	단위별 학습내용 (Week10)
wk10-1	k-인접기법(k-Nearest Neighbor)
wk10-2	k-인접기법 (가중치)
wk10-3	판별분석 I
wk10-4	판별분석 II

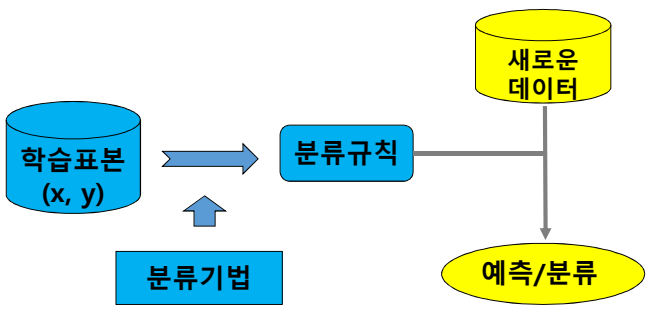
Wk10-1 : k-인접기법(k-Nearest Neighbor) I

1. 분류 (Classification)

모형화	특징	내용	적용기법
예측	• 타겟변수 값이 주어지는 경우 (supervised learning) • 변수간의 관계	주어진 데이터를 기반으로 모델을 만든 후, y값을 예측 (y=continuous value)	• 다중회귀분석 • 주성분 회귀분석 • 부분최소자승법 • 신경망
분류		학습표본을 기반으로 분류 규칙을 생성. 분류규칙의 성능을 검증하기 위해 실제 범주와 추정된 범주를 비교 (y=0/1 혹은 다범주)	• K-인접기법 • 로지스틱 회귀모형 • 의사결정나무 • 선형판별분석 • 서포트벡터머신
군집	• 타겟변수 값이 없는 경우 (unsupervised learning) • 개체간의 관계	주어진 데이터 (X변수들)의 속성으로 군집화	• 계층형 군집 분석 • K-MEANS
연관규칙		연관성 있는 변수관계 도출 (동시 발생 빈도 분석)	• 연관규칙 분석

1. 분류 (Classification)

- 분류(Classification) – 지도학습(Supervised Learning). 타겟범주를 알고 있는 데이터로 분류규칙을 생성하고 새로운 데이터를 특정범주에 분류하는 기법
- 군집화(Clustering) – 비지도학습((Unsupervised learning). 독립변수들의 속성을 기반으로 객체들을 그룹화하는 방법



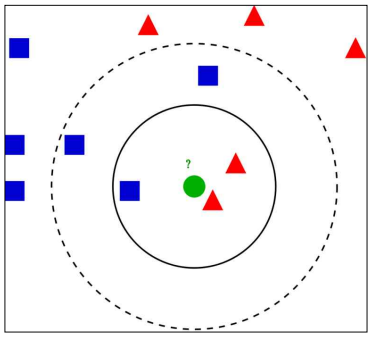
2. k-인접기법 (k-nearest neighbor method)

k-인접방법 (kNN) : k개의 가장 가까운 이웃들을 사용해서 분류하는 방법

- k개의 인접객체를 고려할 때, 새로운 객체 ● 는 어느 범주로 할당?

(1) 만약 k=3로 정하면, 새로운 객체는 ▲ 의 범주로 분류되고,

(2) 만약 k=5로 정하면, 새로운 객체는 ■ 의 범주로 분류된다.



From Wikipedia, the free encyclopedia

2. k-인접기법 (k-nearest neighbor method)

- 최적 k는?
  - k가 너무 크면 데이터 구조를 파악하기 어렵고, 너무 작으면 과적합(overfitting) 위험이 있음
  - 교차검증(cross-validation)으로 정확도가 높은 k를 선정

장점

- 단순하며 효율적
- 데이터 분산을 추정할 필요 없음
- 빠른 훈련 단계

단점

- 모델을 생성하지 않음
- 느린 분류 단계
- 많은 메모리 필요
- 결측치는 추가 작업 필요

2. k-인접기법 (k-nearest neighbor method)

- kNN 을 수행하기 위한 추가 패키지 설치

```
# lec10_12_knn.R

# lec10_1_knn.R
# Classification
# k-Nearest Neighbor

# packages
install.packages("class")#no weighted value knn
install.packages("gmodels")#crosstable
install.packages("scales")#for graph
library(class)
library(gmodels)
library(scales)
```

- kNN 수행을 위한 패키지 : "class"
- 분류분석 후 검증에 사용되는 cross table을 위한 패키지 : "gmodels"
- 최적 k 등 그래프를 위한 패키지 : " scales"

3. train/test 데이터 분할 (cross-validation)

- Iris 데이터 (데이터 불러들이기, 학습데이터와 검증데이터의 분할)

```
# set working directory
setwd("D:/tempstore/moocr/wk10")

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

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

# 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]

train<-iris[tr.idx,]
test<-iris[-tr.idx,]
```

데이터 불러들이기

데이터분할 (학습데이터 2/3, 검증데이터 1/3)

iris.train (독립변수4개를 포함한 100개의 데이터)
iris.test (독립변수4개를 포함한 50개의 데이터)

trainLabels (학습데이터의 타겟변수)
testLabels (검증데이터의 타겟변수)

4. kNN의 수행과 결과

• kNN함수 : knn(train=학습데이터, test=검증데이터, cl=타겟변수, k=)

```
# knn (5-nearest neighbor)
md1<-knn(train=iris.train,test=iris.test,cl=trainLabels,k=5)
md1
```

```
> md1
 [1] setosa     setosa     setosa
 [4] setosa     setosa     setosa
 [7] setosa     setosa     setosa
[10] setosa     setosa     setosa
[13] setosa     setosa     setosa
[16] setosa     versicolor versicolor
[19] versicolor versicolor versicolor
[22] versicolor versicolor versicolor
[25] versicolor versicolor versicolor
[28] versicolor versicolor versicolor
[31] versicolor versicolor virginica
[34] virginica  virginica  virginica
[37] virginica  virginica  virginica
[40] virginica  virginica  virginica
[43] virginica  versicolor virginica
[46] virginica  virginica  virginica
Levels: setosa versicolor virginica
```

• k=5를 한 kNN의 결과
md1에는 test 데이터(50개)들을 예측한 결과가 저장되어 있음

4. kNN의 수행과 결과

• knn의 매뉴얼 : help(knn)

R: k-Nearest Neighbour Classification Find in Topic

k-Nearest Neighbour Classification

Description

k-nearest neighbour classification for test set from training set. For each row of the test set, the k nearest (in Euclidean distance) training set vectors are found, and the classification is decided by majority vote, with ties broken at random. If there are ties for the kth nearest vector, all candidates are included in the vote.

Usage

```
knn(train, test, cl, k = 1, l = 0, prob = FALSE, use.all = TRUE)
```

Arguments

train	matrix or data frame of training set cases.
test	matrix or data frame of test set cases. A vector will be interpreted as a row vector for a single case.
cl	factor of true classifications of training set
k	number of neighbours considered.
l	minimum vote for definite decision, otherwise doubt. (More precisely, less than k-l dissenting votes are allowed, even if k is increased by ties.)
prob	If this is true, the proportion of the votes for the winning class are returned as attribute prob.
use.all	controls handling of ties. If true, all distances equal to the kth largest are included. If false, a random selection of distances equal to the kth is chosen to use exactly k neighbours.

5. kNN(k=5)의 결과 - 정확도


```
# accuracy of 5-nearest neighbor classification
CrossTable(x=testLabels,y=md1, prop.chisq=FALSE)
help(CrossTable)
```

타겟변수의 실제값 예측값

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  50


             | md1
  testLabels |     setosa | versicolor |  virginica |  Row Total |
-------------|------------|------------|------------|------------|
      setosa |         19 |          0 |          0 |         19 |
             |      1.000 |      0.000 |      0.000 |      0.380 |
             |      1.000 |      0.000 |      0.000 |            |
             |      0.380 |      0.000 |      0.000 |            |
-------------|------------|------------|------------|------------|
  versicolor |          0 |         17 |          2 |         19 |
             |      0.000 |      0.895 |      0.105 |      0.380 |
             |      0.000 |      0.944 |      0.154 |            |
             |      0.000 |      0.340 |      0.040 |            |
-------------|------------|------------|------------|------------|
   virginica |          0 |          1 |         11 |         12 |
             |      0.000 |      0.083 |      0.917 |      0.240 |
             |      0.000 |      0.056 |      0.846 |            |
             |      0.000 |      0.020 |      0.220 |            |
-------------|------------|------------|------------|------------|
Column Total |         19 |         18 |         13 |         50 |
             |      0.380 |      0.360 |      0.260 |            |
-------------|------------|------------|------------|------------|
```

- 정확도 : 47/50 -> 94%
- versicolor를 virginica로 오분류(2개)
- virginica를 versicolor로 오분류(1개)
- 오분류율 : 3/50 -> 6%