1. 텍스트 마이닝(Text Mining)

chap17_1_TM_Data 수업내용

1) 텍스트 전처리

2) 문서/단어 행렬

3) 단어/문서 행렬

4) 가중치 설정 방법

1) 텍스트 전처리

- **텍스트 전처리 과정**

```
install.packages('tm')
install.packages('SnowballC') # stemDocument() 함수 제공
library(tm)
library(SnowballC)
sms_corpus = Corpus(VectorSource(sms_data$text)) # 1) 말뭉치 생성(vector -> corpus 변환)
sms_corpus = tm_map(sms_corpus, content_transformer(tolower))  # 2) 소문자 변경
sms_corpus = tm_map(sms_corpus, removeNumbers) # 3) 숫자 제거
sms_corpus = tm_map(sms_corpus, removePunctuation) # 4) 문자부호(콤마 등) 제거
sms_corpus = tm_map(sms_corpus, removeWords, stopwords("SMART")) # 5) 불용어 제거 제거
sms_corpus = tm_map(sms_corpus, stripWhitespace) # 6) 여러 공백 제거(공백 제거)
sms_corpus = tm_map(sms_corpus, stemDocument) # 7) 유사 단어 어근 처리
sms_corpus = tm_map(sms_corpus, stripWhitespace) # 8) 여러 공백 제거(공백 제거)
sms_dtm = DocumentTermMatrix(sms_corpus) # 9) 문서와 단어 집계표 작성
```

2) 문서/단어 행렬

- **DocumentTermMatrix**
  - ✓ 문서와 단어 이용 희소행렬 생성 함수
  - ✓ 줄 단위로 단어 생성

sms_dtm = DocumentTermMatrix(sms_corpus) # 9) 문서와 단어 집계표 작성

sms_dtm

```
<<DocumentTermMatrix (documents: 5558, terms: 6822)>>
Non-/sparse entries: 33629/37883047
Sparsity           : 100%
Maximal term length: 40
Weighting          : term frequency (tf)
```

<DocumentTermMatrix 예>

	a	aa	bb	cc	dd	ee	ff	gg	kk	….	zz
1	1	0	0	1	0	0	0	1	0	…….	0
2	0	1	0	0	0	0	0	0	0	…….	1

3) 단어/문서 행렬

- **TermDocumentMatrix**
  - ✓ DocumentTermMatrix -> TermDocumentMatrix 변경
  - ✓ 행(문서)과 열(단어)를 상호 변경(단어와 문서 구조 변경)
  - ✓ 단어의 출현 빈도수 확인 가능

t(sms_dtm)

```
<<TermDocumentMatrix (terms: 6822, documents: 5558)>>
Non-/sparse entries: 33629/37883047
Sparsity           : 100%
Maximal term length: 40
Weighting          : term frequency (tf)
```

< TermDocumentMatrix 예>

```
    1  2  3  4  5  6  7  8  9 10  …. n
a   1  0  0  1  0  0  0  1  0  0 …… 0
aa  0  1  0  0  0  0  0  0  0  0……..1
bb  0  0  0  0  0  0  0  0  0  0…… 0
:                  :
zz
```

단어/문서 행렬

doc
1. 대한민국은 나의 조국입니다.
2. 나는 홍길동 입니다.

DTM

	나	대한민국	조국	홍길동
1	1	1	1	0
2	1	0	0	1

TDM

	1	2
나	1	1
대한민국	1	0
조국	1	0
홍길동	0	1

4) 출현빈도 가중치

1. TF : 단어 출현 빈도수
2. TFiDF : 단어 출현 빈도수 * 문서 출현빈도수의 역수
   - tf-idf(d,t) = tf(d,t) * idf(t)
   - tf(d,t) : term frequency - 특정 단어 빈도수
   - idf(t) : inverse document frequency
     - 특정 단어가 들어 있는 문서 출현빈도수의 역수

   수식 : TFiDF = tf(d, t) x log(n/df(t)) : 문서 출현빈도수의 역수

4) 출현빈도 가중치

- **TF : 빈도수에 의한 가중치**

```
sms_dtm1 = DocumentTermMatrix(sms_corpus,
                control = list(wordLengths= c(1,8)))
```

```
table(sms_tdm1[2 ,]) # "aah"
#     0    1   <- 가중치
# 5555    3  <- 문서 출현빈도 : 3개 문서에서 1회씩 출현
```

- **TFiDF : 빈도수의 비율에 의한 가중치**

```
sms_dtm2 = DocumentTermMatrix(sms_corpus,
control = list(wordLengths= c(1,8),  weighting = weightTfIdf) )
```

```
table(sms_tdm2[2, ]) # 빈도수에 따라서 가중치 비율 증가
#    0 1.20615417983104 2.17107752369588 2.71384690461985
# 5555              1              1              1
```

4) 출현빈도 가중치

- TFiDF 방식의 가중치 희소행렬

	type	text
0	ham	우리나라 대한민국, 우리나라 만세
1	spam	비아그라 500GRAM 정력 최고!
2	ham	나는 대한민국 사람
3	spam	보험료 15000원에 평생 보장 마감 임박
4	ham	나는 홍길동

```
[[ 0.         0.         0.33939315 0.         0.42066906 0.         0.
   0.         0.         0.84133812 0.         0.         0.         0.   0.         0.        ]
 [ 0.5        0.         0.         0.         0.         0.         0.
   0.5        0.         0.         0.         0.         0.5        0.5  0.         0.        ]
 [ 0.         0.53177225 0.53177225 0.         0.         0.         0.
   0.         0.659118   0.         0.         0.         0.         0.   0.         0.        ]
 [ 0.         0.         0.         0.40824829 0.         0.40824829
   0.40824829 0.         0.         0.         0.40824829 0.40824829
   0.         0.         0.40824829 0.        ]
 [ 0.         0.62791376 0.         0.         0.         0.         0.
   0.         0.         0.         0.         0.         0.         0.   0.         0.77828292]]
```

2. 문서 분류예측 모형

- **문서 분류 예측 모형 유형**

1) Naive Bayes **chap17-2_NB**

2) Support Vector Machine **chap17-3_SVM**

1) Naive Bayes 알고리즘

1. 통계적 분류기
   - ✓ 주어진 데이터가 특정 클래스에 속하는지를 확률을 통해서 예측
   - ✓ 조건부 확률 이용 : P(B|A)
2. 베이즈 이론(Bayes' theorem)을 적용한 기계학습 방법
   - ✓ 두 확률 변수(사전 확률과 사후 확률) 사이의 관계를 나타내는 이론
   - ✓ 사전확률 : 사건이 발생하기 전에 알려진 확률
   - ✓ 사후확률 : 베이즈 이론에 근거한 확률(조건부 확률)
3. 특정 영역에서는 DT나 kNN 분류기 보다 성능이 우수
4. 고차원(텍스트 데이터)인 경우 높은 정확도와 속도 제공
5. 적용분야
   - ✓ **Spam 메일 분류, 문서(주제) 분류, 비 유무**
   - ✓ 컴퓨터 네트워크에서 침입자 분류(악성코드 유무)

조건부 확률(Conditional Probability)

- 사건 A가 발생했다는 전제 하에서 다른 사건 B가 발생할 확률

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad (\text{단 } P(A) > 0)$$

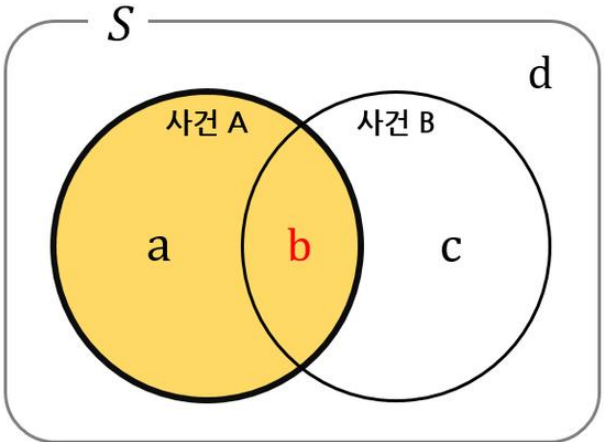
결합확률

$P(A \cap B) = P(B \cap A)$

$P(A \cap B) = P(B|A) \cdot P(A)$

$P(B \cap A) = P(A|B) \cdot P(B)$

- 벤다이어그램 표현



$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{b}{N}}{\frac{a+b}{N}}$$

$N = a+b+c+d$ 일 때

Bayes' Theorem

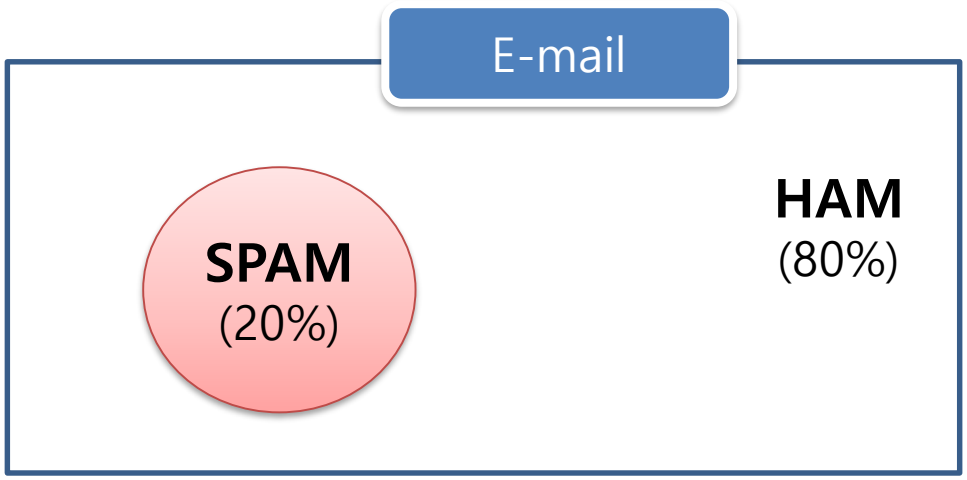
- 베이즈 정리 : 과거의 경험(사건 B)과 현재의 증거(사건 A)를 토대로 어떤 사건의 확률을 예측(추론)하는 이론
- P(A) : 사건 A 사전확률 : 현재의 증거
- P(B) : 사건 B 사전확률 : 과거의 경험
- P(B|A) : 사건 A의 증거에 대한 사후확률 : 사건 A가 일어났다는 것을 알고, 그것이 사건 B로부터 일어난 것이라고 생각되는 조건부 확률

  예) 비아그라 단어(A사건)가 포함될 때 스팸(B사건) 메시지일 확률

  - 사전확률 : 확률 실험 이전에 사건 발생에 대해 이미 알고 있는 사전 지식
  - 사후확률 : 어떤 사건을 인지한 후 이들이 어떤 원인에 의해 출현한 것이라고 생각되는 조건부 확률 지식

SPAM 메시지 분류 예

- 상호배타적, 포괄적 사건
  - ✓ Email 메시지에서 이미 알려진 Spam과 Ham 발생 비율 예



<독립사건인 경우 사전확률>

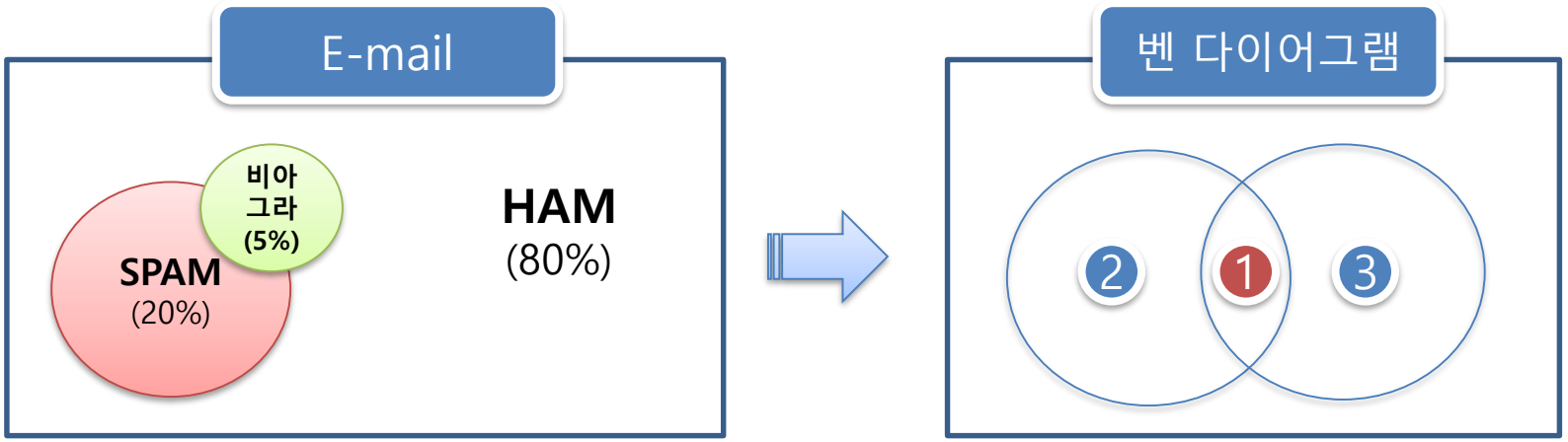
✓ SPAM과 HAM의 확률

- P(spam) = 0.2(20%)
- P(~spam) = 1-0.2=0.8(80%)

1. 상호배타적 : 동시에 두 사건이 일어나지 않음(독립사건)
   - ✓ 예) SPAM이 발생하면 HAM 발생하지 않음
2. 포괄적 : 두 가지 결과만 발생
   - ✓ 예) SPAM 또는 HAM 사건만 발생

● 비 상호배타적 사건

✓ Email 메시지에서 비아그라 속성 추가 예



1. 비상호배타적 : 동시에 두 사건이 일어남
2. 결합확률 : 두 사건이 동시에 일어날 확률

예) SPAM 메일에 비아그라 단어가 포함될 확률

예) HAM 메일에 비아그라 단어가 포함될 확률

● 벤 다이어그램 : 원소 집합의 중첩 표현

① : SPAM/HAM에 비아그라 단어 포함(5%)

② + ① : SPAM 메일(비아그라 단어 출현)

③ + ① : HAM 메일(비아그라 단어 출현)

- **사전확률(주변확률) : 조건 없이 사건A가 발생할 확률**
  - ✓ SPAM 확률 : 20/100 = 0.2
  - ✓ HAM 확률 : 80/100 = 0.8
  - ✓ VIAGRA 확률 : 5/100 = 0.05

- **결합확률 : 두 사건 A와 B가 동시 일어날 확률**
  - ✓ Viagra 단어가 포함된 Spam 메일 확률 : 4/20 = 0.2
  - ✓ Viagra 단어가 포함된 Ham 메일 확률 : 1/80 = 0.0125

구분	VIAGRA		합계
	Yes	No	
SPAM	4	16	20/100
HAM	1	79	80/100
합계	5/100	95/100	100

사전

사전

[사전확률 표]

구분	VIAGRA		합계
	Yes	No	
SPAM	4/20	16/20	20
HAM	1/80	79/80	80
합계	5/100	95/100	100

[결합확률 표]

- 조건부 확률 : 사건 A가 일어나는 조건하에서 사건 B가 일어날 확률
  - ✓ P(B|A) = P(A|B) * P(B) / P(A) = P(A∩B) / P(A)
- **베이즈 이론 적용** : 비아그라(A) 단어가 출현할 때 스팸(B)일 확률 : 80%
  - ✓ P(스팸|비아그라) = P(비아그라|스팸) * P(스팸) / P(비아그라)
  - ✓ 사후확률 = 결합확률 * 사전확률(사건B) / 사전확률(사건A)
  - ✓ P(스팸|비아그라) = (4/20) * (20/100) / (5/100) = 0.8

결합확률: 동시에 일어날 확률

사전확률: 현재의 증거

구분	VIAGRA		합계	주변 확률
	Yes	No		
SPAM	4/20	16/20	20	20/100
HAM	1/80	79/80	80	80/100
합계	5/100	95/100	100	

사전확률 : 과거의 경험

[결합확률 표]

∴ 비아그라 단어가 포함된 Message가 스팸 일 확률은 80%

2) Support Vector Machine 알고리즘

- 2000년대 초반에 많이 사용되는 분류 알고리즘
- SVM 알고리즘 - 이진분류 : 두 범주를 직선으로 분류
- 선형분리 - 2개의 집합(초평면:Hyperplane)을 직선으로 분리
  - 초평면 : 2차원 이상의 고차원 공간을 의미
  - 직사각형의 넓이(Margin) : Margin의 최대값을 구하는 것이 관건
  - Support Vectors : Margin과 가장 가까운 점들
- kNN과 선형회귀 모델링 기법이 적용 : 분류와 수치 예측 가능
- 비선형 분류를 위해서 데이터를 고차원 공간 사상(커널 트릭)

SVM 특징

- 다양한 데이터 셋에서 잘 동작하는 강력한 모델
  - ✓ 저차원, 고차원 데이터 모두 잘 동작
- 데이터의 특징이 적어도 복잡한 결정 경계 생성
- 모든 특징과 스케일이 비슷한 경우 유리함
  - ✓ 데이터 전처리와 매개변수 설정에 주의
- 샘플(관측치)이 많은 경우 불리함(100,000개 이상)
- 모델 분석이 어려움(블랙 box), 예측과정 이해가 어려움

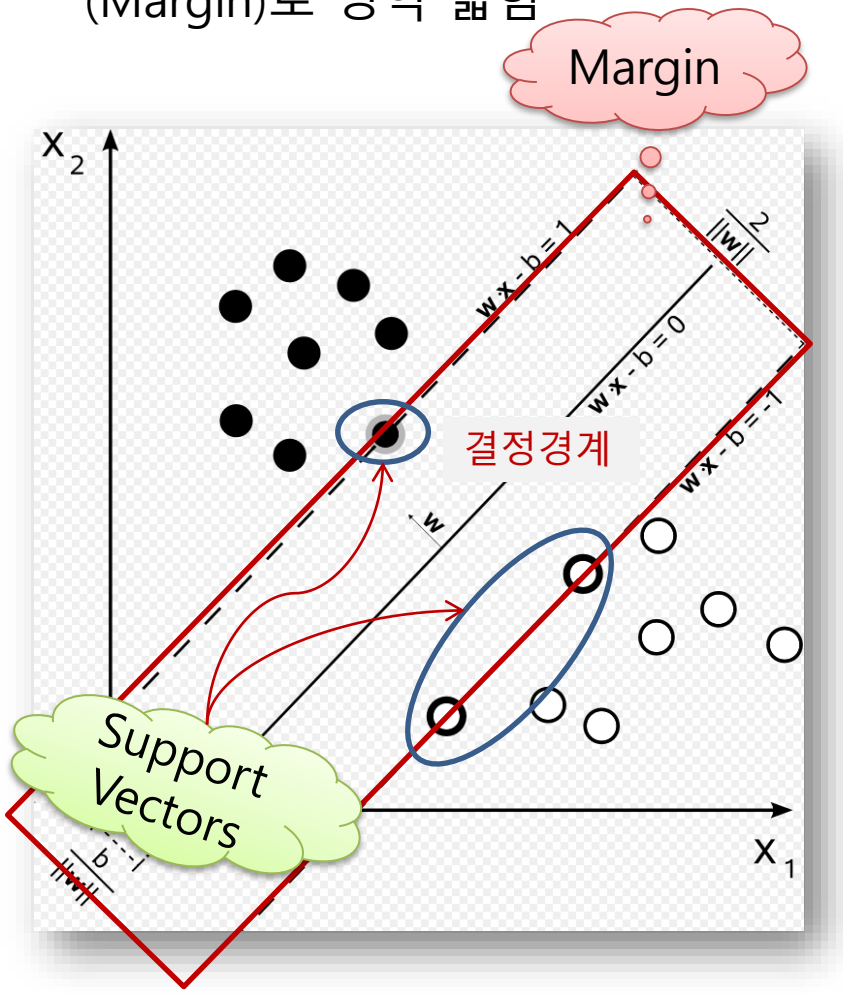
SVM 적용분야

- 적용 분야
  - ✓ 바이오인포매틱스의 마이크로 유전자 데이터 분류
  - ✓ 인간의 얼굴, 문자, 숫자 인식
    - 이미지 데이터 패턴 인식에 적합

    예) 스캐너로 스캔 된 문서 이미지를 문자로 인식

● Linear Classifier Margin

SVM 알고리즘 : 가상 직선 중심으로 거리 계산하여 최대의 직사각형 형태(Margin)로 영역 넓힘



Margin: 점에 닿기 전까지의 선형 분류기의 폭

Support Vectors: Marging과 닿는 점들

W : 초평면 Normal vector

w.x – b = 0 : 두 집합을 분류하는 가상 직선(1 or -1값)

w.x -b = 1 : X^+ 를 지나는 초평면
X^+ : +1 집합에서 가장 가까운 데이터(점)

w.x-b = -1 : X^- 를 지나는 초평면
X^- : -1 집합에서 가장 가까운 데이터(점)

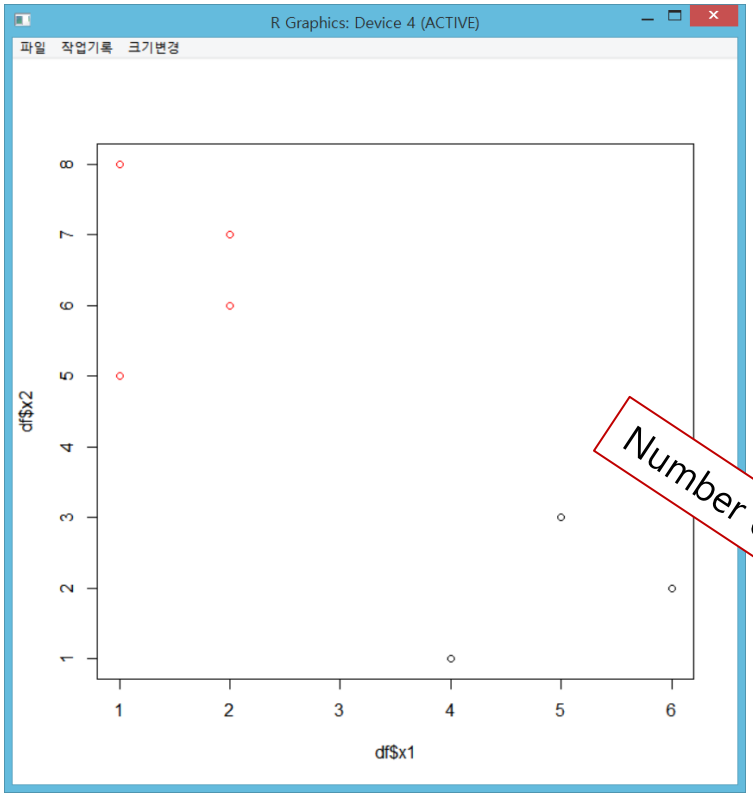
초평면 마진($\frac{2}{\|w\|}$) : 두 초평면 사이의 거리
(각 서포트 벡터를 지나는 초평면 사이의 거리)

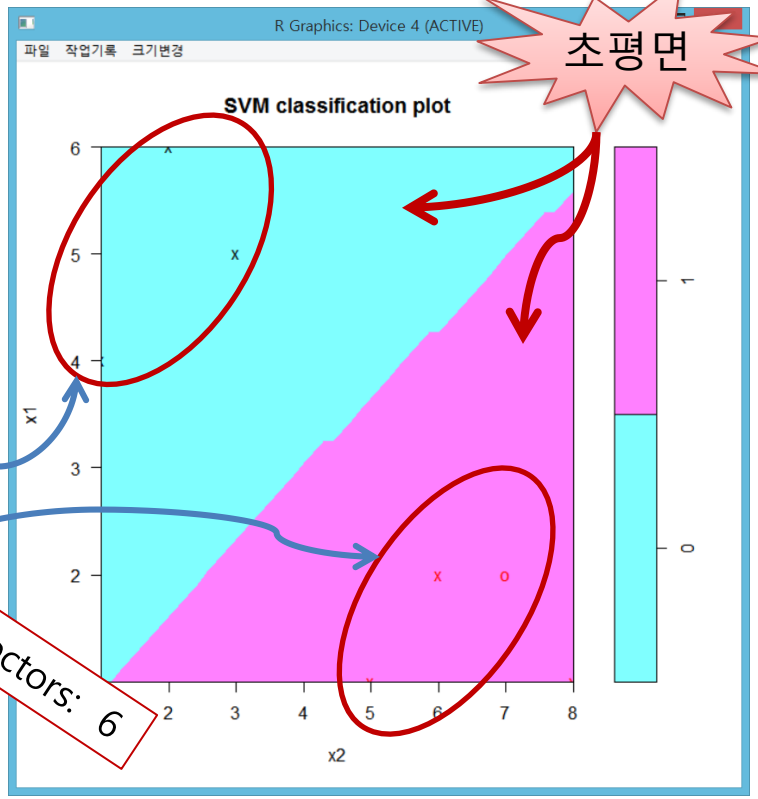
SVM은 이러한 마진을 최대로 만드는 알고리즘

● 초평면(Hyperplane)

서포트 벡터 머신은 분류 또는 회귀 분석에 사용 가능한 초평면(hyperplane) 또는 초평면 들의 집합으로 구성

2개의 집합 직선으로 분리





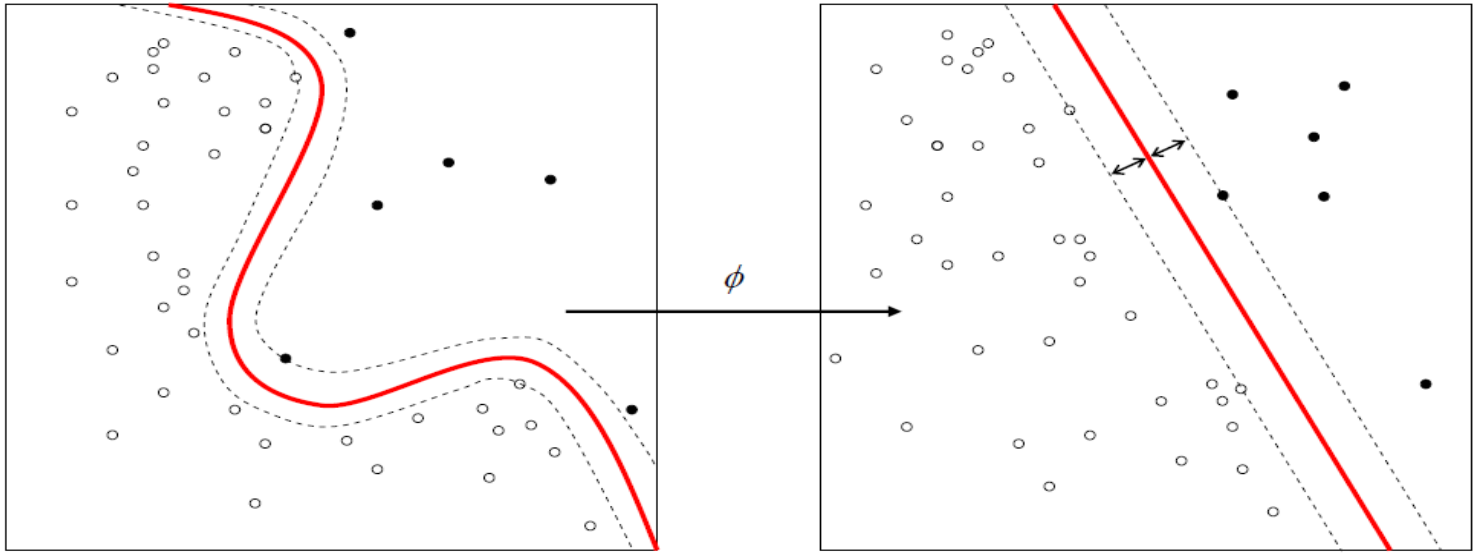
Number of Support Vectors: 6

Margin과 가장 가까운 점 6개

● Non-Linear Classifier Margin

1992년 Bernhard E. Boser, Isabelle M. Guyon and Vladimir N. Vapnik 제안
최대 마진 초평면 문제에 커널 트릭(Kernel Trick) 적용 비선형 분류 제안

- ➢ 커널 트릭(Kernel Trick) : 비선형(Non Linear) 관계를 선형으로 변환하는 역할
- ➢ 커널 함수(kernel function) : 커널 트릭에 사용되는 함수
- ➢커널 함수 종류 : linear(Gaussian), polynomial, radial, sigmoid



● 눈과 비 관계를 커널 트릭 예