17장. SVM 모형

이 장은 대표적 지도학습기계인 SVM, 즉 support vector machine을 다룬다. SVM은 버팀 점의 역할을 하는 소수의 관측 개체들로 분류 및 회귀 모형을 구축하는데 이 방법은 설명변수가 많은 경우에도 잘 작동한다. R의 e1071 팩키지의 svm( ) 함수를 써서 SVM 분류 및 회귀 모형을 만들어 볼 것이다.

1. 선형 SVM 분류

시작에 앞서 이 장부터 다루려하는 기계학습(machine learning)에 대하여 간략히 설명하고자 한다. 기계학습에서 '기계'는 컴퓨터를 지칭하고 '학습'은 경험적 모형 정도로 볼 수 있다. 이 분야가 공학에서 유래한다는 점을 빼고는 실제로 통계적 방법과 다르지 않다. 그러나 데이터마이닝이 뜨면서 기계학습이 마이닝 도구를 지칭하게 되었다.

기계학습(machine learning)은 지도 학습(supervised learning)과 비지도 학습(unsupervised learning)으로 구분된다. 지도 학습은 외적 기준이 되는 변수가 포함되어 있는 데이터(=경험)에서 나오는 경험적 지식이고 비(非)지도 학습은 외적 정보가 없는 데이터(=경험)에서 만들어지는 경험적 지식이다. 여기서 외적 기준 또는 정보라고 함은 통계학적 용어로는 종속변수 Y와 같은 것이다. 따라서 회귀모형이나 분류모형은 지도학습에 속한다. 반면, 군집화나 패턴기술(pattern description) 같은 것은 비지도학습이다.

SVM 분류는 분류규칙을 산출해내는 지도학습 방법이다. n개의 개체들에서 p개의 속성(attribute) $X_1, \cdots, X_p$가 측정되었고 외적 정보로 Y가 -1 또는 +1로 얻어졌다고 하자. 따라서 $Y=-1$인 개체들과 $Y=+1$인 개체들이 뒤섞인 자료로부터, 어떤 속성의 개체들의 Y가 -1이 되고 어떤 속성의 개체들의 Y가 +1이 되는지 분류규칙을 세워보자는 것이 취지이다.

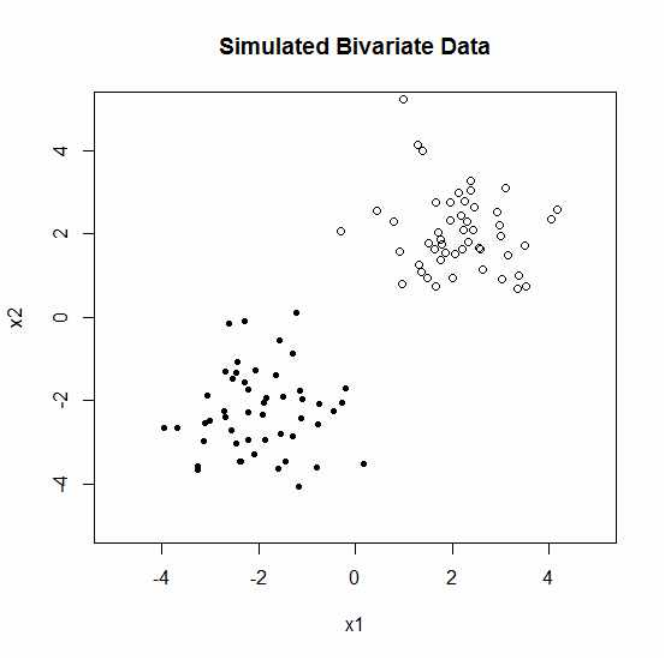

그림 1. 이변량 모의생성자료

선형 SVM 분류란 무엇인가? 단순한 예에서 시작하자. 그림 1의 데이터는 모의생성된 2변량 자료인데 50개의 까만 점과 50개의 하얀 점이 혼합되어 있다. 흑백을 나누는 경계는 무엇인가? 연필과 자로 누구든 그런 선을 그을 수 있을 것이다.

선형 SVM은 그런 경계선 중에서 가장 폭이 넓은 것을 찾아낸다. 그림 2를 보라. 중앙 경계선이 그어졌고 양 쪽 주변으로 경계선과 평행하게 울타리가 세워졌다. 울타리에 버팀 점이 된 관측개체는 3개뿐이다.

SVM은 어떻게 이런 경계를 찾는가? 자료를 $(\boldsymbol{x}_i^t, y_i)$로 표기하자 $(i = 1, \cdots, n)$. 여기서 $\boldsymbol{x}_i$는 $p \times 1$ 설명 벡터이고 y_i는 -1 또는 +1이다. 그림 1과 같이 총 개체군이 선형적으로 분리 가능한 경우, SVM 방법은 다음과 같이 구성된다.

- 찾고자 하는 선형 분류함수를 $f(\boldsymbol{x}) = \boldsymbol{w}^t\boldsymbol{x} + b$ 로 표기하자: $f(\boldsymbol{x}) > 0$이면 y를 +1로 예측하고 $f(\boldsymbol{x}) < 0$이면 y를 -1로 예측한다.

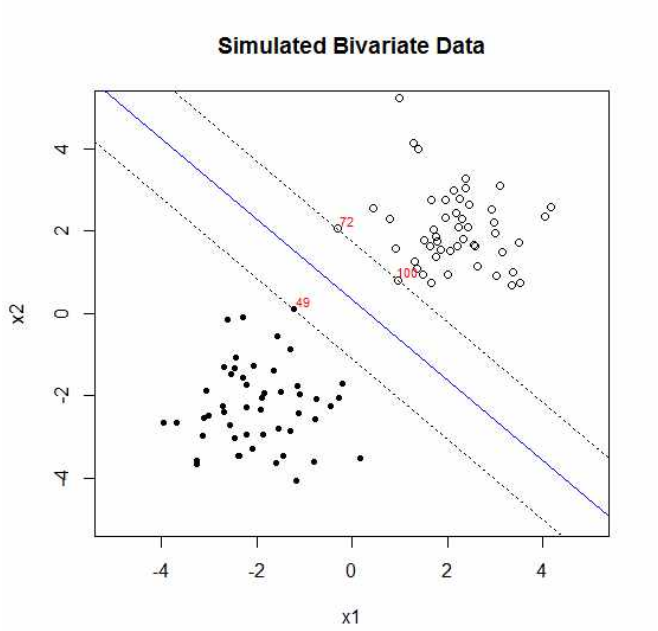

그림 2. SVM 버팀 점(support vector)과 분류선

- 제약조건이 따른다: $i = 1, \cdots, n$에 대하여

  1) $\boldsymbol{w}^t \boldsymbol{x}_i + b \geqq 1$, if $y_i = 1$,

  2) $\boldsymbol{w}^t \boldsymbol{x}_i + b \leqq -1$, if $y_i = -1$.

  이것은 선형적으로 분리되는 상황임을 수학적으로 표현한 것이다. 위의 2개 영역 간 폭은 $\| \boldsymbol{w} \|$의 역에 비례한다. 따라서 $\| \boldsymbol{w} \|$의 최소화가 목표이다.

- 따라서 SVM은 다음과 같이 수학적으로 정식화된다.

  minimize $\frac{1}{2} \| \boldsymbol{w} \|^2$ with respect to $\boldsymbol{w}$

  subject to 1) $\boldsymbol{w}^t \boldsymbol{x}_i + b \geqq 1$, if $y_i = 1$,

  2) $\boldsymbol{w}^t \boldsymbol{x}_i + b \leqq -1$, if $y_i = -1$.

- 라그라지 승수와 quadratic programming으로 다음과 같은 해를 얻는다.

$$\boldsymbol{w} = \sum_{i=1}^{n} \lambda_i y_i \boldsymbol{x}_i, \quad \lambda_1, \cdots, \lambda_n \geq 0.$$

- $\lambda_i > 0$인 개체 i의 설명 벡터 $\boldsymbol{x}_i$를 버팀 벡터(support vector)라고 한다. 버팀 벡터(점)들은 $\boldsymbol{w}^t \boldsymbol{x} + b = \pm 1$의 경계에 놓인다.

- 그림 1의 모의생성 자료에서는 버팀 점 3개(=개체 49, 72, 100)가 탐지되었다. 그림 2를 보라.

많은 상황에서는 데이터가 2개 그룹으로 선형적으로 명확히 나뉘지 않는다. 그림 3을 보라. 연필과 자로는 어떤 선을 긋더라도 흑점과 백점을 분리할 수 없다.

일반적으로 SVM은 어떻게 정식화되는가? 일부 개체들에 대하여 제약조건을 완화하지 않으면 안 된다. 그러나 조건완화에 페널티를 부과하여 해를 구한다.

- SVM 분류의 일반적 정식화:

$$\text{minimize} \quad \frac{1}{2} \| \boldsymbol{w} \|^2 + C \sum_{i=1}^{n} \xi_i \quad \text{w.r.t. } \boldsymbol{w} \text{ and } \xi_1 \geq 0, \cdots, \xi_n \geq 0$$

$$\text{subject to} \quad 1) \ \boldsymbol{w}^t \boldsymbol{x}_i + b \geq 1 - \xi_i, \quad \text{if } y_i = 1,$$

$$2) \ \boldsymbol{w}^t \boldsymbol{x}_i + b \leq -1 + \xi_i, \quad \text{if } y_i = -1.$$

여기서 $\xi_1 \geq 0, \cdots, \xi_n \geq 0$은 조건완화를 위한 여분(slack)이고 $C > 0$는 이 부분에 부과되는 단위비용(unit cost)이다.

그림 4는 그림 3 자료의 SVM 분류 결과이다 ($C = 100$). 네모 표시가 된 점들이 버팀 벡터들이다. 이들 점들은 분류 울타리의 경계와 내부에 위치한다. 그림에서 SVM 분류선은 실선으로 나타나 있다.

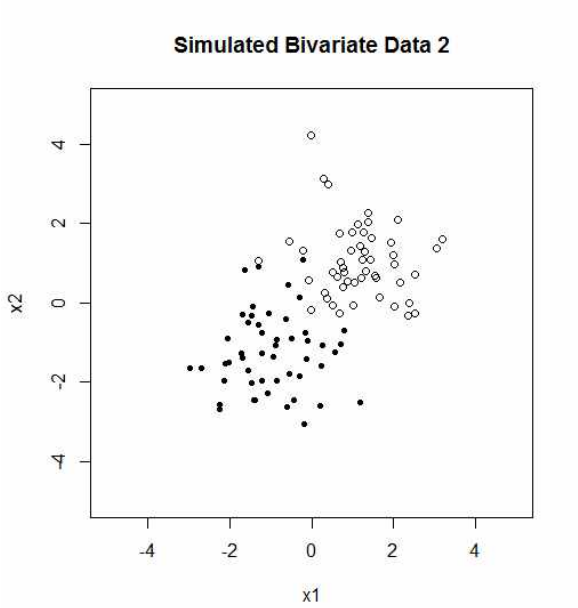

그림 3. 선형적으로 분리가능하지 않은 이변량 자료

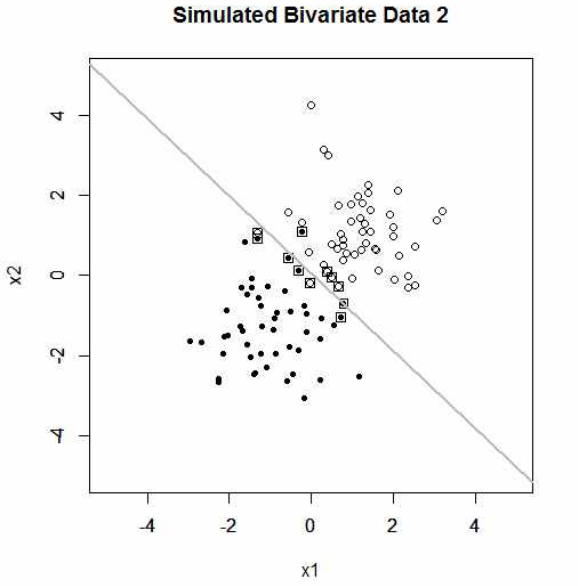

그림 4. SVM 분류의 버팀 벡터들 (네모 표시)

2. 비선형 SVM 분류

선형 SVM 분류는 말 그대로 '선형적'이다. 따라서 그룹 간 경계가 비선형인 경우 잘 작동하지 않게 된다. 예를 들어 그림 5의 케이스를 생각해 보자. 이 케이스에서는 $x_1^2 + x_2^2 \leqq 1.16$ 이면 group=1이고 $x_1^2 + x_2^2 > 1.16$ 이면 group=2이다. 즉 그룹 경계가 원형인 경우이다. 따라서 설명공간의 좌표를 (x_1, x_2) 대신 (x_1^2, x_2^2)으로 바꿀 필요가 있다. 이 케이스는 2변량의 경우이므로 산점도에서 시각적으로 필요한 변환을 찾을 수 있지만, 일반적으로 $p\,(\geqq 3)$변량의 자료에서는 그룹 분류를 위해 어떤 비선형 변환이 필요한지를 알아내기 쉽지 않다.

SVM 방법은 설명공간을 힐버트 공간(Hilbert space)으로 옮겨 비선형 분류의 문제를 해결한다. p차원 설명벡터 $\boldsymbol{x}$를 힐버트 공간의 $\Phi(\boldsymbol{x})$로 옮기자. 즉 Φ는 $\mathbb{R}^p$에서 $\mathbb{H}$로의 함수이다. 여기서 $\mathbb{R}^p$는 p차원의 유클리드 공간이고 $\mathbb{H}$는 힐버트 공간이다.

힐버트 공간 $\mathbb{H}$에서 두 개체 $\Phi(\boldsymbol{x}_1)$과 $\Phi(\boldsymbol{x}_2)$ 간 내적(內積) $<\Phi(\boldsymbol{x}_1), \Phi(\boldsymbol{x}_2)>$는 몇 개의 특정한 커널 함수 $K(\boldsymbol{x}_1, \boldsymbol{x}_2)$로 얻어진다. 가우스 커널이 대표적이다:

가우스 커널(일명 radial kernel):

$$K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp(-\gamma \parallel \boldsymbol{x}_1 - \boldsymbol{x}_2 \parallel^2),\ \gamma > 0.$$

유클리드 공간 $\mathbb{R}^p$에서 두 개체 $\boldsymbol{x}_1$과 $\boldsymbol{x}_2$ 간 내적(內積) $<\boldsymbol{x}_1, \boldsymbol{x}_2>$에 해당하는 커널은 $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \boldsymbol{x}_1^t \boldsymbol{x}_2$이다. 이 커널은 선형 커널(linear kernel)로 불린다. 이 밖에 다항 커널과 로지스틱 커널 등이 있다.

다항 커널(polynomial kernel): $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = (\gamma\, \boldsymbol{x}_1^t \boldsymbol{x}_2 + c_0)^d,\ d = 1, 2, \cdots$.

로지스틱 커널(logistic kernel): $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \tanh(\gamma\, \boldsymbol{x}_1^t \boldsymbol{x}_2 + c_0)$.

여기서 $c_0 \geqq 0,\ \gamma > 0$이다.

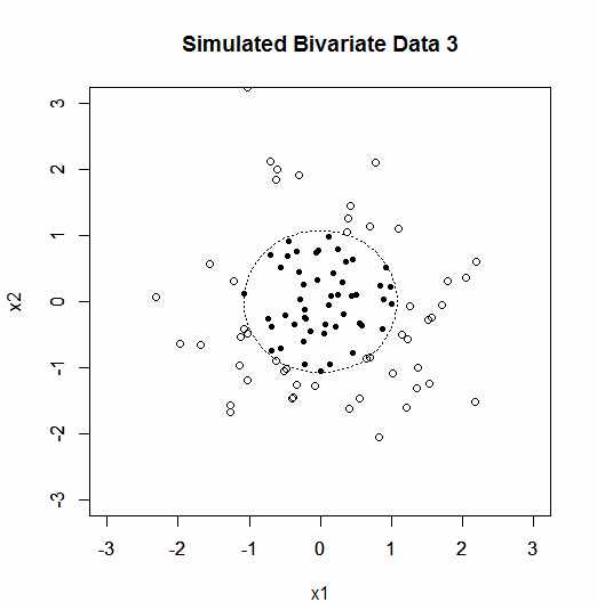

그림 5. 이변량 모의생성 자료: 비선형적 케이스

비선형 SVM 방법은 분류함수를 $f(\boldsymbol{x}) = \boldsymbol{w}^t \Phi(\boldsymbol{x}) + b$ 로 놓고 $\Phi(\boldsymbol{x}_i)$들의 선형결합, 즉 $\sum_{i=1}^{n} w_i \Phi(\boldsymbol{x}_i)$ 중에서 계수 $\boldsymbol{w}$를 다음 정식화에 따라 찾는다.

$$\text{minimize} \quad \frac{1}{2} \| \boldsymbol{w} \|^2 + C \sum_{i=1}^{n} \xi_i \quad \text{w.r.t. } \boldsymbol{w} \text{ and } \xi_1 \geq 0, \cdots, \xi_n \geq 0$$

$$\text{subject to} \quad 1) \ \boldsymbol{w}^t \Phi(\boldsymbol{x}_i) + b \geq 1 - \xi_i, \quad \text{if } y_i = 1,$$

$$2) \ \boldsymbol{w}^t \Phi(\boldsymbol{x}_i) + b \leq -1 + \xi_i, \quad \text{if } y_i = -1.$$

여기서 $\xi_1 \geq 0, \cdots, \xi_n \geq 0$은 여분(slack)이고 $C > 0$는 단위비용이다.

이 정식화를 풀어서 얻어지는 분류함수는 다음 형태가 된다.

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \lambda_i y_i < \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}) > + b = \sum_{i=1}^{n} \lambda_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b .$$

그림 6은 그림 5 자료에 가우스 커널을 적용하여 분류한 결과이다 ($\gamma = 0.5$,

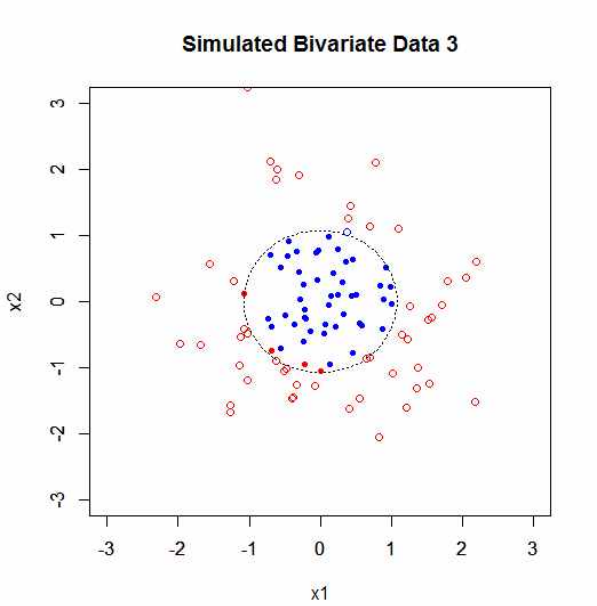

그림 6. 비선형적 케이스에 대한 SVM 분류 결과: 가우스 커널

$C = 1$). 청색 점이 group=1로 분류된 개체들이다. 원의 경계에서 일부 혼동이 있기는 하지만 대체로 잘 분류되어 있다.

다음은 SVM 분류에 쓰인 R 스크립트인데 e1071 팩키지의 `svm( )` 함수를 사용하였다 (simulated 2-dim nonlinear.r). `svm( )`에서 `kernel`은 "`radial`"로 지정하였고 `gamma`는 0.5로 하였다. `scale`은 `FALSE`로 두어 원 척도를 썼다.

```
set.seed(123)
x <- matrix(rnorm(200),100,2)
grp <- ifelse(apply(x*x,1,sum) <= 1.16, 1, 2)
table(grp)

library(e1071)
y <- as.factor(grp)
svm.model <- svm(y ~ x, kernel="radial", scale=F, gamma=0.5)
summary(svm.model)
```

```
windows(height=8,width=7)
plot(x, pch=c(20,21)[grp], col=c("blue","red")[svm.model$fitted],
    xlim=c(-3,3), ylim=c(-3,3), xlab="x1", ylab="x2",
    main="Simulated Bivariate Data 3")
theta <- seq(0,1,0.01)*2*pi; r <- sqrt(1.16)
par(new=T); plot(r*cos(theta), r*sin(theta), lty="dotted", type="l",
    xlim=c(-3,3), ylim=c(-3,3), xlab="", ylab="")
```

스팸 메일 사례. kernlab 팩키지의 spam 자료는 휴렛-팩커드 사에서 수집된 4,601개의 e메일에 대한 스팸 분류(spam 또는 nonspam; 변수 명 type)와 57개 변수 x1-x57 (단어 및 기호의 빈도 • 특성)으로 구성되어 있다. x1-x57을 써서 type에 대한 비선형 SVM 분류를 해보자. 다음과 같이 R 스크립트를 작성하였다 (spam.r).

```
> library(e1071)
> library(kernlab)
> data(spam); str(spam)
> svm.model <- svm(type ~ ., data=spam, gamma=1, cost=1)

> addmargins(table(spam$type, svm.model$fitted))
          nonspam spam  Sum
  nonspam    2788    0 2788
  spam         27 1786 1813
  Sum        2815 1786 4601
```

적합모형에 자료를 넣어 예측한 결과 1,813개 스팸 중에서 27개를 nonspam으로 분류하여 1.5%의 오류율을 보였다.[1] 2,788개 비(非)스팸에 대해서는 모두 옳게 예측함으로써 제로 오류율을 기록하였다. 총(總)오류율은 0.6%(=(27+0)/4601)이다. 인상적이지 않은가? 그렇게 보이지만, 앞의 분석에서 분류 정확도는 과장되어 있다. 왜냐하면 모형적합에 쓰인 자료(training data, 훈련자료)와 모형성과의 테스트에 쓰인 자료(test data, 테스트 자료)가 같기 때문이다.

---

1) 표에서 행은 Y의 실제 값이고 열은 Y의 예측 값이다.

전체자료를 일정 비율로 분할(partition)하여 훈련자료와 테스트 자료를 겹치지 않게 할 필요가 있다. 다음은 전체자료를 3:1로 분할하는 예이다.

```
> n <- nrow(spam)
> sub <- sample(1:n, round(0.75*n))
> spam.1 <- spam[sub,]
> spam.2 <- spam[-sub,]
```

spam.1은 훈련자료로서 3,451개의 개체를 가져가고 spam.2는 테스트 자료로서 나머지 1,150개의 자료를 가져간다. 이제 spam.1으로 SVM 분류 모형을 적합하고 spam.2 개체들의 Y 예측치를 만든 다음 실제 Y와 대조해보자.

```
> svm.model.1 <- svm(type ~ ., data=spam.1, gamma=1, cost=1)
> svm.predict.2 <- predict(svm.model.1, newdata=spam.2)

> addmargins(table(spam.2$type, svm.predict.2))
          svm.predict.2
          nonspam spam  Sum
  nonspam     682    1  683
  spam        236  231  467
  Sum         918  232 1150
```

spam.1으로 적합한 SVM 분류 모형에 spam.2 자료를 넣어 예측한 결과 467개 스팸 중에서 236개를 nonspam으로 분류하여 50%가 넘는 오류율을 보였다. 그리고 683개 비(非)스팸 중에 대해서는 1개를 스팸으로 예측함으로써 0.15%의 오류율을 기록하였다. 총(總)오류율은 20.6%(=(236+1)/1150)이다. 앞에서 총오류율이 0.6%였는데 이것이 얼마나 과소한 것이었던가를 실감할 수 있다.

이제까지 `svm( )`에서 `kernel`은 `gamma=1, cost=1`로 하였는데 이들 파라미터를 다르게 두면 모형성과가 다르게 나타날 수 있다. e1071 팩키지의 `tune.svm( )`으로 SVM 모형의 최적 파라미터를 찾아보자. 이 함수는 10-겹 교차 타당성 평가(10-fold cross-validation) 방법을 쓴다.[2] 가우스 커널의 `gamma` 값 0.1, 1,

2) 전체 자료를 10개로 나누어 9개 부(副,sub) 자료를 합하여 모형적합을 하고 나머지 1개

10과 cost 값 0.1, 1, 10의 총 9개 조합 중에서 최적의 (gamma, cost) 조합을 찾아보자.

```
> tune.svm <- tune(svm, type ~ .,  data=spam.1,
                   ranges=list(gamma=c(0.1,1,10),cost=c(0.1,1,10)))
> summary(tune.svm)

  - sampling method: 10-fold cross validation

  - best parameters:
    gamma cost
     0.1   10

  - best performance: 0.09329068

  - Detailed performance results:
    gamma cost      error dispersion
1   0.1  0.1 0.18336669 0.02816067
2   1.0  0.1 0.38158358 0.02980236
3  10.0  0.1 0.39027006 0.02932452
4   0.1  1.0 0.09676559 0.01431572
5   1.0  1.0 0.21293232 0.02368920
6  10.0  1.0 0.25639253 0.02279772
7   0.1 10.0 0.09329068 0.01889967
8   1.0 10.0 0.20829877 0.02424337
9  10.0 10.0 0.25320661 0.02237121
```

위 출력에 의하면, "radial"(가우스) 커널의 SVM 분류 모형에서 gamma=0.1과 cost=10의 조합이 평균 총 오류율 9.3%로 최적이다.[3)]

부자료로 적합 모형을 테스트하는 방법이다. 1개 부자료의 선택에 10개 경우가 있으므로 총 10번의 테스트를 하는 셈이 된다.

3) 계산처리 시간이 상당히 걸린다. 저자의 노트북에서는 744 sec만에 결과가 나왔다.

```
p.time <- proc.time()
tune.svm <- tune(svm, type ~ .,  data=spam.1,
                 ranges=list(gamma=c(0.1,1,10),cost=c(0.1,1,10)))
summary(tune.svm)
proc.time()-p.time
```

3. 선형 및 비선형 SVM 회귀

선형 SVM 회귀에서 회귀함수는 $f(\boldsymbol{x}) = \boldsymbol{w}^t\boldsymbol{x} + b$의 꼴이다. SVM 회귀는 y_i가

$$(f(\boldsymbol{x}_i) - \epsilon, f(\boldsymbol{x}_i) + \epsilon) \tag{1}$$

내에 오도록 회귀함수 $f(\boldsymbol{x})$를 잡는 것을 이상으로 한다 ($\epsilon > 0$, $i = 1, \cdots, n$).

많은 실제 사례에서는 이것이 가능하지 않다. 그런 사례에서는 $y_i - f(\boldsymbol{x}_i) \geqq \epsilon$이면 y_i를 $y_i - \xi_i\ (= y_i^*)$로 조정하고 $y_i - f(\boldsymbol{x}_i) \leqq -\epsilon$이면 y_i를 $y_i + \xi_i\ (= y_i^*)$로 조정하여, y_i^*가 (1)의 구간 $(f(\boldsymbol{x}_i) - \epsilon, f(\boldsymbol{x}_i) + \epsilon)$에 담기도록 $f(\boldsymbol{x})$를 잡는다. 여기서 "여분"(slack) ξ_i는 비음(nonnegative)이다 ($i = 1, \cdots, n$).

그림 7은 $x \sim N(0,1)$, $f(x) = 0.8x$, $y = f(x) + e$, $e \sim N(0, 0.6)$에서 생성된 2변량 자료이다. 참 회귀선을 중심으로 폭이 $\epsilon\ (= 1)$인 띠의 밖에 있는 관측점의 경우 관측점에서 띠의 경계까지의 수직 선분의 길이가 "slack"(여분)이다.

여분의 총 크기 $\sum_{i=1}^{n} \xi_i$는 가급적 작아야 한다. 따라서 선형 SVM 회귀는 다음과 같이 정식화된다.

$$\text{minimize } \frac{\|\boldsymbol{w}\|^2}{2} + C\sum_{i=1}^{n}\xi_i \text{ with respect to } \boldsymbol{w} \text{ and } \xi_1, \cdots, \xi_n.$$

$$\begin{aligned} \text{subject to } \quad & y_i - f(\boldsymbol{x}_i) - \xi_i \leqq \epsilon, \quad \text{if} \quad y_i - f(\boldsymbol{x}_i) \geqq \epsilon, \\ & y_i - f(\boldsymbol{x}_i) + \xi_i \geqq -\epsilon, \text{ if } y_i - f(\boldsymbol{x}_i) \leqq -\epsilon, \\ & \text{for } i = 1, \cdots, n. \end{aligned}$$

SVM 방법론에 의하면 이것의 해는 다음 형태로 주어진다.

$$\boldsymbol{w} = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\boldsymbol{x}_i, \qquad f(\boldsymbol{x}) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\boldsymbol{x}_i^t\boldsymbol{x} + b.$$

여기서 α_i와 α_i^*는 비음의 Lagrange 승수이다.

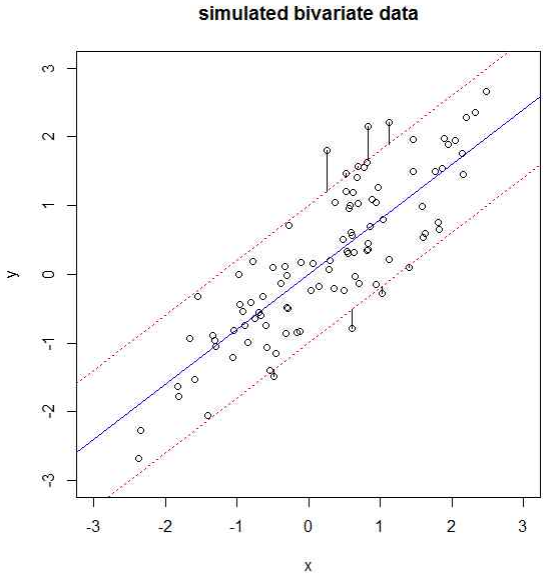

그림 7. 모의생성 2변량 자료에서의 선형 SVM 회귀 개념도

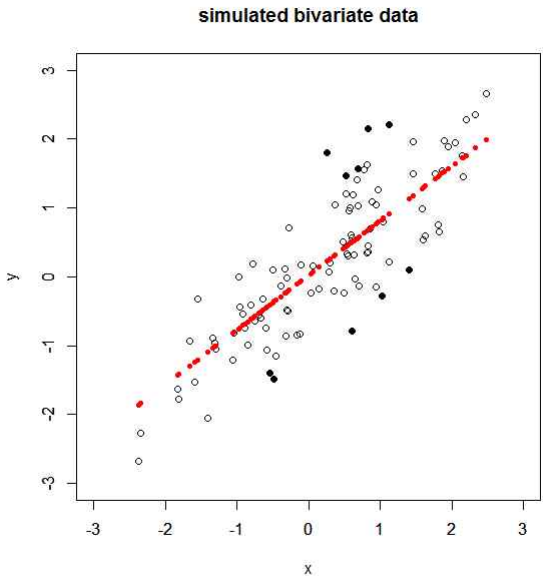

그림 8. 모의생성 2변량 자료에서 선형 SVM 회귀 적합모형

그림 8은 그림 7의 자료의 선형 SVM 회귀를 보여준다. 예측값들은 대각의 경향 직선에 놓인다. 그 직선을 중심으로 $\pm\epsilon$의 띠 밖의 개체 점들은 청색 컬러로 채워졌는데 ($\epsilon = 1$) 이들이 받침 벡터(support vector)들이다. 그림 8의 작성을 위해 다음 R 스크립트가 쓰였다 (파일 명: simulated 2-dim linear regression.r).

```
set.seed(12345)
x <- rnorm(100)
y <- 0.8*x + rnorm(100,0,0.6)
library(e1071)
svm.model <- svm(y ~ x, kernel="linear", epsilon=1, scale=F)
summary(svm.model)

windows(height=7.5,width=7)
plot(y ~ x, main="simualted bivariate data",xlim=c(-3,3),ylim=c(-3,3))
par(new=T)
plot(svm.model$fitted ~ x, main="",xlim=c(-3,3),ylim=c(-3,3),
     xlab="",ylab="",col="red",pch=20)
points(x[svm.model$index],y[svm.model$index],pch=
```

비선형 SVM 회귀에서 회귀함수는 $f(\boldsymbol{x}) = \boldsymbol{w}^t \Phi(\boldsymbol{x}) + b$로 바뀌고 적합모형은 다음 형태로 주어진다.

$$\boldsymbol{w} = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \Phi(\boldsymbol{x}_i), \qquad f(\boldsymbol{x}) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(\boldsymbol{x}_i, \boldsymbol{x}) + b.$$

여기서 α_i와 α_i^*는 비음의 Lagrange 승수이고 $K(\boldsymbol{x}, \boldsymbol{x}')$은 커널 함수이다.

그림 9는 $x \sim N(0,1)$, $f(x) = 0.8x^2$, $y = f(x) + e$, $e \sim N(0, 0.6)$에서 생성된 2변량 자료이다. 중앙의 속이 채워진 적색 점들이 2차식인 회귀함수 $f(x)$이다. $y = f(x)$를 중심으로 폭이 $\epsilon\,(=1)$인 띠의 밖에 있는 개체들이 몇 개 있는데 이들은 청색으로 채워져 있다.

그림 10은 그림 9의 자료의 비선형 SVM 회귀(가우스 "radial" 커널, epsilon=1, gamma=0.5, cost=1)를 보여준다. 예측값들은 채워진 적색 점으로 표지되었는데

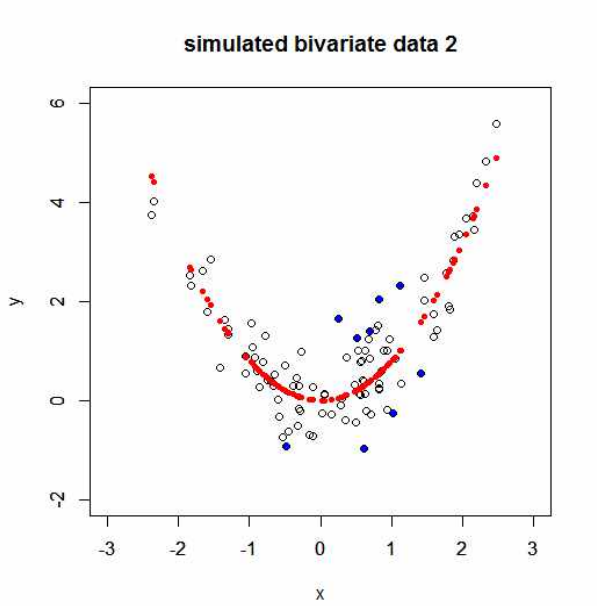

그림 9. 2차 곡선 패턴의 모의생성 자료: 회귀곡선과 띠의 안팎

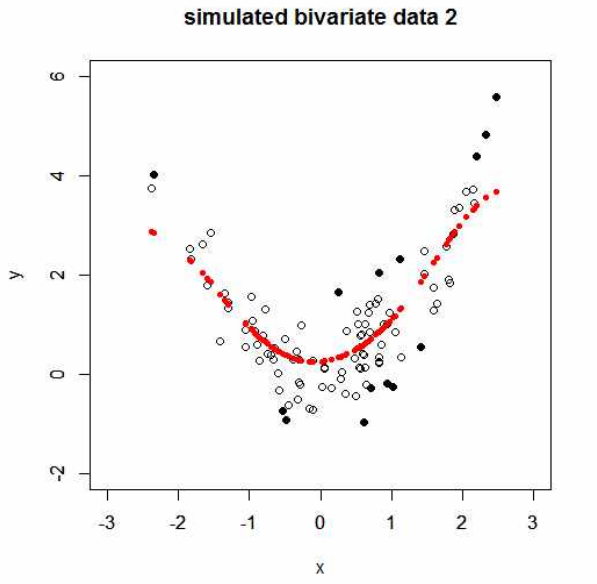

그림 10. 모의생성 2변량 자료에서 비선형 SVM 회귀 적합모형

2차식에 가까운 곡선 상에 놓여 있다. 이 경향곡선을 중심으로 $\pm\epsilon$의 띠 밖에 있는 채워진 청색 점들이 비선형 SVM 회귀를 결정하는 받침 벡터이다 ($\epsilon = 1$).

그림 10의 작성을 위해 다음 R 스크립트가 쓰였다 (파일 명: simulated 2-dim nonlinear regression.r).

```
set.seed(12345)
x <- rnorm(100);  y.fit <- 0.8*x^2
y <- y.fit + rnorm(100,0,0.6)

library(e1071)
svm.model <- svm(y ~ x, gamma=0.5, epsilon=1, scale=F)
summary(svm.model)

windows(height=7.5,width=7)
plot(y ~ x, main="simuated bivariate data 2",xlim=c(-3,3),ylim=c(-2,6))
par(new=T)
plot(svm.model$fitted ~ x, main="",xlim=c(-3,3),ylim=c(-2,6),
     xlab="",ylab="",col="red", pch=20)
points(x[svm.model$index],y[svm.model$index],pch=
```

오존 연구 사례. gclus 팩키지의 ozone 자료는 미국 Los Angeles 시 인근에서 330일에 걸쳐 측정된 오존 및 기상 변인에 대한 기록이다. 다음 변수들로 구성되어 있다.

```
Ozone:  Ozone conc., ppm, at Sandbug AFB.
Temp:   Temperature F. (max?).
InvHt:  Inversion base height, feet
Pres:   Daggett pressure gradient (mm Hg)
Vis:    Visibility (miles)
Hgt:    Vandenburg 500 millibar height (m)
Hum:    Humidity, percent
InvTmp: Inversion base temperature, degrees F.
Wind:   Wind speed, mph
```

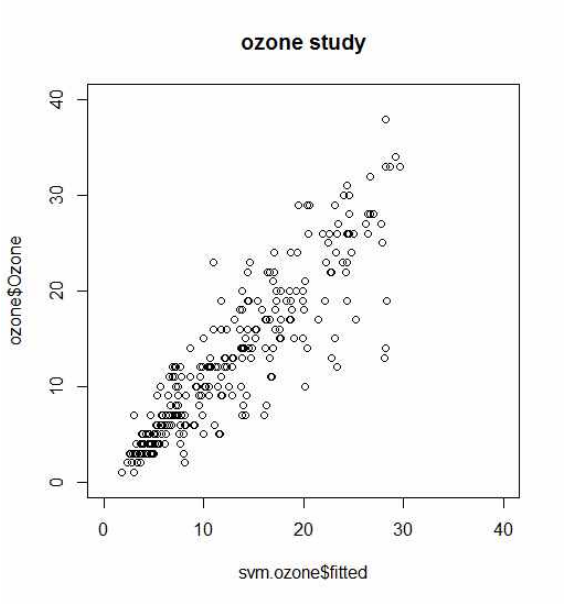

그림 11. 오존에 대한 비선형 SVM 회귀 1

종속변수로 Ozone을, 그리고 나머지 변수들을 설명변수로 하자. 그림 11이 오존의 비선형 SVM 회귀 적합값 대 관측값의 플롯이다.

적합 SVM 회귀에 쓰인 파라미터는 cost=1, gamma=0.125, epsilon=0.1이다. R 스크립트는 다음과 같다 (ozone.r).

```
library(gclus)
data(ozone); str(ozone)
library(e1071)
svm.ozone <- svm(Ozone ~ ., data=ozone, cost=1)
summary(svm.ozone)
windows(height=7.6, width=7)
plot(ozone$Ozone ~ svm.ozone$fitted, main="ozone study",
     xlim=c(0,40),ylim=c(0,40))
cor(svm.ozone$fitted,ozone$Ozone)
```

비선형 SVM 회귀 적합값과 관측값 간 상관은 0.91로 좋은 편이지만 버팀 벡터의 수가 무려 250개나 된다. 버팀 벡터 수를 줄이기 위해서는 epsilon을 크게 할 필요가 있다.

연습문제 1. spam 사례에서 spam 수는 1,813개이고 nonspam 수는 2,788개로 불균형하다 (0.8대 1.2). 따라서 spam에 1.2의 가중치를 주고 nonspam에 0.8의 가중치를 주어 클래스 크기의 불균형을 교정하는 방안을 고려할 수 있다. 그렇게 하는 경우, 최선의 SVM 분류모형은?
[힌트: `svm( )` 함수에서 `class.weights` 옵션을 써볼 것]

연습문제 2. ozone 사례에서 “linear” 커널을 쓰는 경우 SVM 회귀 적합값과 ozone 값 간 상관은 어느 정도가 되는가?