데이터과학을 위한 R프로그래밍

10주차. 서포트벡터머신

이혜선 교수
포항공과대학교 산업경영공학과

목차

10주차. 서포트벡터머신

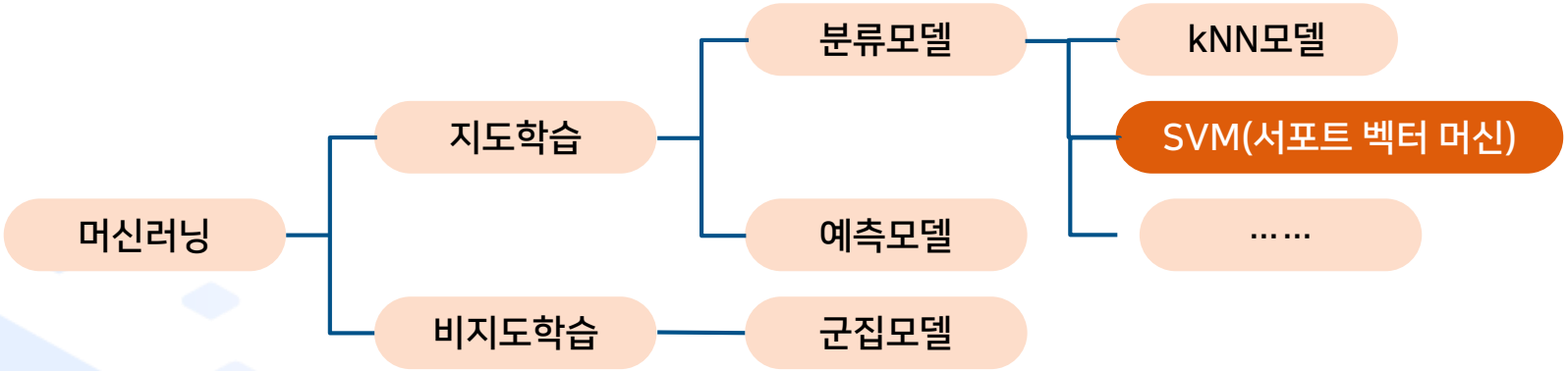
10주차

1차시

서포트벡터머신 I

POSTECH

서포트벡터머신(Support Vector Machine)



장점

- 정확도가 상대적으로 좋음
- 다양한 데이터(연속형, 범주형) 사용가능

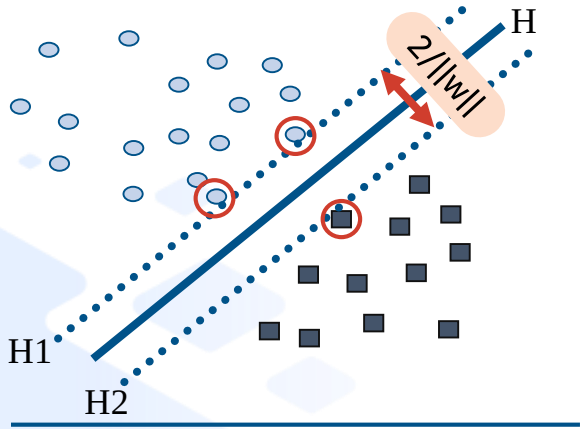
단점

- 해석하기 어려움
- 데이터가 많을 때 속도가 걸림

POSTECH

서포트벡터머신(Support Vector Machine)

선형 SVM



<분리 가능한 경우의 하이퍼플레인>

H1: w' x + b = 1
H2: w' x + b = -1

} H와 평행인 두 하이퍼플레인 (단, H1과 H2 사이에 객체 X)

H1, H2는 각각의 범주에서 H(분리 하이퍼플레인)와 가장 가까운 객체를 포함하는 평면

H1와 H2간의 거리(margin): 2/||w||

POSTECH

서포트벡터머신(Support Vector Machine)

선형 SVM

- H1와 H2간의 거리를 최대로 하는 분리 하이퍼플레인을 찾자!

$$\text{Max} \quad \frac{2}{w'w}$$

Subject to

$$w'x_i + b \geq 1 \quad for\ yi = 1 \qquad (i = 1, \ldots., \text{N})$$

$$w'x_i + b \leq -1 \quad for\ yi = -1 (i = 1, \ldots., \text{N})$$

최적화 문제 변환

$$\text{Min} \quad \frac{w'w}{2}$$

Subject to

$$y_i(w'x_i + b) \geq 1 \quad (i = 1, \ldots., \text{N})$$

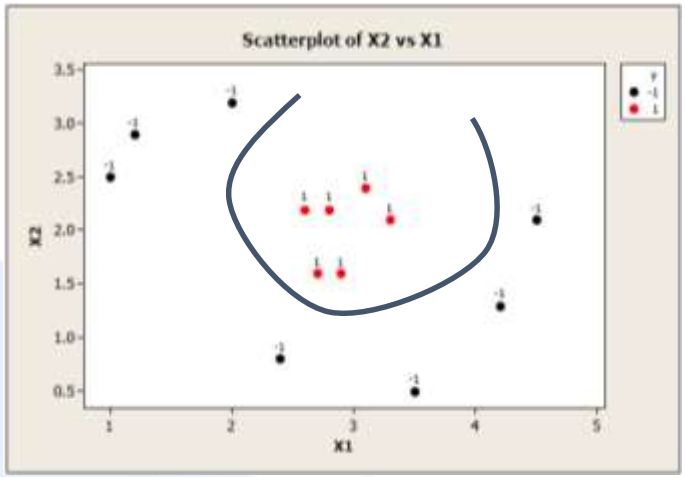
학습 표본 객체: N개

- x_i : N개의 변수로 이루어진 i번째 객체 벡터
- y_i : x_i 분류에 대응하는 범주에 대한 값(두 가지의 범주를 갖는다고 가정, y_i =1 or –1)

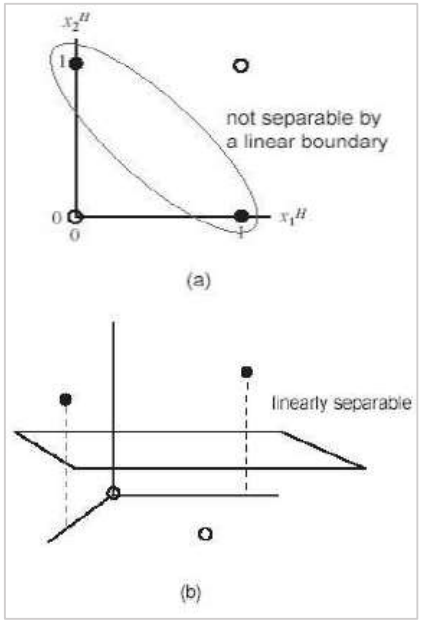
POSTECH

서포트벡터머신(Support Vector Machine)

비선형 SVM



<비선형 하이퍼플레인 도출>



<고차원(커널) 공간으로의 변환>

대부분의 패턴은 선형적으로 분리 불가능

- 비선형 패턴의 입력공간을 선형 패턴의 'feature space'로 변환
- Kernel method로 비선형 경계면 도출

POSTECH

iris 데이터 설명

iris 데이터(iris.csv)

input변수(독립변수)

output변수(종속변수, 타겟변수)

A	B	C	D	E
Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3	1.4	0.1	setosa

타겟변수(y) : setosa, versicolor, virginica

Iris setosa Iris versicolor Iris virginica

POSTECH

서포트벡터머신 패키지와 함수

- 서포트벡터머신을 수행하기 위한 패키지 : e1071
- 서포트벡터머신 함수 : svm

```
# install package for support vector machine
install.packages("e1071")
library (e1071)
#help(svm)
```

e1071 : svm함수 사용을 위한 패키지

```
# install package for confusionMatrix
#install.packages("caret")
library(caret)
```

caret : confusionMatrix 사용을 위한 패키지

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

# read data
iris<-read.csv("iris.csv",stringsAsFactors = TRUE)
attach(iris)
```

데이터 불러오기

POSTECH

서포트벡터머신 결과

- 서포트벡터머신 함수 : svm(y변수~x변수, data=)
- iris 데이터의 서포트벡터머신 결과(전체 데이터를 사용한 결과)

```
## classification
# 1. use all data
m1<- svm(Species ~., data = iris)
summary(m1)
```

svm의 결과 요약

```
> summary(m1)

Call:
svm(formula = Species ~ ., data = iris)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1

Number of Support Vectors:  51

 ( 8 22 21 )


Number of Classes:  3

Levels:
 setosa versicolor virginica
```

- svm에서 주어지는 옵션(default)

kernel=radial basis function,
gamma=1/(# of dimension) (1/4=0.25)

POSTECH

서포트벡터머신 결과

svm모델에 적용한 예측범주와 실제범주 비교(전체 데이터를 사용한 결과)

```
# classify all data using svm result (m1)
# first 4 variables as attribute variables
x<-iris[, -5]
pred <- predict(m1, x)

# Check accuracy (compare predicted class(pred)
# y <- Species or y<-iris[,5]
y<-iris[,5]
library(caret)
confusionMatrix(pred, y)
```

```
> confusionMatrix(pred, y)
Confusion Matrix and Statistics

            Reference
Prediction   setosa versicolor virginica
  setosa         50          0         0
  versicolor      0         48         2
  virginica       0          2        48
```

x<-iris[, -5]

iris데이터에서 타겟값인 5번째 열을 제외한 데이터,
즉 4개의 독립변수들만 있는 데이터

pred<-predict(m1, x)

svm모델 m1을 적용하여 예측된 범주값을 pred로 저장

오분류율

(2+2)/150=0.0266 (2.66%)

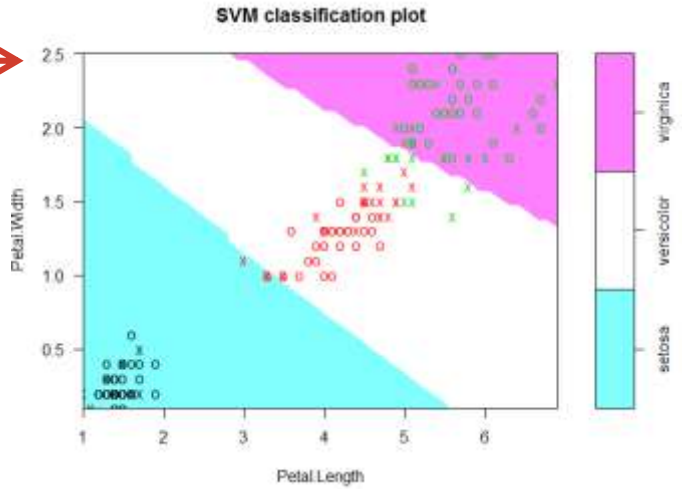
POSTECH

서포트벡터머신 결과 - 시각화

- iris 데이터의 서포트벡터머신 결과(전체 데이터를 사용한 결과)

svm결과를 그림으로 시각화

```
# visualize classes by color
plot(m1, iris,  Petal.Width~Petal.Length, slice=list(Sepal.Width=3,
```



4개 변수 중 petal.width와 petal.length가 중요 변수