군집분석

문제 1. 다음 중 비계층적 군집화의 장점이 아닌 것은?

① 분석 방법의 적용이 용이하다.

② 결과 해석이 용이하다.

③ 다양한 형태의 데이터에 적용이 가능하다.

④ 데이터에 대한 내부 정보 없이 의미있는 자료구조를 찾을 수 있다.

다음의 데이터는 a ~ e 점들의 x 및 y좌표값들에 대한 것이다. 다음 데이터를 보고 문제 2~3에 답하시오.

In [10]:
```
data = data.frame(x = c(1, 2, 4, 4, 5), y = c(4, 1, 6, 3, 1))
rownames(data)  = c("a", "b", "c", "d", "e")
data
```

	x	y
a	1	4
b	2	1
c	4	6
d	4	3
e	5	1

문제 2. a 점과 e 점의 거리(유클리드 거리)를 계산하면 얼마인가?

① 1

② 3

③ 5

④ 7

문제 3. b 점과 e 점의 거리(유클리드 거리)를 계산하면 얼마인가?

① 1

② 2

③ 3

④ 4

문제 4. 다음 중 관측값들간 거리를 구하는 방법을 의미하지 않는 것은 무엇인가?

① 유클리드

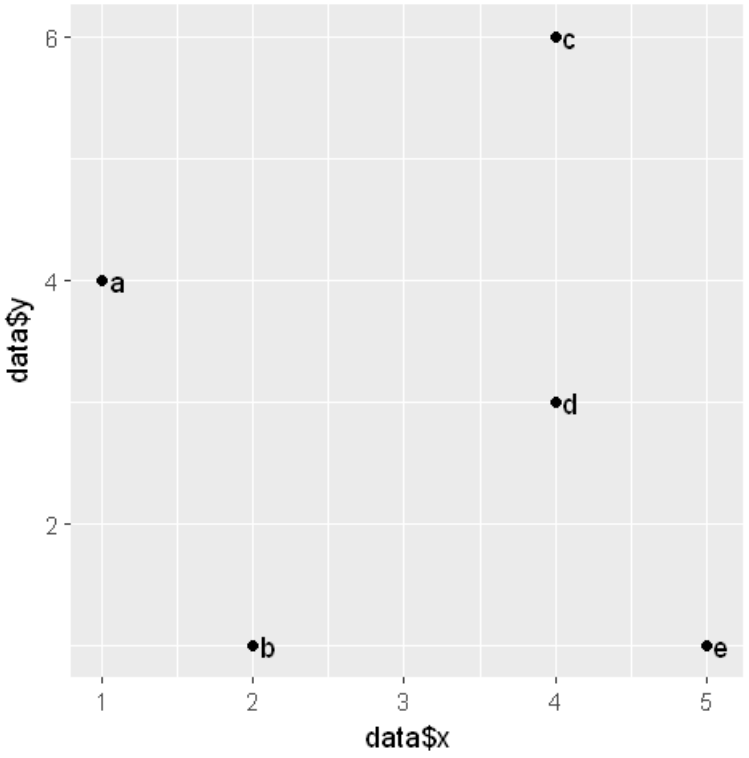
② 와드

③ 맨하탄

④ 마할라노비스

다음과 같이 분포되어 있는 데이터를 보고 문제 5~12번에 답하시오.

In [11]:
```
data = data.frame(x = c(1, 2, 4, 4, 5), y = c(4, 1, 6, 3, 1))
rownames(data)  = c("a", "b", "c", "d", "e")
library(ggplot2)
ggplot(data, aes(data$x,data$y)) + geom_point() + geom_text(aes(label  =  rownames(data)), hjust=0, nudge_x =
library(repr)
options(repr.plot.width=4, repr.plot.height=5)
```



문제 5. b와 d 점간의 거리를 계산하시오.(제곱근 형태로 표현)

문제 6. c와 d 점간의 거리를 계산하시오.(제곱근 형태로 표현)

문제 7. d와 e 점간의 거리를 계산하시오.(제곱근 형태로 표현)

문제 8. b와 e 점간의 거리를 계산하시오.

문제 9. c와 e 점간의 거리를 계산하시오.(제곱근 형태로 표현)

문제 10. 다음 중 데이터를 최단연결법을 이용해 계층적 군집화를 하려 할 때 가장 먼저 생긴 군집을 구성하는 관측값들은 무엇인가?

① a, b

② b, c

③ c, d

④ d, e

문제 11. 다음 중 데이터를 최단 연결법을 이용해 계층적 군집화를 하려 할 때 두 번째 생긴 군집을 구성하는 관측값들은 무엇인가?

① a, c, d

② b, d, e

③ c, e, a

④ a, b, c

문제 12.다음 중 데이터를 최장 연결법을 이용해 계층적 군집화를 하려 할 때 두 번째 생긴 군집을 구성하는 관측값들은 무엇인가?

① a, c, d

② b, d, e

③ c, e, a

④ a, b, c

문제 13. 다음 중 비계층적 군집화의 단점으로 볼 수 없는 것은?

① 가중치와 거리정의가 어렵다.

② 초기 군집수를 결정하기 어렵다.

③ 다양한 형태의 데이터에 적용하기 어렵다.

④ 결과 해석이 어렵다.

문제 14. 계층적 군집화 방법 중 하나로, 군집 내 편차들의 제곱합을 고려하며 군집간 정보의 손실을 최소화 하는 군집연결 방법은 무엇인가?

문제 15. 비계층적 군집화의 대표적인 알고리즘 중 하나로, 데이터를 k개의 클러스터로 묶으며 각 클러스터와의 거리 차이의 분산을 최소화하는 방식으로 동작하는 알고리즘을 무엇이라 하는가?

문제 16. 비계층적 군집화 알고리즘 중 하나로, k-means와는 달리 데이터를 군집화할 때 특정 클러스터에 속할지를 강제(hard assignment)하지 않고, 가우시안 혼합 모형을 이용해 데이터의 소속 여부를 해당 클러스터에서 데이터가 나타날 확률 값으로 표현하여 강제하지 않는(soft assignment) 알고리즘을 무엇이라 하는가?

① PAM 알고리즘

② 혼합군집분포 알고리즘

③ K-means 알고리즘

④ hclust 알고리즘

문제 17. 비계층적 군집화 알고리즘의 하나로, 결측값을 허용하고 이상치에 영향을 덜 받으며 연속형 변수가 아닌 여러 형태의 변수들에게도 적용할 수 있는 일종의 강화된 k-means라 할 수 있으며, 데이터 군집 시 비복원추출된 데이터에서 대표되는 중심점을 기준으로 군집화를 반복하는 알고리즘을 무엇이라 하는가?

문제 18. 다음 중 비계층적 군집화 알고리즘이 아닌 것은 무엇인가?

① PAM 알고리즘

② EM 알고리즘

③ K-means 알고리즘

④ hclust 알고리즘

문제 19. 다음 중 K-means를 수행하는 적절한 절차는?

가. 원하는 군집의 개수가 초기값(seed)을 정해 seed 중심으로 군집을 형성한다.

나. 각 군집의 seed 값을 다시 계산한다.

다. 각 데이터를 거리가 가장 가까운 seed가 있는 군집으로 분류한다.

라. 모든 개체가 군집으로 할당될 때까지 위 과정을 반복한다.

① 가 --> 나 --> 다 --> 라

② 다 --> 가 --> 나 --> 라

③ 가 --> 다 --> 나 --> 라

④ 다 --> 나 --> 가 --> 라

문제 20. 모형기반(Model-based)의 군집방법으로 데이터가 k개의 모수적 모형의 가중합으로 표현되는 모집단 모형으로부터 나왔다는 가정 하에서 모수와 함께 가중치를 자료로부터 추정하는 방법으로 사용하는 군집 방법은?

① k-평균군집(K-means Clustering)

② 혼합 분포 군집(Mixture Distribution Clustering)

③ 계층적 군집(Hierarchical Clustering)

④ 분리 군집(Partitioning Clustering)

문제 21. 비계층적 군집 방법의 기법인 k-means Clustering의 경우 이상값(Outlier)에 민감하여 군집 경계의 설정이 어렵다는 단점이 존재한다. 이러한 단점을 극복하기 위해 등장한 비계층적 군집 방법으로 가장 적절한 것은?

① PAM(Partitioning Around Meroids)

② 혼합 분포 군집(Mixture Distribution Clustering)

③ Density based Clustering

④ Fuzzy Clustering

문제 22. 혼합군집분포(mixture distribution clustering)은 모형 기반의 군집 방법으로서 데이터가 k개의 모수적 모형의 가중합으로 표현되는 모집단 모형으로부터 나왔다는 가정 하에서 분석을 하는 방법이다. k개의 각 모형은 군집을 의미하며 이 혼합모형의 모수와 가중치의 최대가능도(Maximum Likelihood) 추정에 사용되는 알고리즘은 무엇인가?

답안

1. 2
2. 3
3. 3
4. 2
5. $\sqrt{8}$
6. $\sqrt{9}$
7. $\sqrt{5}$
8. 3
9. $\sqrt{26}$
10. 4
11. 2
12. 2
13. 3
14. 와드연결법
15. K-means
16. 2
17. PAM
18. 4
19. 3
20. 2
21. 1
22. EM 알고리즘(기대값 최대화 알고리즘)