

Wk9-4 : 데이터마이닝과 분류

-학습데이터와 검증데이터-

3. iris 데이터 설명

• Iris 데이터 (붓꽃 데이터)

1. 꽃잎의 폭과 길이에 대한 **4개 변수**로 꽃의 종류를 예측하는 것이 목적
2. **타겟변수(y)** : setosa, versicolor, virginica



Iris setosa

Iris versicolor

Iris virginica



데이터마이닝 : 분류(classification)

3. iris 데이터 설명

9.4 학습데이터와 검증데이터

• iris 데이터 (iris.csv)

input변수(독립변수) output변수(종속변수, 타겟변수)

| | A | B | C | D | E |
|----|--------------|-------------|--------------|-------------|---------|
| | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width | Species |
| 1 | 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 2 | 4.9 | 3 | 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 | 3.6 | 1.4 | 0.2 | setosa |
| 6 | 5.4 | 3.9 | 1.7 | 0.4 | setosa |
| 7 | 4.6 | 3.4 | 1.4 | 0.3 | setosa |
| 8 | 5 | 3.4 | 1.5 | 0.2 | setosa |
| 9 | 4.4 | 2.9 | 1.4 | 0.2 | setosa |
| 10 | 4.9 | 3.1 | 1.5 | 0.1 | setosa |
| 11 | 5.4 | 3.7 | 1.5 | 0.2 | setosa |
| 12 | 4.8 | 3.4 | 1.6 | 0.2 | setosa |
| 13 | 4.8 | 3 | 1.4 | 0.1 | setosa |

```
# lec9 4.R
# classification
# training data and test data

# set working directory
setwd("D:/tempstore/moocr/wk9")

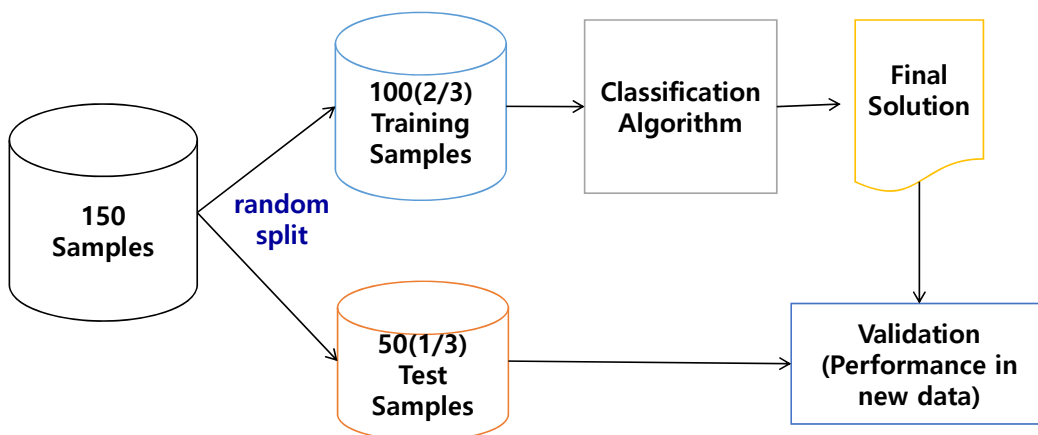
# read csv file
iris<-read.csv(file="iris.csv")
head(iris)
str(iris)
attach(iris)
```

```
> str(iris)
'data.frame': 150 obs. of 5 variables:
 $ Sepal.Length: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9
 $ Sepal.Width : num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3
 $ Petal.Length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4
 $ Petal.Width : num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2
 $ Species : Factor w/ 3 levels "setosa","versicolor"
```

4. 학습데이터와 검증데이터

9.4 학습데이터와 검증데이터

• k- fold cross-validation (k=3, 5, 10)



(Example) 3- fold cross-validation (n=150)

5. 학습데이터와 검증데이터 생성

9.4 학습데이터와 검증데이터

• iris 데이터 (iris.csv) – 150개 데이터

input변수(독립변수) output변수(종속변수, 타겟변수) -> y=iris[,5]

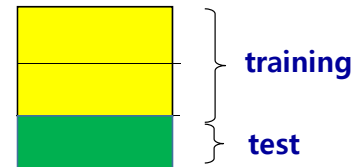
| | A | B | C | D | E |
|----|--------------|-------------|--------------|-------------|---------|
| | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width | Species |
| 1 | 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 2 | 4.9 | 3 | 1.4 | 0.2 | setosa |
| 3 | 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| 4 | 4.6 | 3.1 | 1.5 | 0.2 | setosa |
| 5 | 5 | 3.6 | 1.4 | 0.2 | setosa |
| 6 | 5.4 | 3.9 | 1.7 | 0.4 | setosa |
| 7 | 4.6 | 3.4 | 1.4 | 0.3 | setosa |
| 8 | 5 | 3.4 | 1.5 | 0.2 | setosa |
| 9 | 4.4 | 2.9 | 1.4 | 0.2 | setosa |
| 10 | 4.9 | 3.1 | 1.5 | 0.1 | setosa |
| 11 | 5.4 | 3.7 | 1.5 | 0.2 | setosa |
| 12 | 4.8 | 3.4 | 1.6 | 0.2 | setosa |
| 13 | 4.8 | 3 | 1.4 | 0.1 | setosa |

```
# training/ test data : n=150
set.seed(1000, sample.kind="Rounding")
N=nrow(iris)
tr.idx=sample(1:N, size=N*2/3, replace=FALSE)
tr.idx
```

tr.idx는 100개의 무작위로 선정된 100개의 데이터 아이디

3-fold cross-validation

- set.seed는 난수 생성 시 처음 시작값을 주어 동일한 훈련표본 사용 (set.seed를 지정하지 않으면 매번 다른 훈련표본 생성)
- train/test를 2:1로 랜덤 분할(100/50, n=150)



5. 학습데이터와 검증데이터 생성

9.4 학습데이터와 검증데이터

• iris 데이터를 cross-validation을 위해 분할함

```
# training/ test data : n=150
set.seed(1000, sample.kind="Rounding")
N=nrow(iris)
tr.idx=sample(1:N, size=N*2/3, replace=FALSE)
tr.idx

# attributes in training and test
iris.train<-iris[tr.idx,-5]
iris.test<-iris[-tr.idx,-5]
```



| Environment | History | Connections |
|--------------------|------------------------------------|-------------|
| Global Environment | | |
| Data | | |
| iris.test | 50 obs. of 4 variables | |
| iris.train | 100 obs. of 4 variables | |
| Values | | |
| N | 150L | |
| testLabels | Factor w/ 3 levels "setosa", "v... | |
| tr.idx | int [1:100] 50 114 17 102 76 1... | |
| trainLabels | Factor w/ 3 levels "setosa", "v... | |

iris.train=iris[tr.idx, -5] 5번째 열의 종속변수를 제외한 100개의 데이터

iris.test=iris[-tr.idx, -5] 5번째 열의 종속변수를 제외한 50개의 데이터

5. 학습데이터와 검증데이터 생성

9.4 학습데이터와 검증데이터

| | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width |
|----|--------------|-------------|--------------|-------------|
| 2 | 4.9 | 3.0 | 1.4 | 0.2 |
| 8 | 5.0 | 3.4 | 1.5 | 0.2 |
| 14 | 4.3 | 3.0 | 1.1 | 0.1 |
| 16 | 5.7 | 4.4 | 1.5 | 0.4 |
| 19 | 5.7 | 3.8 | 1.7 | 0.3 |
| 20 | 5.1 | 3.8 | 1.5 | 0.3 |
| 23 | 4.6 | 3.6 | 1.0 | 0.2 |
| 24 | 5.1 | 3.3 | 1.7 | 0.5 |
| 25 | 4.8 | 3.4 | 1.9 | 0.2 |
| 29 | 5.2 | 3.4 | 1.4 | 0.2 |
| 30 | 4.7 | 3.2 | 1.6 | 0.2 |
| 32 | 5.4 | 3.4 | 1.5 | 0.4 |
| 34 | 5.5 | 4.2 | 1.4 | 0.2 |
| 39 | 4.4 | 3.0 | 1.3 | 0.2 |
| 41 | 5.0 | 3.5 | 1.3 | 0.3 |
| 42 | 4.5 | 2.3 | 1.3 | 0.3 |
| 43 | 4.4 | 3.2 | 1.3 | 0.2 |
| 46 | 4.8 | 3.0 | 1.4 | 0.3 |
| 48 | 4.6 | 3.2 | 1.4 | 0.2 |

Showing 1 to 20 of 50 entries

| Environment | History | Connections |
|--------------------|-----------------------------------|-------------|
| Global Environment | | |
| Data | | |
| iris.test | 50 obs. of 4 variables | |
| iris.train | 100 obs. of 4 variables | |
| Values | | |
| N | 150L | |
| tr.idx | int [1:100] 50 114 17 102 76 1... | |

- iris.test를 열어보면 50개의 데이터를 볼 수 있음
- iris.train을 열어보면 100개의 데이터를 볼 수 있음

5. 학습데이터와 검증데이터 생성

9.4 학습데이터와 검증데이터

- iris 데이터의 타겟변수 (학습데이터의 타겟변수, 검증데이터의 타겟변수)

```
# training/ test data : n=150
set.seed(1000, sample.kind="Rounding")
N=nrow(iris)
tr.idx=sample(1:N, size=N*2/3, replace=FALSE)
tr.idx

# attributes in training and test
iris.train<-iris[tr.idx,-5]
iris.test<-iris[-tr.idx,-5]

# target value in training and test
trainLabels<-iris[tr.idx,5]
testLabels<-iris[-tr.idx,5]
```



| Environment | History | Connections |
|--------------------|-------------------------------------|-------------|
| Global Environment | | |
| Data | | |
| iris.test | 50 obs. of 4 variables | |
| iris.train | 100 obs. of 4 variables | |
| Values | | |
| N | 150L | |
| testLabels | Factor w/ 3 levels "setosa", "v..." | |
| tr.idx | int [1:100] 50 114 17 102 76 1... | |
| trainLabels | Factor w/ 3 levels "setosa", "v..." | |

