03 빅데이터와 통계를 결합한 경우의 기대효과 측정

1 예측방식 유형

1 예측방식 유형

		관련된 통계적 요소
비지도학습 Unsupervised Learning	군집화/세분화 Clustering	요인분석(Factor Analysis) 판별분석(Discriminant Analysis)
	연관성 분석 Assocation Analysis	지지도(Support) 신뢰도(Confidence) 향상도(Lift)
지도학습 Supervised Learning	회귀분석 Regression Analysis	상관분석(Correlation) 분산분석(ANOVA)
	분류식별 분석 Classification	카이스퀘어 검증

1 예측방식 유형

기준	연속형 값	유형 값
사례	강수량, 매출액, 주가	반응/무반응, 증가/감소
정확도 평가	R Square, RMSE, MAPE 등	Accuracy, Precision, Recall Rate
장점	값의 크기가 중요한 경우 도움	분류한 기준이 의사결정 기준인 경우 유용
단점	정확한 값 예측이 어려움	잘못 분류할 수 있음
활용	제한된 분야에 활용되고 있음	다양한 분야에 활용 가능
제약조건	정규분포, 독립성 등 가정 조건이 있음	가정 조건이 거의 없음
빅데이터 분석 활용	높은 전문성 필요	낮은 전문성으로도 활용 가능

03 빅데이터와 통계를 결합한 경우의 기대효과 측정

2

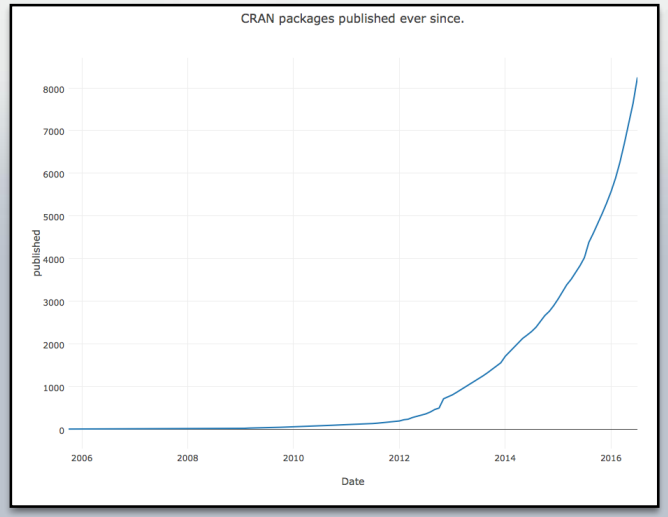
Open Source R을 이용한 분류식별 모델링

1 Open Source R 개요

- 1993년에 개발, 23년의 역사
- 통계 계산과 그래픽을 위한 프로그래밍 언어이자 소프트웨어 환경
- 로버트 젠틀맨(Robert Gentleman), 로스 이하카(Ross Ihaka)에 의해 시작, 현재 R 코어 팀이 개발
- GPL 하에 배포되는 S 프로그래밍 언어의 구현으로 GNU S라고도함(GNU GPL v2)

1 Open Source R 개요

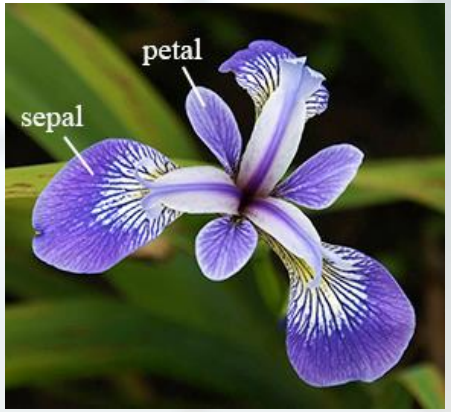
- 통계 소프트웨어 개발과 자료 분석에 널리 사용
- 패키지 개발 용이 → 통계학자들 사이에서 통계 소프트웨어 개발에 많이 사용됨
- Data 가공, Statistical Analysis, Data Mining, Machine Learning, Deep Learning, Visualization, Spatial Analysis, Simulation, Optimization 등



[Open Source R package 증가추이]

3 분류식별에서의 성과측정

1 분류식별 데이터 예




```
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

- 붓꽃 종(Species) : Setosa, Versicolor, Virginica
- Sepal, Petal의 폭과 길이로 구분될 수 있을 것 같아 해당 데이터를 종별로 50건씩 총 150개 데이터 수집

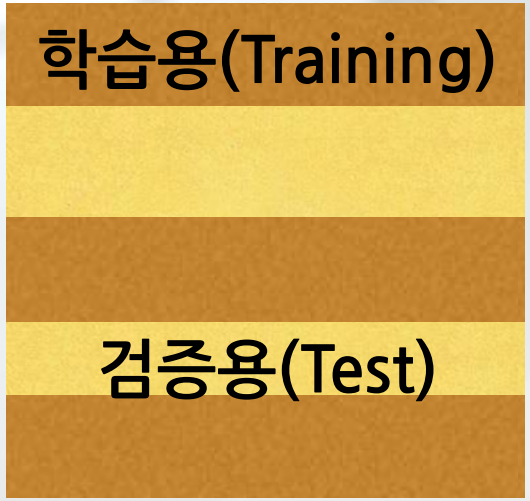
2 분류식별 데이터 기초통계

```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

종별로 동일한 개수가
있어서 임의로 선택하면
제대로 맞출 확률이 33.3%

3 분류식별 검증방법

[전체 데이터]



무작위 추출
(Random Sampling)

	비중	목적
학습용 데이터	70%	패턴을 파악하여 학습하여 예측 성능 높임
검증용 데이터	30%	학습된 패턴이 유사한 성능이 나오는지 검증

3 분류식별 검증방법

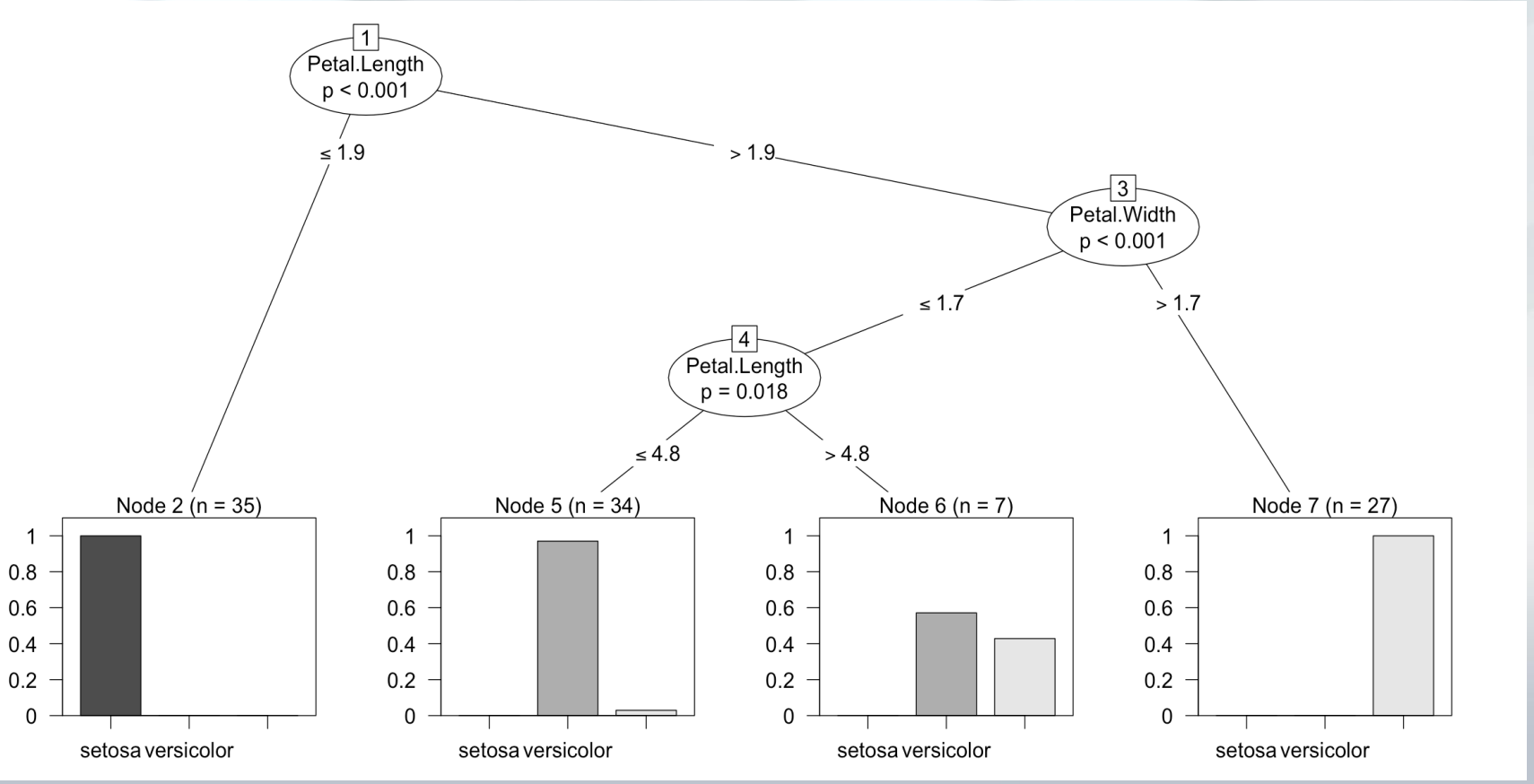
```
# 1 또는 2를 70%, 30%로 생성
ind <- sample(2, nrow(iris), replace=TRUE,prob=c(0.7,0.3))
# 1인 경우만 데이터를 추출해서 tr에 저장하여 학습용으로 활용
tr <- iris[ind==1,]
# 2인 경우만 데이터를 추출해서 ts에 저장하여 검증용으로 활용
ts <- iris[ind==2,]
```

4 분류식별 모형의 예

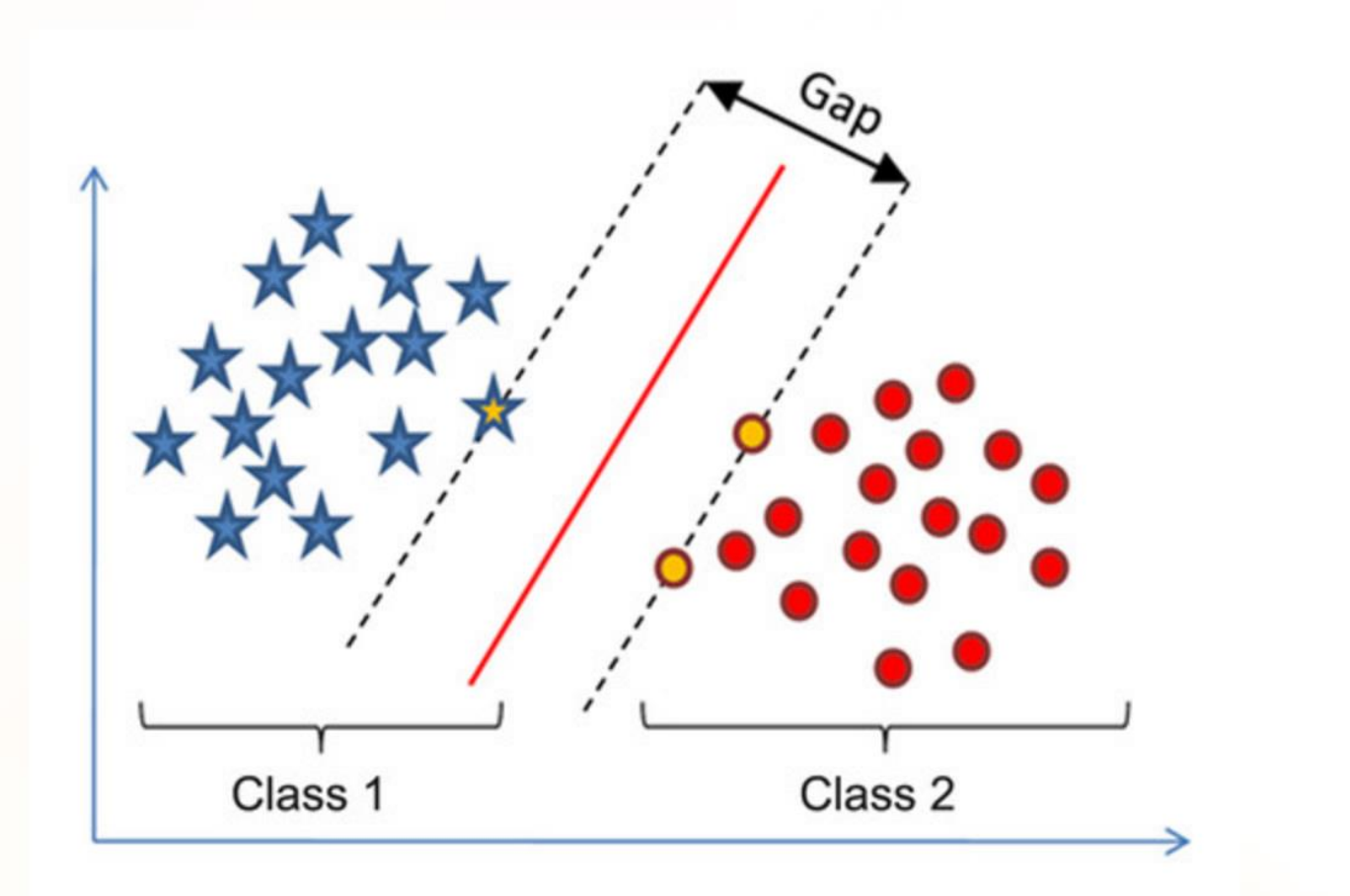
```
> ind <- sample(2,nrow(iris),replace=TRUE,prob=c(0.7,0.3))
> tr <- iris[ind==1,]
> ts <- iris[ind==2,]
> party1 <- ctree(Species~.,data=tr)
> plot(party1)
```

- 학습용 데이터 tr을 이용해서 ctree 함수로 Species가 어떤 변수들로 에측할 수 있는지 자동으로 학습해서 party1에 저장
- 결과를 그래프로 출력

5 분류식별 모형의 결과 및 해석



6 분류식별 모형의 다양한 알고리즘



[Support Vector Machine]

7 분류식별에서의 성과측정

```
> table(predict(party1,newdata=ts),ts$Species)

             setosa versicolor virginica
  setosa         15          0         0
  versicolor      0         12         1
  virginica       0          1        18
```

Accurary = (15 + 12 + 18) / (15 + 0 + 0 + 0 + 12 + 1 + 0 + 1 + 18) = 95.7%

Precision(versicolor) = 12 / (0 + 12 + 1) = 92.3%

Recall Rate(versicolor) = 12 / (12 + 1) = 92.3%

03 빅데이터와 통계를 결합한 경우의 기대효과 측정

4 재무적 효과분석

1 재무적 효과분석 예

```
> table(predict(party1,newdata=ts),ts$Species)

             setosa versicolor virginica
  setosa         15          0         0
  versicolor      0         12         1
  virginica       0          1        18
```

- 정확한 예측 시 수익 1, 잘못 예측 시 -1 비용 발생
- 임의로 150개 분류하라고 하면 33.3%는 정확하고 나머지는 잘못 분류해서 비용 발생 →
  수익 33.3, 비용 66.6으로 -33.33의 손실 발생

1 재무적 효과분석 예

```
> table(predict(party1,newdata=ts),ts$Species)

             setosa versicolor virginica
  setosa         15          0         0
  versicolor      0         12         1
  virginica       0          1        18
```

- 모델 이용 시 Accuracy는 95.7% → 수익 95.7, 비용 4.3으로 91.4의 이득 발생

-33.33과 91.4의 차이는 매우 큼!