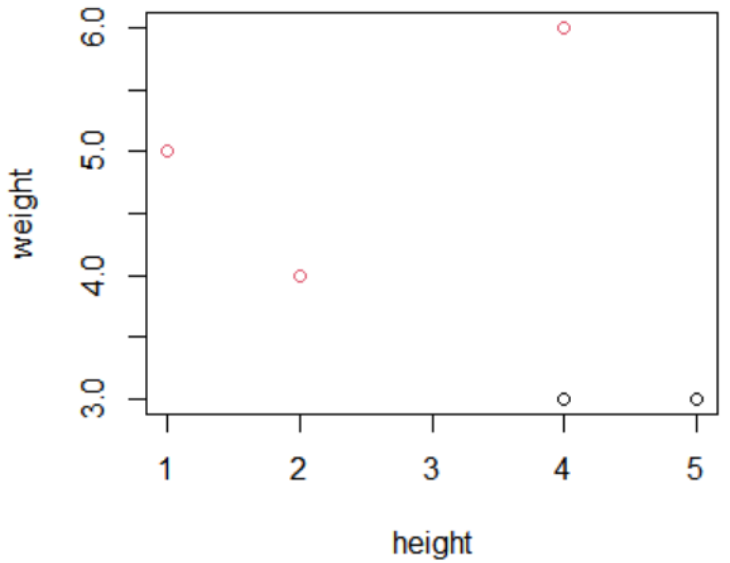
비계층적 군집의 예

- 아래는 학생들의 키와 몸무게를 정규화 한 데이터이다.
  K-means를 사용하여 비계층적 군집을 실행하라

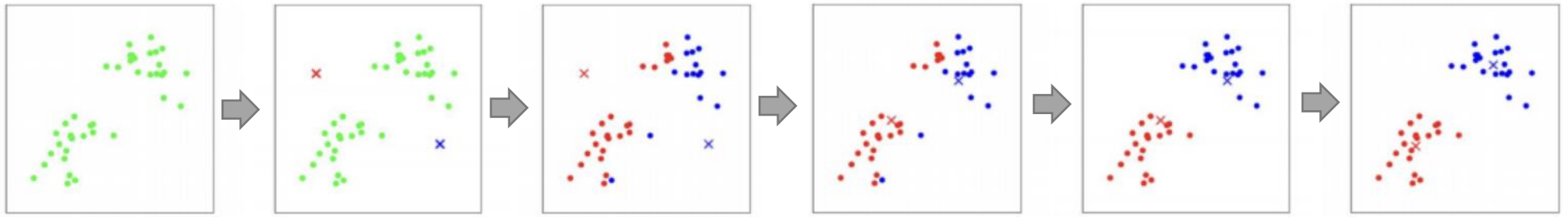
학생	(키, 몸무게)
A	(1, 5)
B	(2, 4)
C	(4, 6)
D	(4, 3)
E	(5, 3)

```
K-means clustering with 3 clusters of sizes 2, 2, 1

Cluster means:
  height weight
1    4.5    3.0
2    1.5    4.5
3    4.0    6.0

Clustering vector:
A B C D E
2 2 3 1 1

Within cluster sum of squares by cluster:
[1] 0.5 1.0 0.0
 (between_SS / total_SS =  91.5 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

```
# 비계층적 군집
km = kmeans(student, 3)
km
plot(student, col=km$cluster)
```

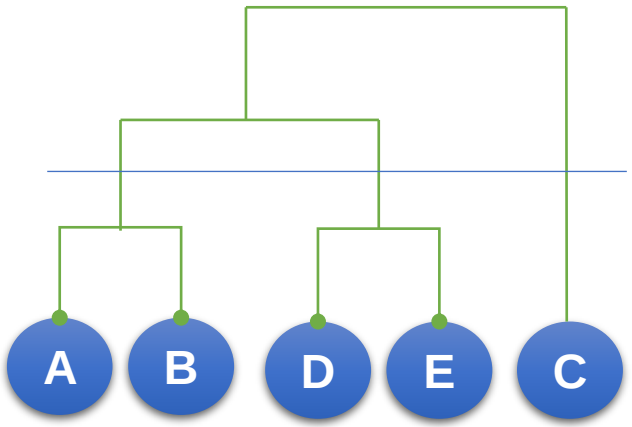


계층적 vs 비계층적 군집

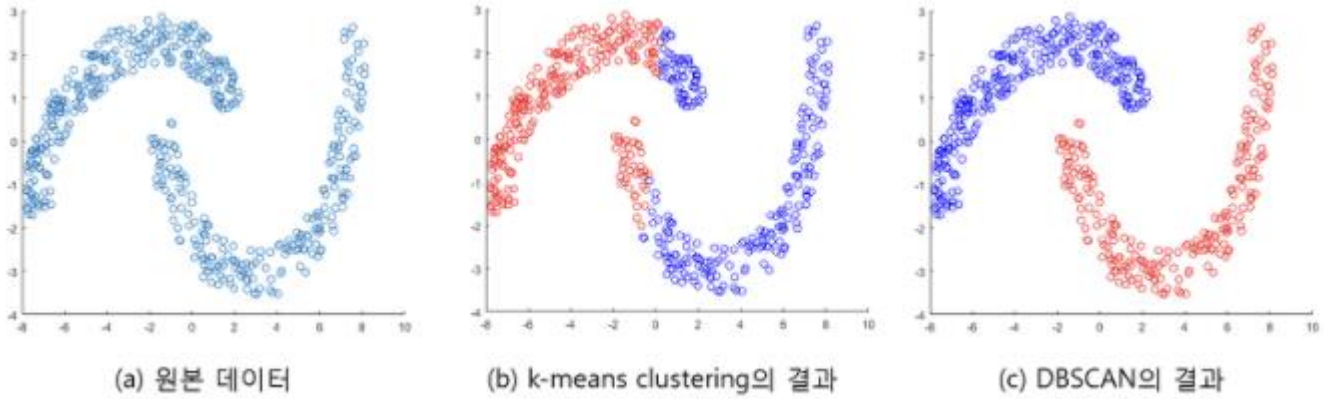


출처 : http://stanford.edu/~cpiech/cs221/img/kmeansViz.png

비계층적 군집



계층적 군집



이미지출처 :
https://untitledtblog.tistory.com/146

결과 해석

2020년, 2021년 기출 관련

의사결정나무 - 불순도 측정

- 지니 지수, 엔트로피 지수 :
  - '자체가 불순도를 의미'함
  - 따라서 작은 값을 선택해야 함, 불순도가 작다는 것이 순수도가 높아지며
  - '순수도가 높은 것을 분류를 잘한 것'으로 봄
  - 엔트로피 지수는 p=0.5 (확률이 0.5)일 때 가장 불순도가 높은 것임

- 카이제곱통계량의 p-value :
  - 카이제곱통계량을 구할 때 '동질성 검사'를 사용함, 즉 동질성 검사의 결과임
  - 동질성 검사의 가설
    귀무 : A의 종류에 따라 B가 같다
    대립 : A의 종류에 따라 B가 다르다 - '대립가설이 채택' 되어야 함
    (대립가설이 채택되면 : 어떤 독립변수를 사용해서, 종속변수의 구분이 잘 된다는 뜻)
  - Chi-Square 값이 클 수록 p-value가 작음의 의미
  - p값이 작을수록 자식노드 내의 이질성이 큼을 나타냄 ('카이제곱 이용시, 이질성이 커야 좋음')

의사결정나무 - 불순도 측정

의사결정나무에서 [A], [B], [C] 노드가 있고 [A노드 Good:50, Bad:50], [B노드 Good:10, Bad:40], [C노드 Good:20, Bad:30] 일 때, B노드의 지니 지수를 구하시오.

- B노드의 지니지수를 구하라고 했으니 [B노드 Good:10, Bad:40] 으로 구하시면 됩니다.
- 지니지수 = $1 - \sum *”C_{\text{ㄴ}} = 1 - ((1/5)^2 + (4/5)^2) = 1 - (1/25 + 16/25) = 1-17/25 = 8/25 = 0.32$

- Good 과 Bad의 합이 50 이고, Good이 10, Bad 가 40 이어서
  Good은 10/50 = 1/5 로 보고 Bad 는 40/50 = 4/5 로 한 것입니다.

정밀도(Precision)

오분류표에서 실제/예측 True와 실제/예측 False가 100으로 동일하다고 한다.
민감도가 0.8이라고 할 때, 정밀도(Precision)은 얼마인가?

민감도(Sensitivity) = 실제 True인 것 중에 예측도 True로 한 것

$= \frac{TP}{TP+FN} = 0.8$ $\frac{10*(TP+FN)*TP}{TP+FN} = \frac{8}{10} * (TP+FN) * 10$, $10TP = 8TP + 8FN$ ➔ TP = 4FN

민감도가 0.8이면 TP=4FN이다.
실제/예측 True와 실제/예측 False가 100으로 통일이라면 아래와 같이 된다. TP+FP=100, FN+TN=100, TP+FN=100, FP+TN=100 이 중 TP+FN=100에서 FN=20, TP=80이다. 그러면 FP=20, TN=80이므로 Precision(예측이 True인 것 중 실제도 True인 것) 을 구하면 TP/(TP+FP)=80/100=0.8이다.

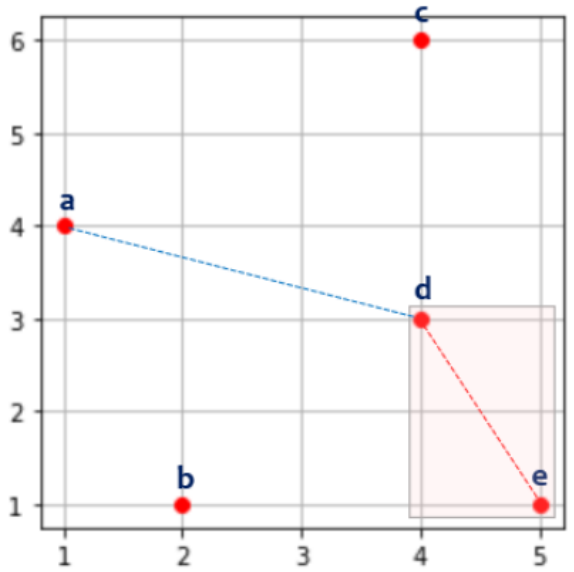
실\예	True	False
True	TP	FN
False	FP	TN

계층적 군집 - 최단연결법

데이터셋 x는 두 개의 변수와 5개의 관측치를 가지며 아래는 데이터와 관측치 간의 유클리드 거리를 나타낸다. 최단연결법을 사용하여 계층적 군집화를 할 때 첫 단계에서 형성되는 군집과 관측치 a와의 거리를 구하시오

```
> x
  x1 x2
a  1  4
b  2  1
c  4  6
d  4  3
e  5  1
```

```
> round(dist(x), 1)
    a   b   c   d
b 3.2
c 3.6 5.4
d 3.2 2.8 3.0
e 5.0 3.0 5.1 2.2
```



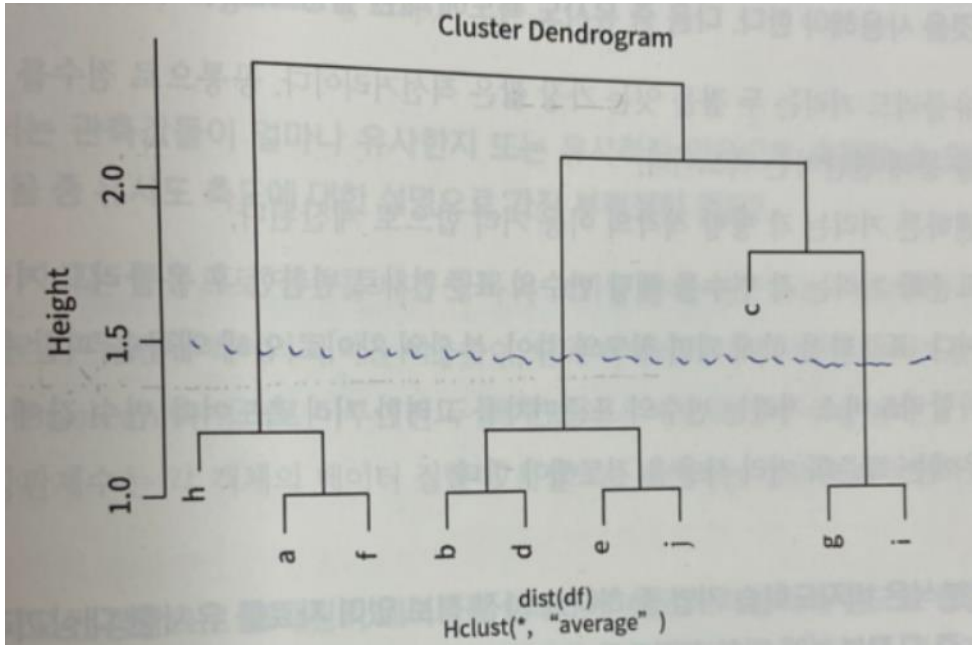
1. 첫 단계에서 형성되는 군집을 dist(x)의 결과에서 찾습니다.
   ➔ 값이 가장 작은 것 {d, e} : 2.0
2. 문제에서 '최단연결법'이라고 했기 때문에 {d, e} 군집과 a의 거리는 d 와 e 중에서 a와 더 가까운 것과의 거리가 됩니다.
   ➔ dist(x)의 결과를 보면 a와 d의 유클리드 거리가 3.2 라고 표기 되어 있습니다.

- 결과 해석- 04 -

답 : 3.2

계층적 군집 - 덴드로그램

아래 그림은 평균연결법을 통한 계층적 군집화 예제이다. 데이터 분석 목적 상 Height값을 1.5를 기준으로 하위 군집을 구성할 때 다음 중 생성된 하위 군집을 가장 잘 나타낸 것은?

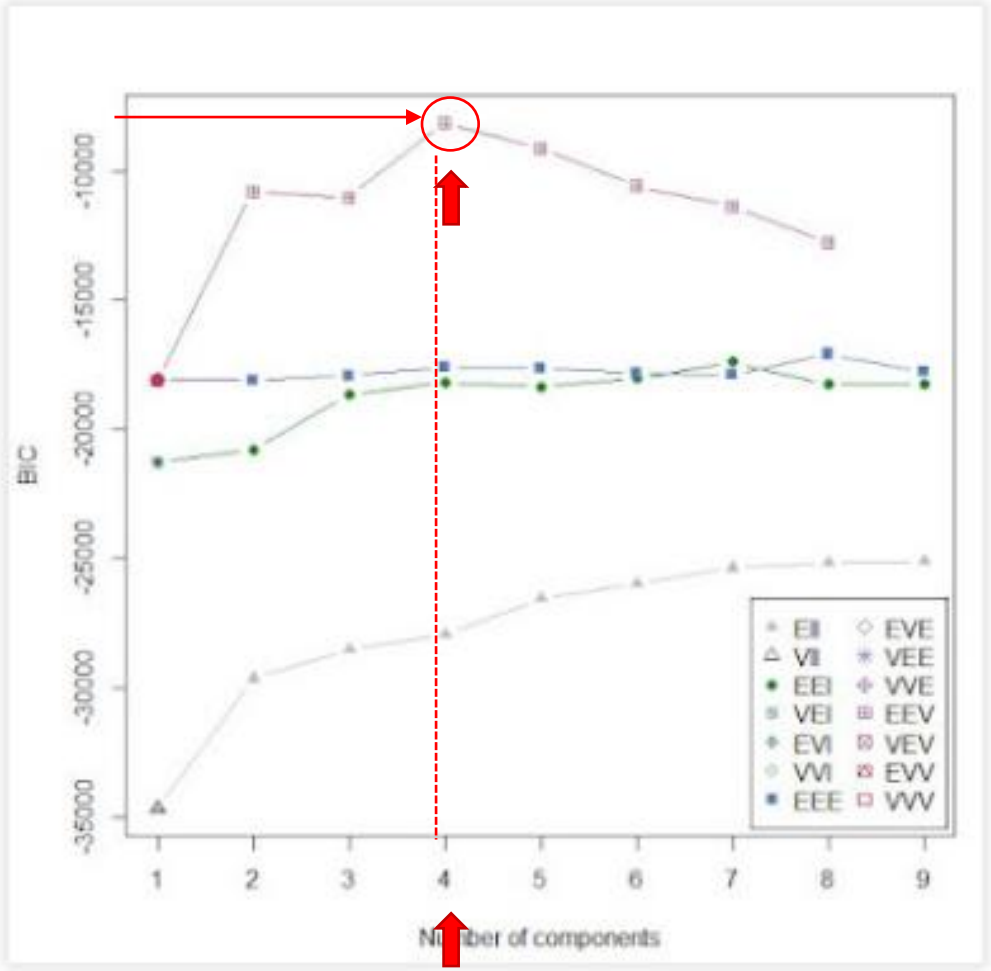


- 덴드로그램에서 c 쪽이 위에 걸려 있지만 아래로 선을 그려서 생각해 주셔야 합니다.
- 즉, Height 1.5의 경우 4개의 군집이 되는 것입니다.
- 그래프가 그려질 때 더 이상 분리 되지 않는 경우 선이 아래까지 그려지지 않아서 그런 것입니다.
- 문제 오류 아님

{c} 가 1.5 선 위에 있지만 그렇다고 군집이 아닌 것은 아니다 (= 군집이다!)

답: {h,a,f}, {b,d,e,j}, {c}, {g,i}

군집분석 결과



최적의 군집수는?

BIC 값이 가장 큰 것의 x축 값을 읽어 주시면 됩니다.
BIC 값이 Y 축에 있으며, 그림에서 위쪽으로 갈 수록 값이 커지는 것을 볼 수 있습니다.

군집, 분류 분석의 구분

분류분석과 군집분석을 어떻게 구분 할 수 있나요??

- 이질적인 모집단을 세분화 ==> 군집, 서로 다른 특성을 갖는 것으로 나누는 것입니다.
- 분류 분석은 '지도 학습'이기 때문에 정답에 맞는 특성을 학습하고 그 구분되는 특성에 따라 class를 나누는 것입니다.
- 군집은 기준이 없고, 분류는 기준이 있습니다.

다음 데이터 마이닝의 대표적인 기능 중 이질적인 모집단을 세분화하는 기능으로 적절한 것은?
➔ 군집분석

에어컨 회사에서 지역별 온도, 습도에 따라 고객군을 나눠서 마케팅 전략을 수립할 때
분석 방법은?
➔ 분류분석