고고학 자료 통계분석

Week 8 : 군집분석과 판별분석

숭실대학교 사학과 석사과정 1학기
주 찬 혁

계획

주차	제목	내용
1	Intro	소개, R 설치
2	기초 통계(1)	모집단과 표본, 기술통계량
3	기초 통계(2)	변수의 종류, 가설과 검정, 오류, 분석절차
4	전처리	데이터 전처리
5	시각화	다양한 종류의 그래프
6	검정과 상관분석	검정, 상관분석
7	회귀분석	선형회귀, 다중선형회귀, 로지스틱회귀
8	군집분석과 판별분석	K-means, LDA, QDA
9	주성분분석	PCA

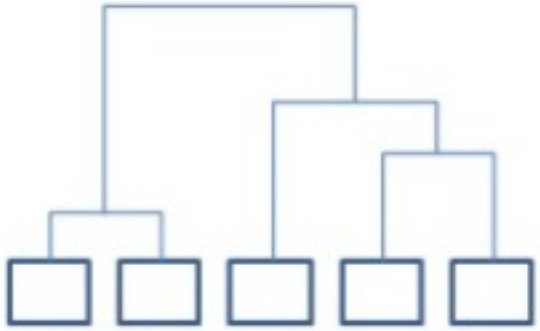
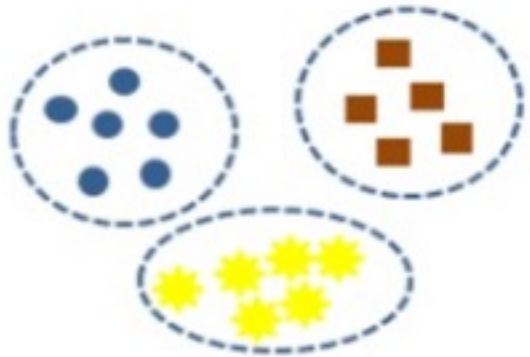
복습

- 회귀분석이 무엇인지 안다.
- 회귀분석의 해석방법, 주의점에 대해 안다.
- R로 단순선형회귀분석을 시행할 수 있다.
- R로 다중선형회귀분석을 시행할 수 있다.
- R로 로지스틱회귀분석을 시행할 수 있다.

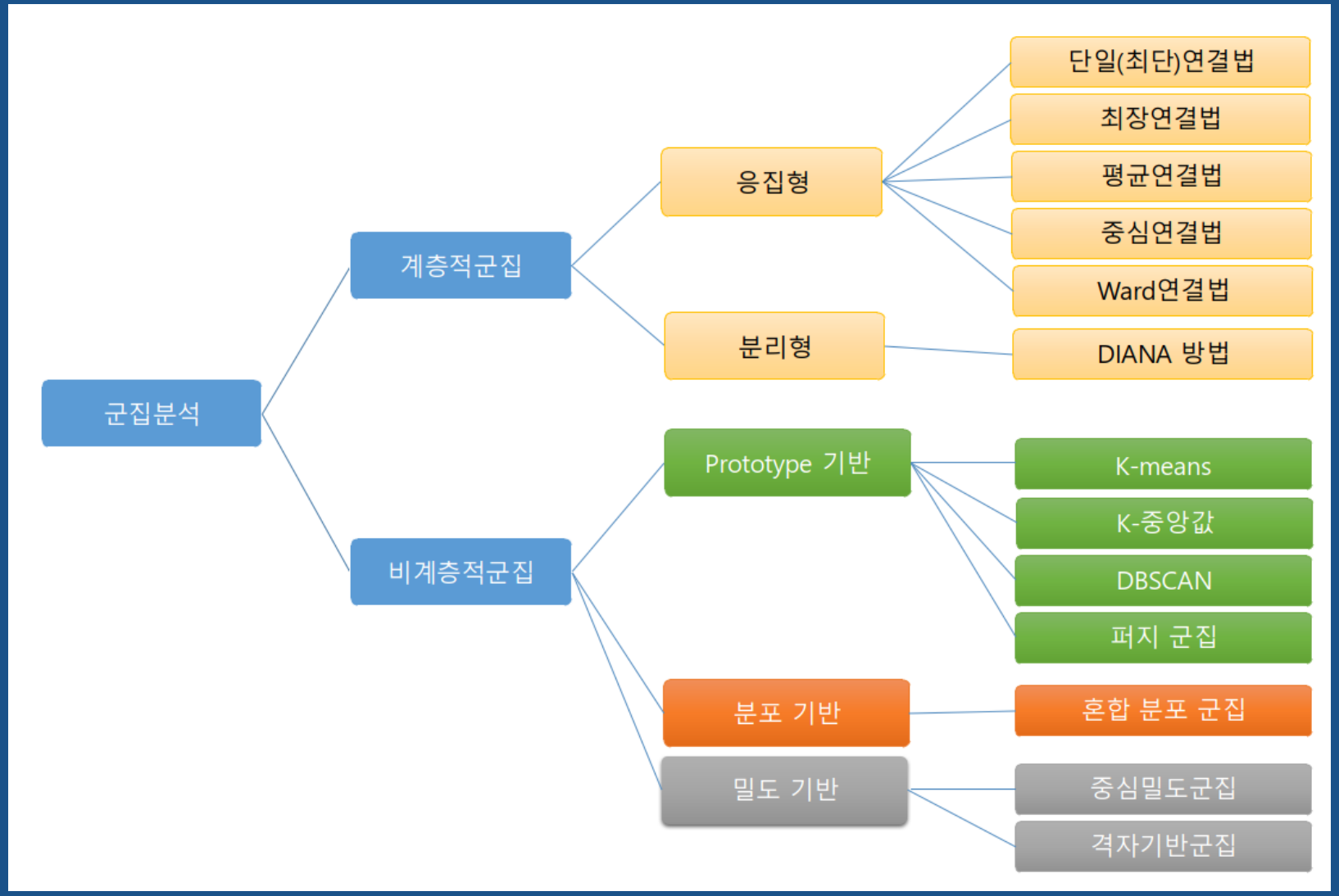
군집분석(Clustering)

- 각 데이터의 유사성을 측정하여 여러 군집으로 구분하고 각 군집 사이의 상이성을 파악하는 분석방법
- 크게 계층적 군집(Hierarchical Clustering)과 분할적 군집(Partional Clustering)으로 구분
- 분할하는 방법과 사용하는 군집 사이의 거리(척도)에 따라 다양한 방법이 존재

군집분석(Clustering)

계층적 군집 (Hierarchical Clustering)	분할적 군집 (Partional Clustering)
	

군집분석(Clustering)의 종류



다양한 거리(=척도)(Distance Function)

- 데이터의 유사도를 측정하기 위해 다양한 거리(척도)가 활용

- 유클리디안 거리 : 유클리드 공간에서 **기하학적 최단 거리**
  - 피타고라스 정리
  - 이상치에 민감
- 맨하탄 거리 : 격자 형태의 공간에서 최단 거리
- 마할라노비스 거리 : 정규분포에서 데이터가 평균으로부터 얼마나 멀리 위치하고 있는지를 나타낸 거리
  - **데이터의 분포 형태를 고려**하여 거리를 측정
  - 이상치 탐지 가능
- 표준화 거리 : 변수의 **측정단위를 표준화**하여 계산한 거리

...

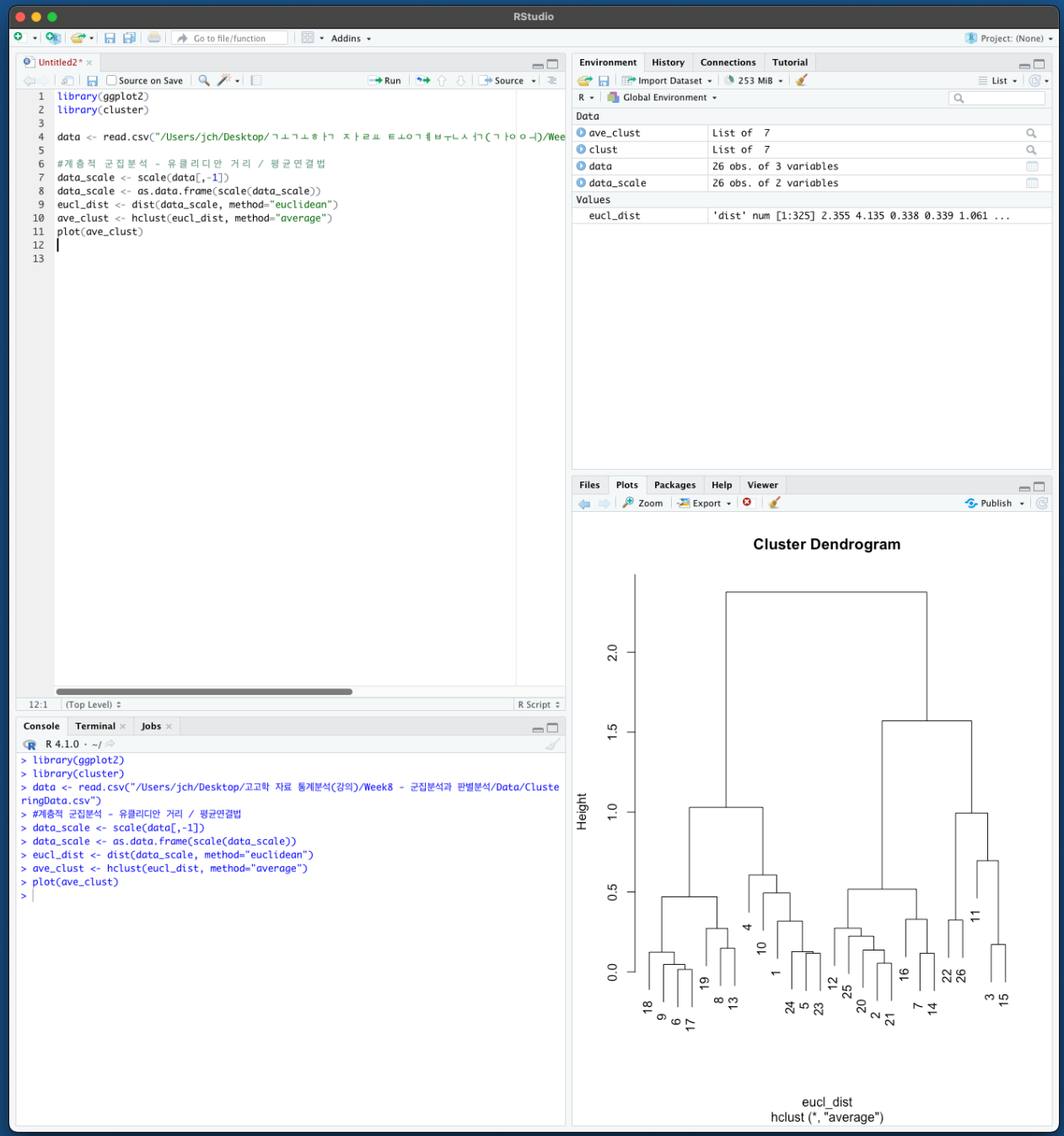
계층적 군집분석

- 다양한 연결방법을 사용하여 데이터 사이의 계층적인 관계를 분석
- 덴드로그램(Dendrogram)을 활용하여 시각화
- 다양한 연결 방법
  - 최단연결법(Single) : 각 군집에서 하나의 관측값을 추출했을 때 나타날 수 있는 최소거리
  - 최장연결법(Complete) : 각 군집에서 하나의 관측값을 추출했을 때 나타날 수 있는 최대거리
  - 중심연결법(Centroid) : 각 군집의 중심간의 거리
  - 평균연결법(Average) : 모든 항목에 대한 거리의 평균을 계산하여 군집화
  - Ward's 연결법(Ward) : 군집 내의 오차제곱합을 통해 군집화

계층적 군집분석

> A <- dist(데이터, method="거리")
> B <- hclust(A, method="연결방법")
> plot(B)

- dist 함수를 사용하여 거리 측정, method에서 사용할 거리 지정
- hclust 함수를 사용하여 군집화, method에서 사용할 연결방법 지정



계층적 군집분석

- 계층적 군집분석의 경우 **데이터의 수가 조금만 많아져도 시각화하는 것이 불가능**
- 데이터의 구조, 형태를 파악하고 **적절한 거리(척도)**와 **연결방법**을 활용하는 것이 필요

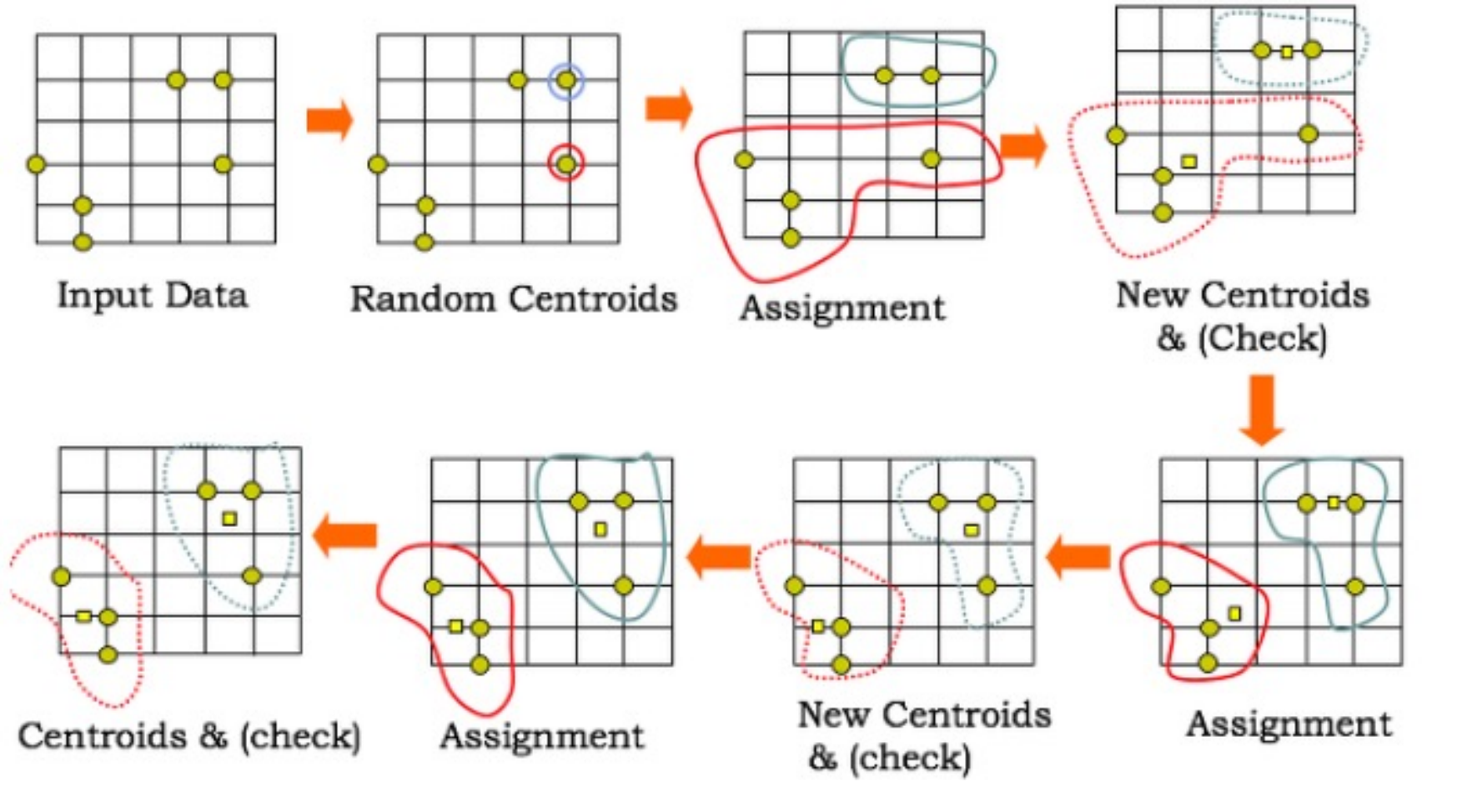
분할적 군집분석

- **다양한 기준**을 사용하여 데이터를 분류하는 방법
- **산점도**를 활용하여 시각화
- 다양한 기준
  - 프로토타입 기반 : k-평균, k-중앙값, k-메도이드, 퍼지
  - 분포 기반 : 혼합분포
  - 밀도 기반 : 중심밀도, 커넬, 격자
  - 그래프 기반 : 코호넨

K-means 군집분석(K-means Clustering)

- 분할적 군집분석 중 하나
- **군집의 개수를 지정**하고 그 개수만큼 군집화하는 방법
- 다양한 기준 중, **평균**에 기반하여 군집화

K-means 군집분석(K-means Clustering)



K-means 군집분석

> 모델명 <- kmeans(데이터, 군집 수)
>
> plot(데이터(c("변수1", "변수2")],
>
> col=모델명$cluster, pch=점 스타일)

- kmeans 함수를 사용하여 k-means 군집화 시행
- 산점도로 시각화 한 뒤, 군집화를 바탕으로 색을 지정하여 구분

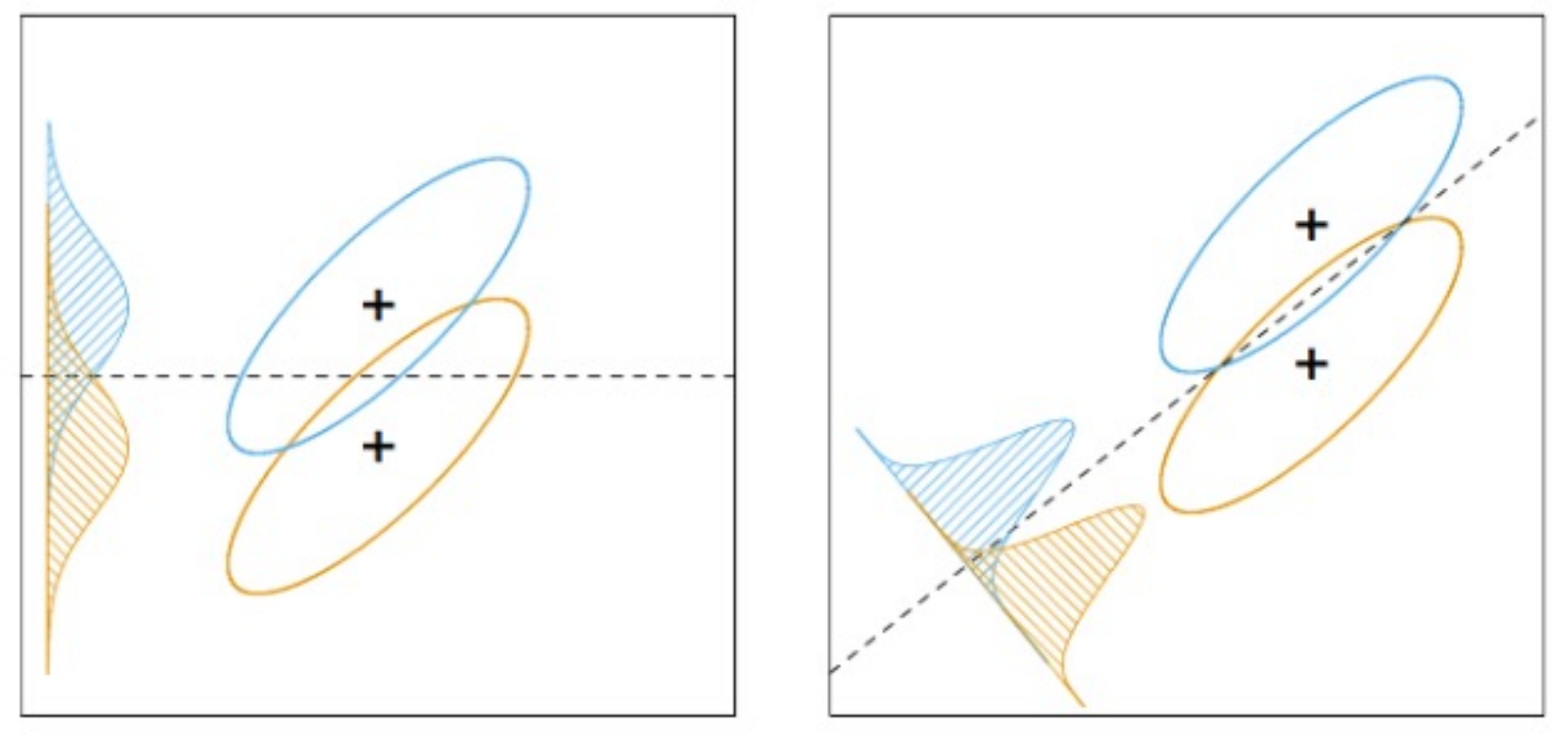
판별분석

- 데이터가 두 개 이상의 군집으로 구분되는 것을 알고 있을 때, 새로운 값을 사전에 알고 있는 정보를 통해 해당 군집들 중 하나로 **분류**하는 방법
- 낮은 **다중공선성**, 변수들의 **정규성** 등이 충족되어야함
- 선형판별분석(LDA), 이차판별분석(QDA)

선형판별분석(Linear Discriminant Analysis)

- 하나의 차원에 사영(Projection)하여 가장 잘 분류할 수 있는 직선을 찾음
- 두 범주의 평균이 서로 멀고, 각 범주의 분산이 작은 것이 좋은 분류 기준
- 공분산행렬이 동일해야함
- 많은 변수(차원)들로 구성된 데이터의 차원을 축소하여 분석하는 차원축소(Dimension Reduction) 방법 중 하나

선형판별분석(Linear Discriminant Analysis)



Iris Dataset

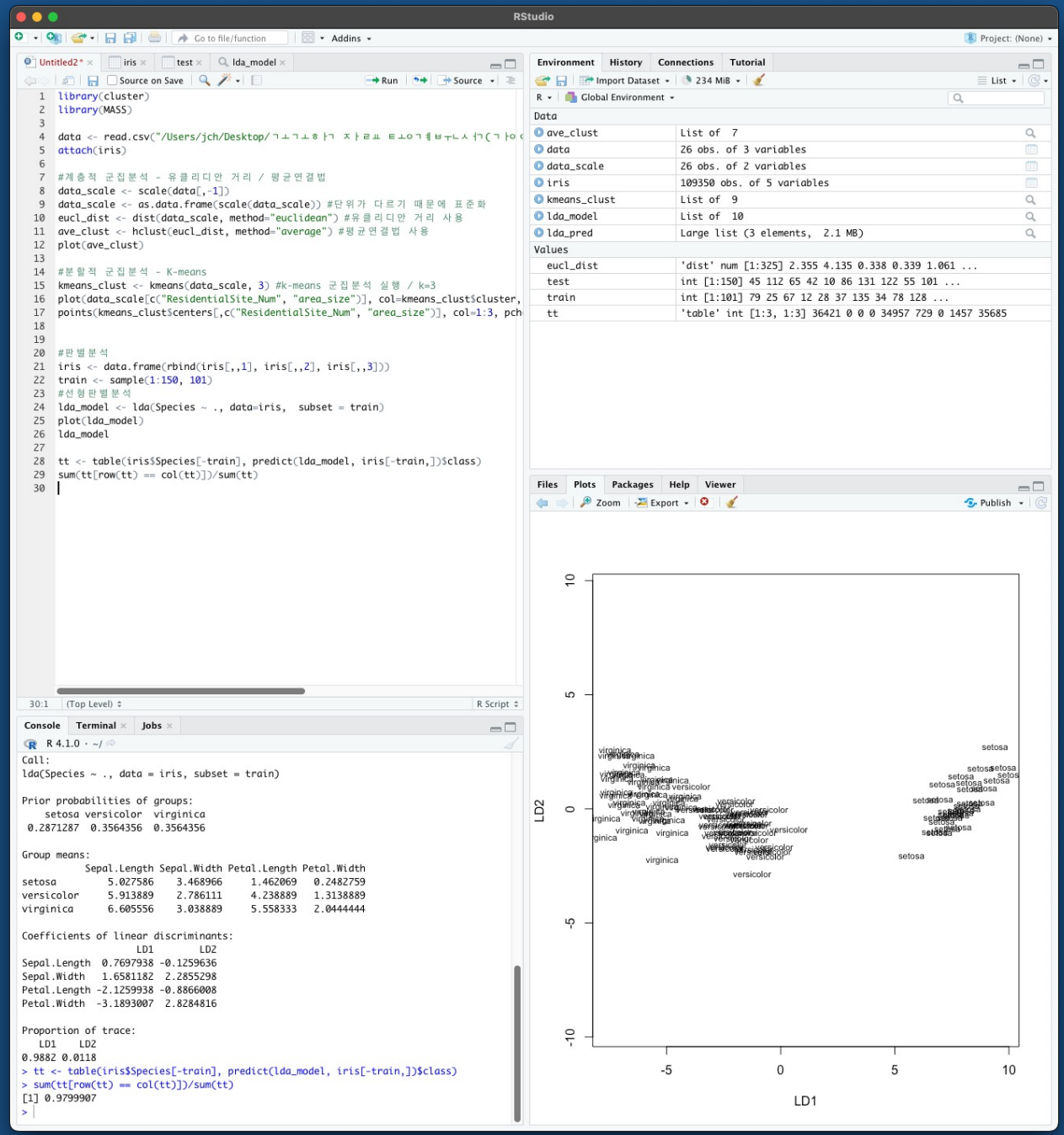
- 붓꽃(iris)에 대한 데이터
- 꽃 잎의 길이, 너비와 같은 계측치를 중심으로 구성된 데이터
- R, Python 등을 활용한 통계, 머신러닝 등에서 가장 기초적으로 사용되는 데이터
- Google에 분석방법을 검색할 경우 대부분 이 데이터셋을 샘플로 사용하여 설명하기 때문에 익숙해질 필요성이 있음

선형판별분석

```
> A <- lda(변수1 ~ ., 데이터)
> plot(A)
```

- lda 함수를 선형판별분석 실행
- 이후 정분류율, 오분류율, ROC 등을 통해 모델 평가

*prior probabilities of groups : 범주별 비율
*Group means : 그룹별 평균값
*Coefficients of linear discriminants : 판별계수



판별분석과 군집분석의 차이

군집분석	판별분석
무언가를 분류하기 위해 시행	
데이터들이 속한 집단을 모를 때, 유사한 속성을 데이터들끼리 군집을 지어 분류하는 분석방법	이미 집단을 알고 있는 데이터를 이용해 모델을 구성하고 이를 활용하여 집단을 모르는 데이터를 분류하는 분석방법