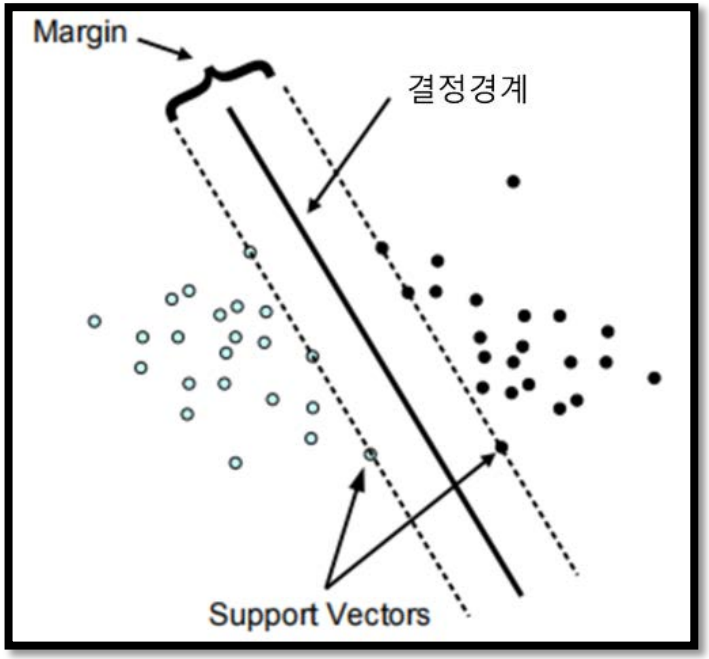
6장 데이터 분석기법(분류 모델)

- 머신 러닝
- 분류 모델
  - 의사결정트리
  - 랜덤 포리스트
  - SVM
- 성능 평가

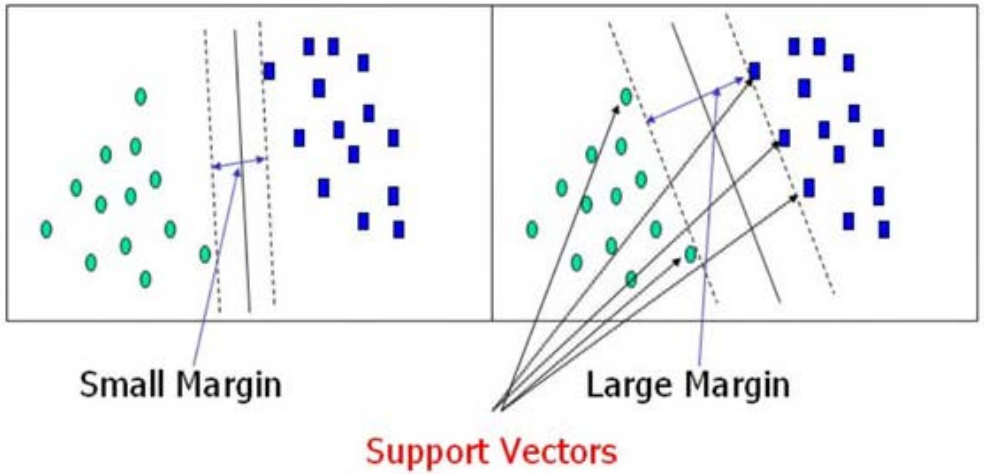
04 서포트 벡터 머신 (Support Vector Machines, SVM)

- 분류와 회귀를 수행 가능하며, 딥러닝이 나온 이후에도 여전히 활발히 사용되고 있는 머신 러닝 알고리즘이다.
- 분류를 위한 결정 경계 (Decision Boundary)를 정의하는 모델이다.
- 직관적으로 자료를 클래스별로 가장 잘 분리하는 결정 경계는 가장 가까운 훈련용 자료까지의 거리(이를 마진(margin)이라 함)가 가장 큰 경우이며, 마진이 가장 큰 결정 경계를 분류기(classifier)로 사용할 때, 새로운 자료에 대한 오류율이 가장 낮아진다.
- 즉, 최대 마진을 가지는 선형분류에 기초하며, 속성들 간의 의존성은 고려하지 않는 방법이다.
- 서포트 벡터들은 두 클래스 사이의 경계에 위치한 데이터 포인트들을 말한다. 이 데이터들이 결정경계를 지지(support)하고 있다고 말할 수 있으므로 서포트 벡터라고 부른다.

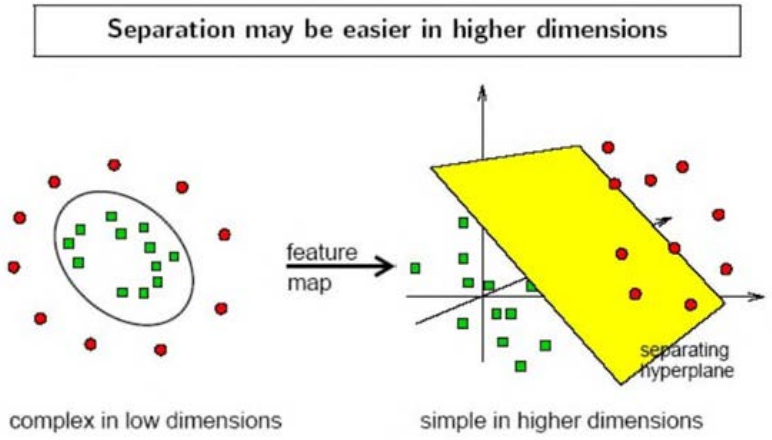


04 서포트 벡터 머신 (Support Vector Machines, SVM)

- 지도학습기법으로 비-중첩(non-overlapping) 분할을 제공하며 모든 특성(속성, features, attributes)을 활용하는 전역적(global) 분류 모형이다.
- 선형분류 뿐 아니라, 커널 트릭(kernel trick)이라 불리는 다차원 공간상으로의 맵핑(mapping) 기법을 사용하여 비선형분류도 효율적으로 수행한다
- 수학적으로 잘 정의되어 있으며, 복잡도를 조정할 수 있다.



Linear SVM classification



Nonlinear SVM classification

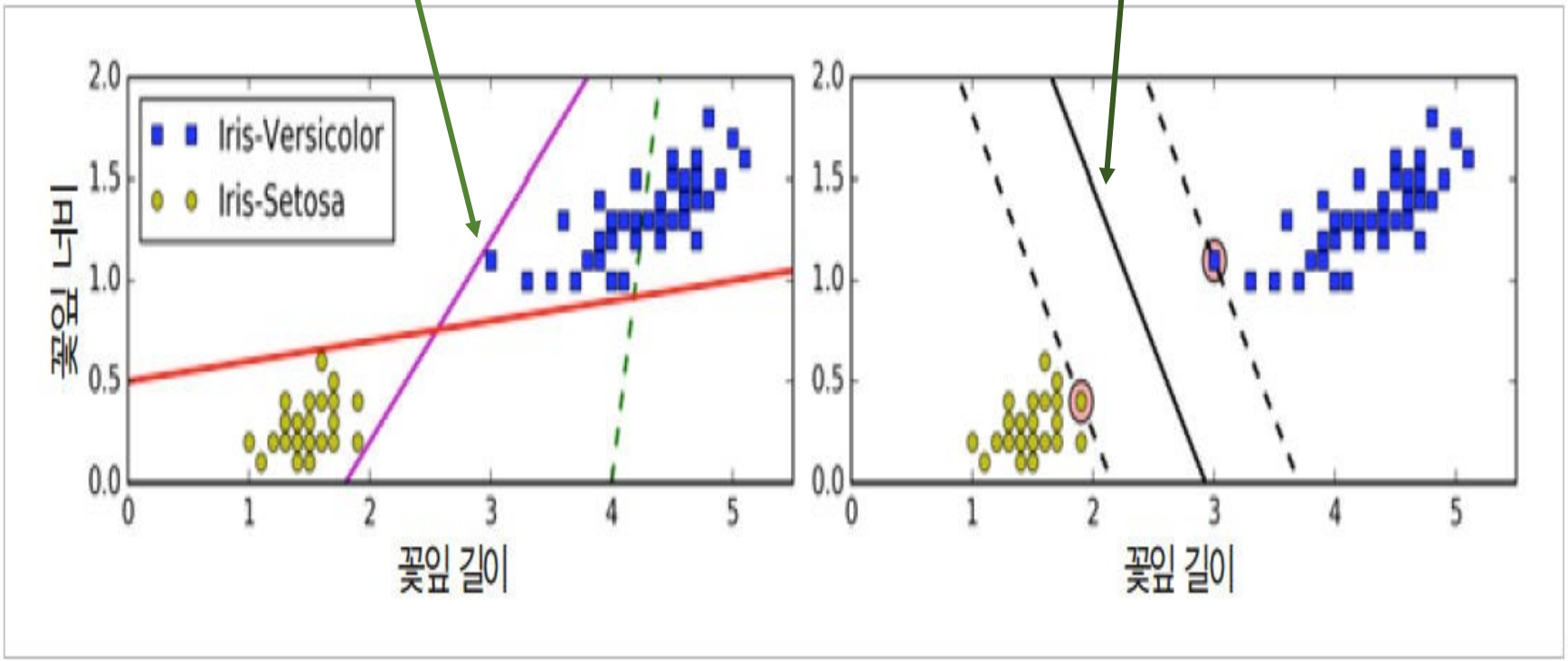
04-1 SVM의 기본 개념: 최대 마진 분류기

- 마진(Margin)은 결정경계와 서포트 벡터 사이의 거리를 의미한다.
- 마진을 최대화하도록 결정경계를 설정한다 (=서포트벡터를 선택한다)

➔ Maximal Margin Classifier

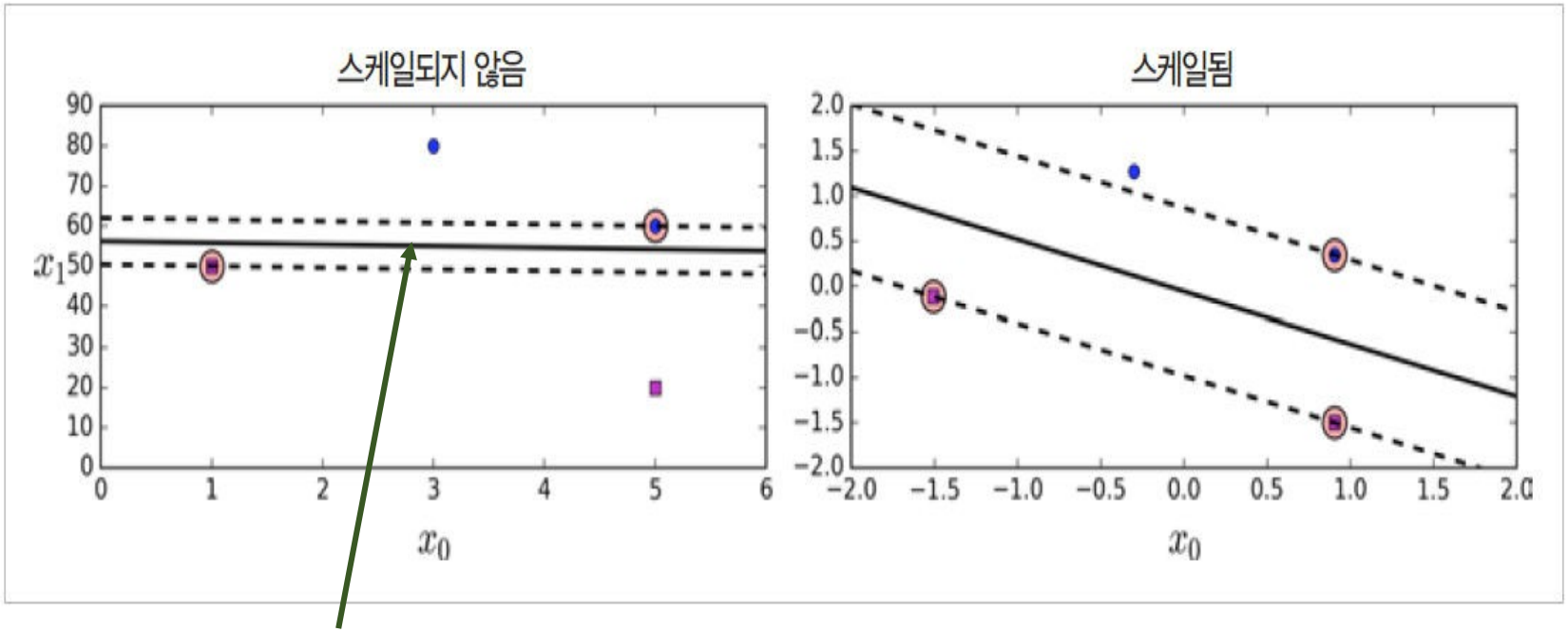
결정 경계가 샘플에 너무 가깝다

결정 경계가 가능한 샘플에서 멀리 떨어져 있다.
라지 마진(large margin) 분류라고도 부른다



04-1 SVM의 기본 개념: 스케일링

- SVM은 특성의 스케일에 영향을 많이 받음
  ➔ 특성을 정규화 한 후 SVM 을 적용하여야 함

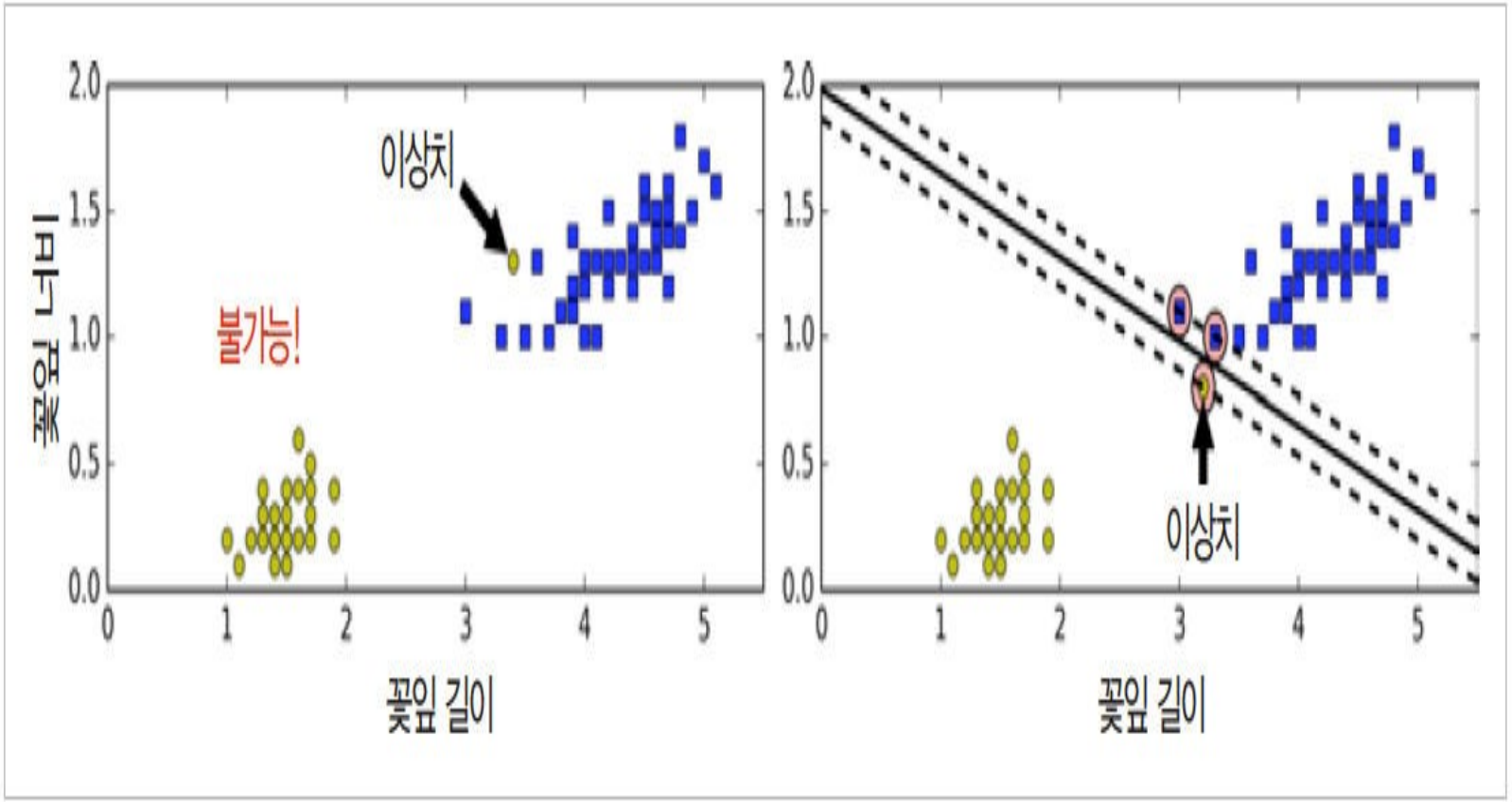


스케일이 큰 x1에 영향을 많이 받아 결정 경계가 수평에 가깝게 된다.

R의 svm() 함수 자체에서 스케일링(평균 0, 표준편차 1)을 수행하므로 별도의 스케일링을 할 필요가 없다. 단. 필요 없을 경우, scale 인자에 FALSE를 입력한다

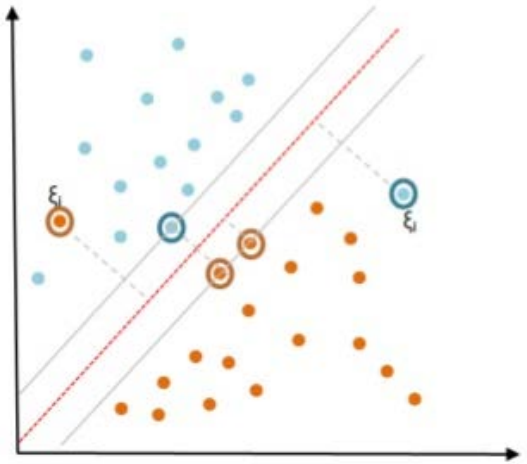
04-1 SVM의 기본 개념: 하드 마진 분류 (hard margin classification)

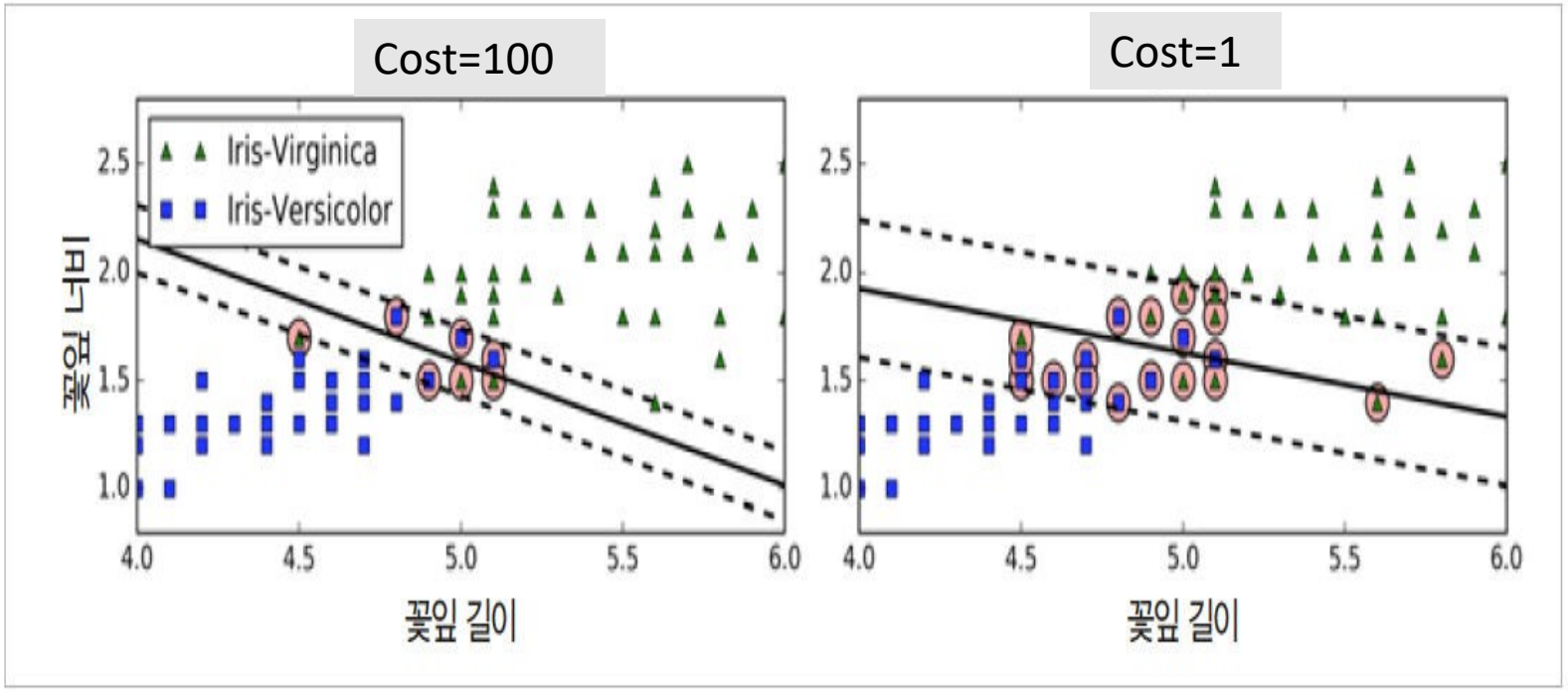
- 마진 안이나 밖에 기준과 맞지 않는 데이터를 절대 허용하지 않는 방식이다.
- 데이터가 선형적으로 구분될 수 있어야 한다.
- 이상치(outlier)에 민감하다
- 오버피팅의 문제가 발생할 수 있다.
- 몇 개의 노이즈로 인해 두 클래스를 구분하는 결정경계를 잘 못 구할 수도 있고, 경우에 따라서는 찾지 못하는 문제가 발생한다. 따라서 현실세계에서는 적용하기 힘들다.



04-1 SVM의 기본 개념: 소프트 마진 분류 (soft margin classification)

- 약간의 오분류를 허용하여 전체적인 성능을 높이는 방법 (margin violation)
- 마진 안에 데이터의 존재를 허용하며, 반대편 마진(영역)에 데이터가 존재하는 것조차도 허용함.
- 어느 정도의 과적합을 방지할 수 있음.
- 어느 정도의 오류를 허용해 줄 것인가: Cost 파라메터에 의하여 조정 가능함
- Cost가 큰 값을 가지면 마진의 폭이 좁아지고, Cost가 작은 값을 가지면 마진의 폭은 넓어진다.
- 즉, Cost가 커지면 모델 복잡도가 증가하고, Cost가 작아지면 복잡도가 감소한다.



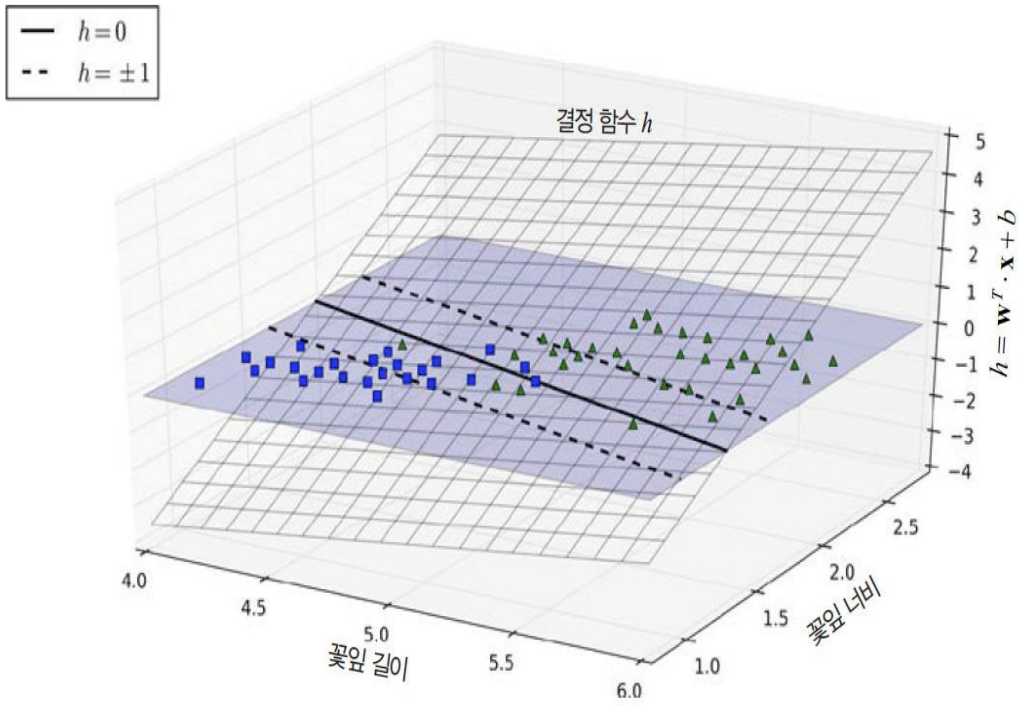


04-2 선형 SVM 모델

마진 오류가 전혀 없거나 (하드 마진), 어느 정도 오류를 가지면서 (소프트 마진) 최대한 마진을 크게 하는 w와 b 값을 찾는 것이다.

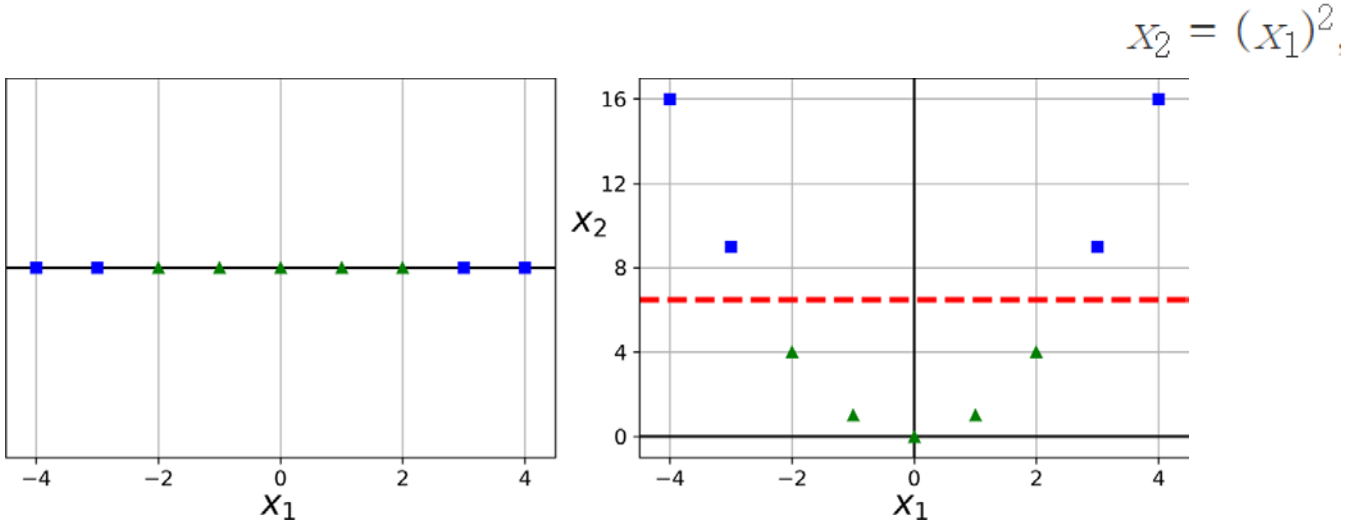
결정함수 : $\mathbf{w}^T \cdot \mathbf{x} + b = w_1 x_1 + \cdots + w_n x_n + b$

예측 : $\hat{y} = \begin{cases} 0 & \mathbf{w}^T \cdot \mathbf{x} + b < 0 \text{일 때} \\ 1 & \mathbf{w}^T \cdot \mathbf{x} + b \geq 0 \text{일 때} \end{cases}$



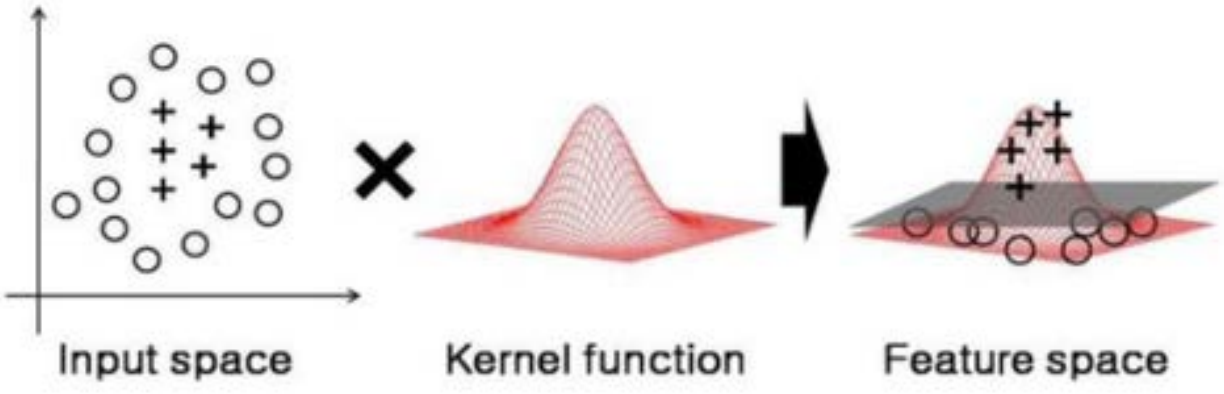
04-3 비선형 SVM

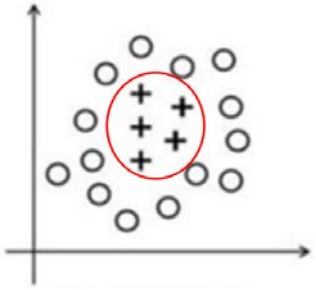
- SVM은 비선형 결정 경계(non-linear decision boundaries)를 가지는 문제도 해결할 수 있다.
- 핵심 아이디어는 원래의 D-차원 공간을 D'-차원(D' >D)으로 맵핑하여 그 점들이 선형적으로 분리 가능하도록 하는 것이다.
- PolynomialFeatures + LinearSVC: 다항 특성 수의 증가로 모델링 속도가 저하될 수 있다.



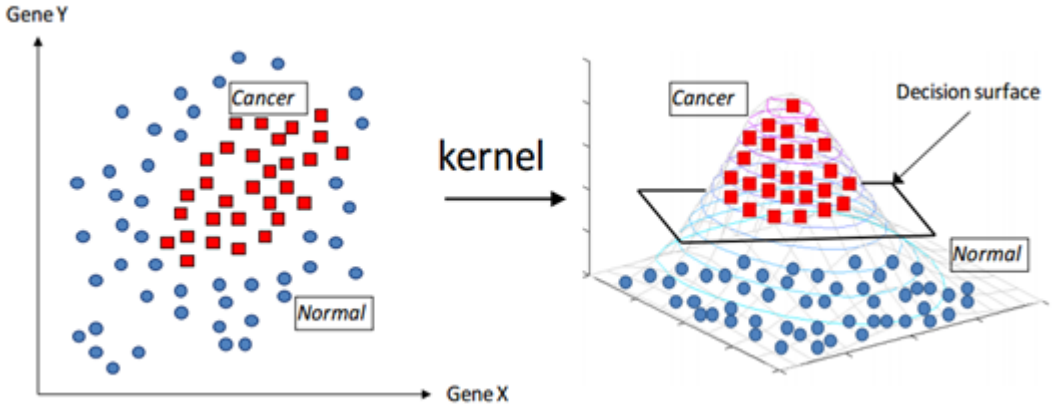
04-4 비선형 SVM – 커널 트릭 (Kernel Trick)

- 실제 다항 특성을 추가하지 않고 비슷한 효과를 만드는 수학적 트릭
- 특성을 직접 변환하는 대신 두 샘플사이의 유사도를 의미하는 커널을 정의/도입
- 다항커널과 가우시안 커널이 주로 사용됨.





가우시안 커널로 데이터를 분류한 예



04-4 비선형 SVM – 커널 트릭 (Kernel Trick)

- 다항커널(Polynomial Kernel): $K(x_i, x_j) = (x_i^T x_j + 1)^q$, q: 다항식의 차수

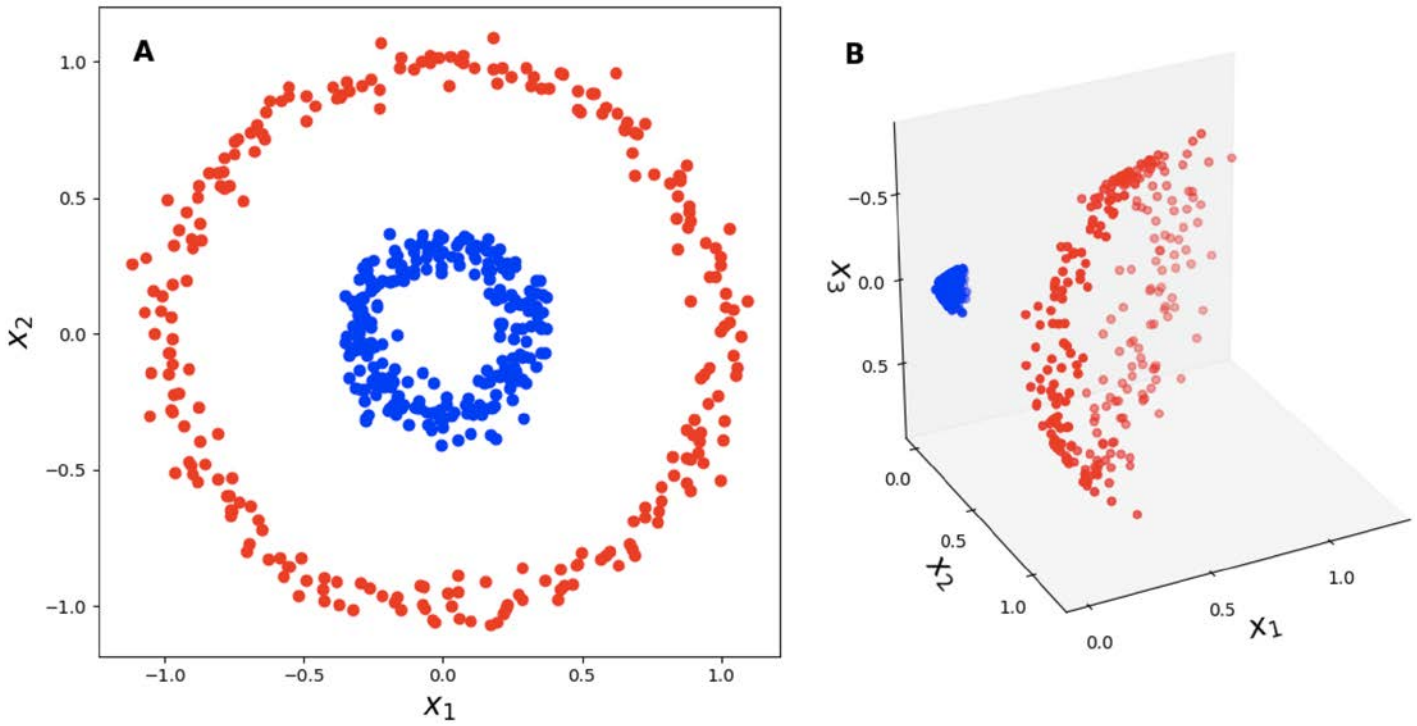


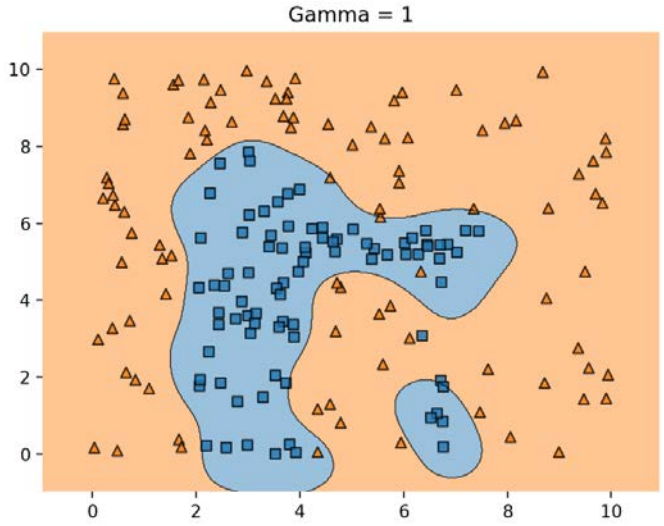
Figure 1: The "lifting trick". (a) A binary classification problem that is not linearly separable in $\mathbb{R}^2$. (b) A lifting of the data into $\mathbb{R}^3$ using a polynomial kernel, $\varphi([x_1 \quad x_2]) = [x_1^2 \quad x_2^2 \quad \sqrt{2}x_1x_2]$.

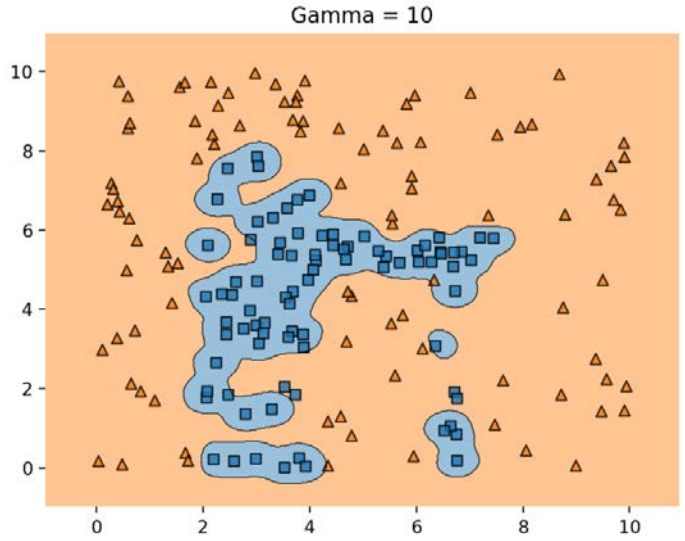
04-4 비선형 SVM – 커널 트릭 (Kernel Trick)

- 가우시안 커널 (Gaussian Kernel)
  - RBF(Radial Basis Function)을 사용. $K(x_i, x_j) = e^{-\gamma\|x_i - x_j\|^2}, \quad \gamma \geq 0$
  - 가장 많이 쓰는 커널
  - 하이퍼 파라메터

- 감마 (gamma): 하나의 데이터 샘플이 영향력을 행사하는 거리를 결정한다. 가우시안 함수의 표준편차와 관련되어 있어, 그 값이 클수록 작은 표준편차를 갖는다.

즉, 결정경계의 곡률을 조정한다고 할 수 있어, 감마 값을 높이면 학습 데이터에 많이 의존해서 결정 경계를 구불구불 굿게 된다. 반대로 감마 값을 낮추면 학습 데이터에 별로 의존하지 않고 결정 경계를 직선에 가깝게 긋게 된다. 감마 값이 커지면 모델 복잡도가 커진다.





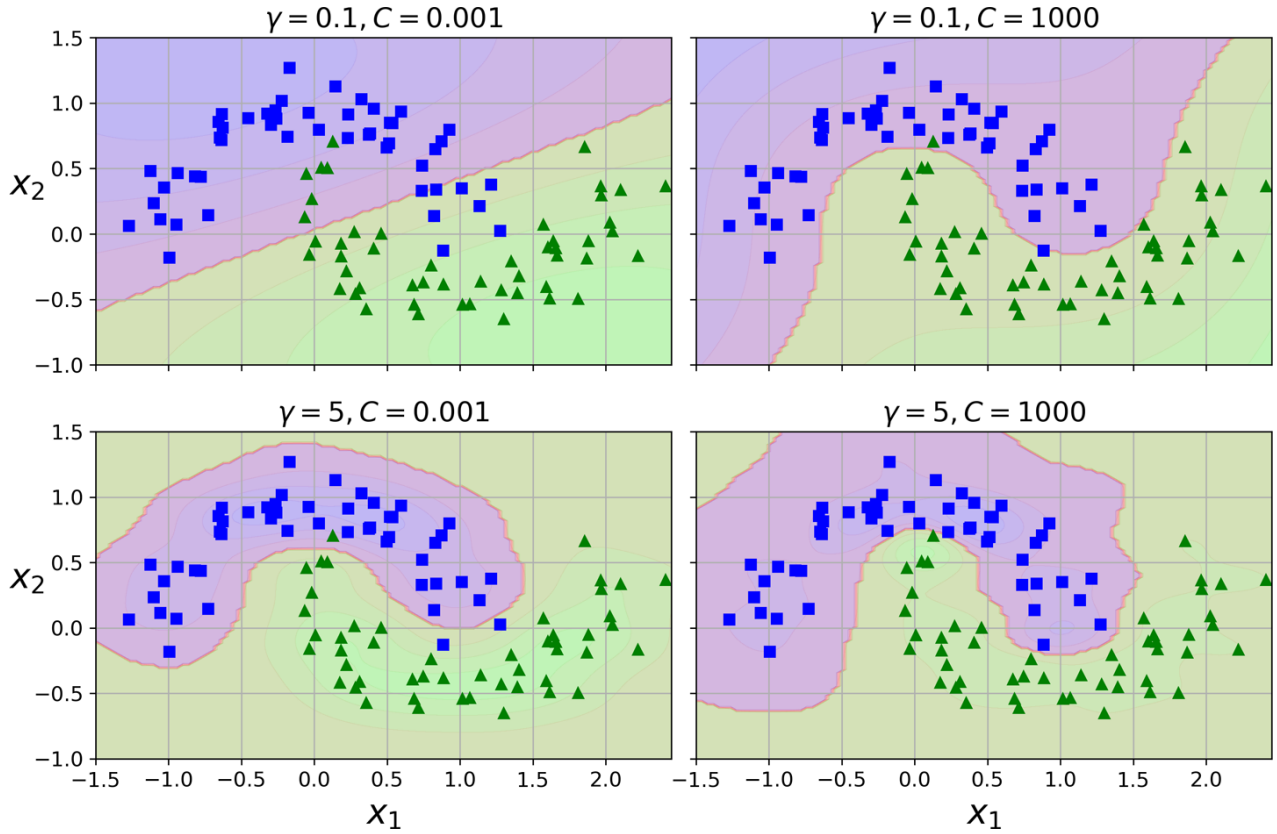
04-4 비선형 SVM – 커널 트릭 (Kernel Trick)

가우시안 RBF 커널

$$K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}, \qquad \gamma \geq 0$$

하이퍼 파라메터

- γ (gamma): 값이 커지면 모델 복잡도가 커진다.
- C (cost) : 값이 커지면 모델 복잡도가 커진다.



04-4 비선형 SVM – 커널 트릭 (Kernel Trick)

1. Linear Function

$$k(x_i, x_j) = x_i * x_j$$

2. Polynomial Function

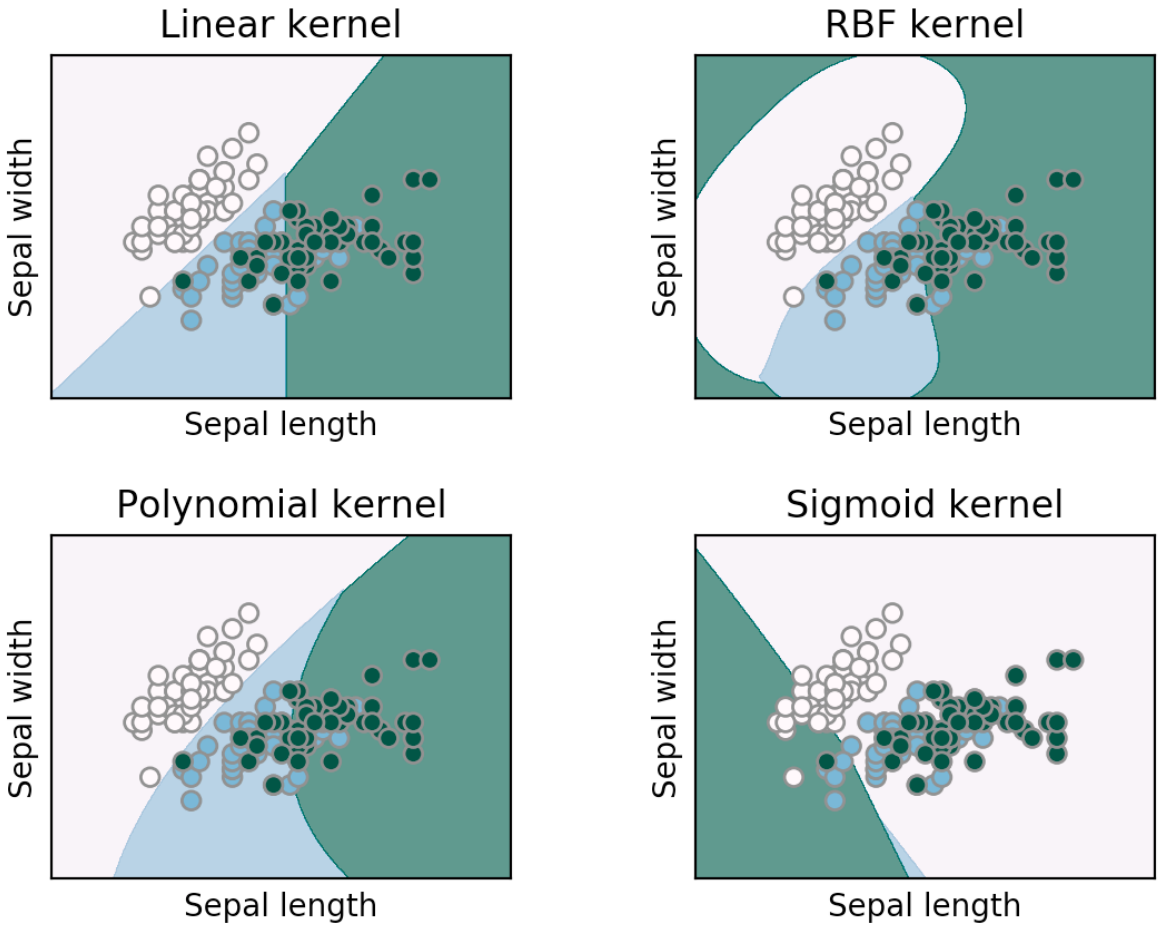
$$k(x_i, x_j) = (1 + x_i * x_j)^d$$

3. Radial Basis Function (RBF)

$$k(x_i, x_j) = exp(-\gamma \|x_i - x_j\|^2)$$

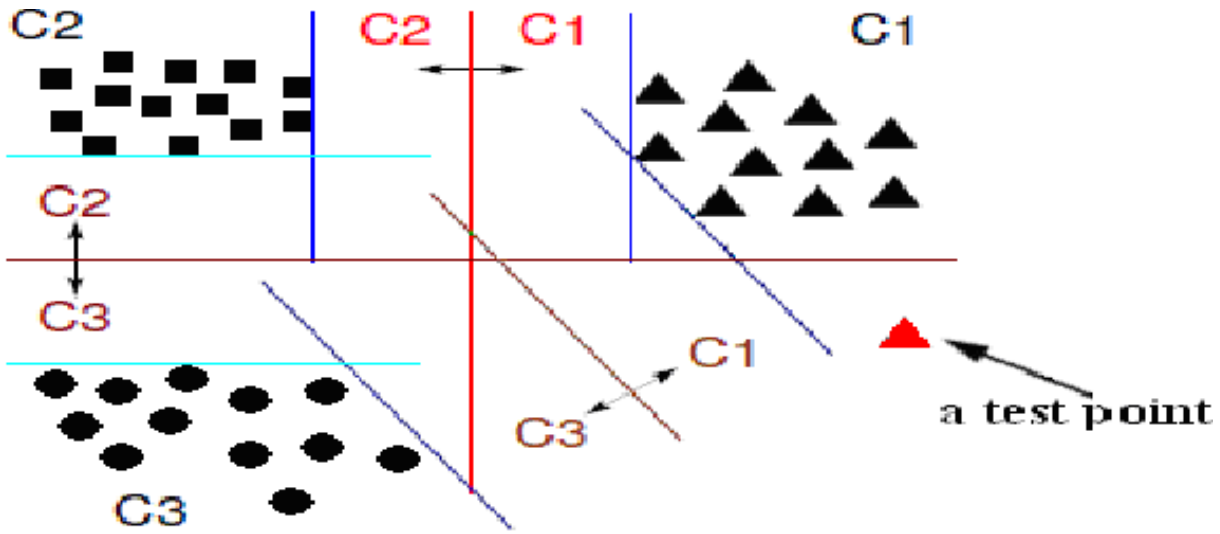
4. Sigmoid Function

$$k(x_i, x_j) = \tanh(\alpha x^T y + c)$$



04-5 SVM – 다중 분류 (Multiclass Classification)

- SVM은 기본적으로 이진분류(binary classification)를 위하여 고안되었으나, 다중 분류를 지원한다.
- SVM 이진 분류기를 여러 차례 반복적으로 사용하여 다중 분류기처럼 사용하는 방법이다.
- 여러 그룹을 마치 두 그룹인 경우처럼 분류를 한 뒤, voting mechanism 등을 사용하여 결과들을 결합한다.
- 일대일 SVM(one-versus-one, OVO SVM), 일대 다 SVM(one-versus-rest, OVR SVM) 방법 등을 사용할 수 있다.
- OVO SVM의 경우, n이 분류 클래스의 개수일 때, n*(n-1)/2 번 이진 분류기를 반복적으로 돌려 결과를 얻는다.



04-6 SVM – svm() 함수 이용

```
> library("e1071")
> data(iris)
> svm.e1071 <- svm(Species ~ . , data = iris,
                   type = "C-classification", kernel = "radial",
                   cost = 10, gamma = 0.1)
> summary(svm.e1071)
```

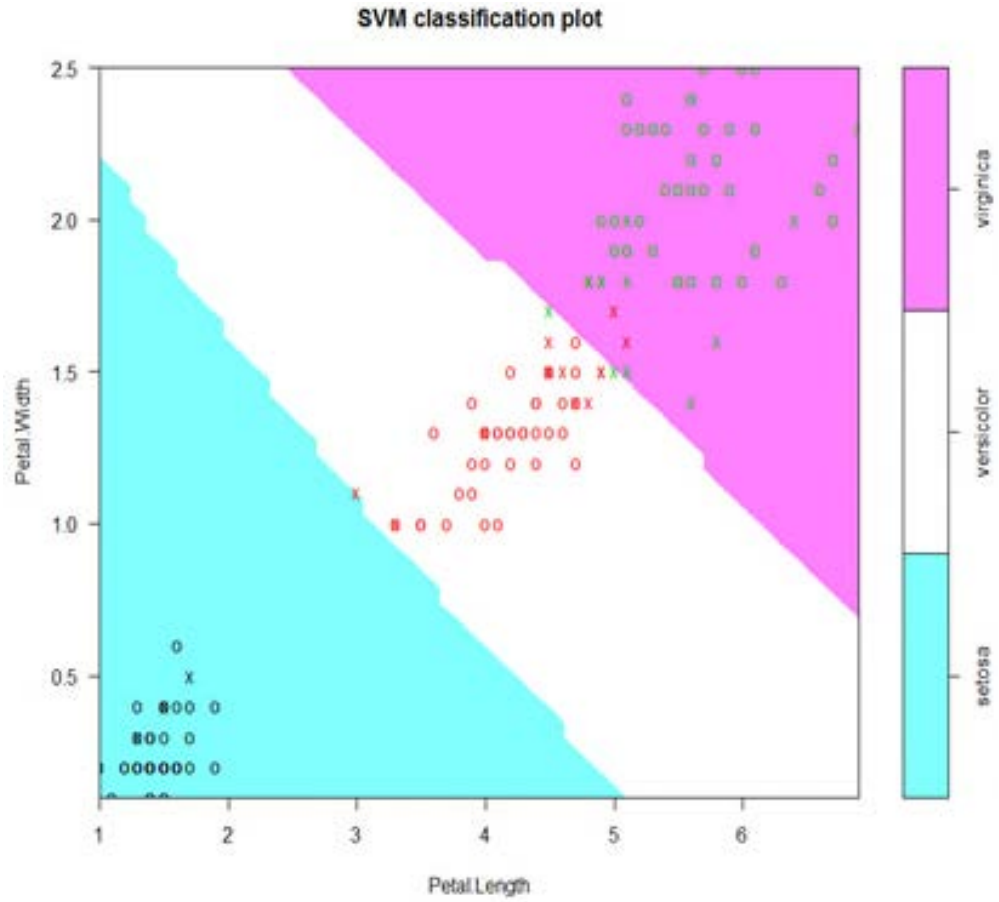
```
Call:
 svm(formula = Species ~ ., data = iris, type = "C-classification",
         kernel = "radial", cost = 10, gamma = 0.1)

 Parameters:
    SVM-Type: C-classification
  SVM-Kernel: radial
        cost: 10
       gamma: 0.1

 Number of Support Vectors: 32
   ( 3 16 13 )

 Number of Classes: 3

 Levels:
  setosa versicolor virginica
```

svm() 함수의 주요 옵션

- type : svm()의 수행 방법(분류, 회귀 또는 novelty detection)을 정한다.
- kernel : 훈련과 예측에 사용되는 커널로, "radial" 옵션은 가우시안 RBF를 의미한다.
- degree : 다항커널이 사용될 경우의 모수(차수)이다.
- gamma : 선형을 제외한 모든 커널에 요구되는 모수로, 디폴트는 1/(데이터 차원) 이다.
- coef0 : 다항 또는 시그모이드 커널에 요구되는 모수로, 디폴트는 0 이다.
- cost : 제약 위배의 비용으로, 디폴트는 1 이다.
- cross : k- 중첩 교차타당도에서 k값을 지정한다.

04-6 SVM – plot() 함수 이용 (시각화)

```
> plot(svm.e1071, iris, Petal.Width ~ Petal.Length,
       slice = list(Sepal.Width = 3, Sepal.Length = 4))
```



04-6 SVM – predict() 함수 이용 (예측)

```
> pred <- predict(svm.e1071, iris, decision.values = TRUE)
> (acc <- table(pred, iris$Species))
pred          setosa versicolor virginica
  setosa          50          0         0
  versicolor       0         47         0
  virginica        0          3        50
```

04-6 SVM – tune() 함수 이용 (hyperparameters tuning)

```
> tuned <- tune.svm(Species~., data = iris, gamma = 10^(-6:-1),
                    cost = 10^(1:2))
> # 6× 2 = 12개의 조합에서 모수조율이 이루어짐
> summary(tuned)

 Parameter tuning of ‘svm’:
 - sampling method: 10-fold cross validation
 - best parameters:
 gamma cost
  0.01  100
```

tune() 함수는 제공된 파라미터 영역에서 격자(grid) 탐색을 수행하여 hyperparameters를 조절(tune)할 수 있도록 지원한다.
이 함수는 최적의 파라미터를 제공해 주며, 동시에 여러 파라미터 값에 대한 검정의 자세한 결과를 제공해 준다.

```
 - best performance: 0.03333333
 - Detailed performance results:
    gamma cost      error dispersion
1   1e-06   10 0.78666667 0.07568616
2   1e-05   10 0.78666667 0.07568616
3   1e-04   10 0.63333333 0.15153535
4   1e-03   10 0.10666667 0.10976968
5   1e-02   10 0.04000000 0.04661373
6   1e-01   10 0.04000000 0.04661373
7   1e-06  100 0.78666667 0.07568616
8   1e-05  100 0.63333333 0.15153535
9   1e-04  100 0.10000000 0.10999439
10  1e-03  100 0.04000000 0.04661373
11  1e-02  100 0.03333333 0.04714045
12  1e-01  100 0.04666667 0.04499657
```

error: classification error for the 10-fold cross validation
dispersion: the standard deviation of the aggregated training results based on the training data.

05-1 성능 평가 – 교차 검증 (K-fold Cross validation)

- 모델 성능 평가의 목적:
  - 새로운 데이터에 대한 성능을 예측하기 위하여
  - 더 좋은 모델을 선택하기 위하여 (혹은 Hyperparameter tuning을 위하여)

- 교차 검증:
  - 통계학에서 모델을 평가하는 한 가지 방법이다.
  - 방법:
    1. 데이터를 K개로 쪼갠다.
    2. 하나는 검증 데이터, 나머지는 훈련 데이터로 사용해 성능을 구한다.
    3. 또 다른 부분을 검증 데이터, 나머지를 훈련 데이터로 사용해 성능 (정확률:accuracy) 을 구한다.
    4. K번 반복한다.
    5. K번의 성능 (정확률:accuracy) 의 평균을 구한다.

05-2 정확도/정밀도/재현률

- 정확률(accuracy)이 의미가 없는 상황
  - 예) 1000명당 1명꼴로 암환자라면 의사가 무턱대고 정상이라고 판정해도 정확률이 99.9% (오진률은 0.1%)인 명의가 됨
  - 부류 불균형인 상황에서는 다른 성능 척도를 사용해야 함
- 정밀도(precision)와 재현률(recall): 정상인과 환자, 정상품과 불량품, 승인과 거부처럼 2부류 분류 문제에서 모델의 성능을 보다 세밀하게 측정해주는 척도로 사용됨

- 두 부류를 긍정$_{positive}$과 부정$_{negative}$으로 구분하는 방법
  예) 의사의 진단: 목적은 환자를 가려내기 위함 → 환자를 긍정, 정상을 부정으로 봄
  예) 반도체 불량품 검사: 불량품이 긍정, 정상품이 부정
  예) 신용 카드 승인 시스템: 불승인이 긍정, 승인이 부정

- 예측의 네 가지 경우

  TP(True Posiive): True를 True로 잘 예측한 것 # 예: 환자를 환자로 분류

  TN(True Negativet): False를 False로 잘 예측한 것 # 예: 정상인을 정상인으로 분류

  FP(False Positive): False를 True로 잘 못 예측한 것 # 예: 정상인을 환자로 분류

  FN(False Negative): True를 False로 잘 못 예측한 것 # 예: 환자를 정상인으로 분류

05-3 혼동 행렬 (Confusion Matrix)

		True condition	
	Total population	**Condition positive**	**Condition negative**
Predicted Condition	**Predicted condition positive**	**True positive**	**False positive**
	Predicted condition negative	**False negative**	**True negative**

Confusion Matrix를 통해 측정할 수 있는 모델의 성능지표

- **Accuracy (정확도): (TP + TN) / Total**
  전체 클래스 중 True는 True로 False는 False로 잘 예측했는지?
  정확도는 타겟값이 불균형한 데이터에서는 적절한 평가지표가 되지 못한다.

- **Precision (정밀도): TP / (TP + FP)**
  모델이 True로 예측한 값들 중에서 정말로 예측한 값이 맞는지?
  정밀도는 양성 예측도라고도 부른다. 정밀도가 상대적으로 더 중요한 지표인 경우는 실제 Negative 음성인 데이터 예측을 Positive로 잘못 판단하게 되면 업무상 큰 불이익이 발생하는 경우이다.

- **Recall (재현율): TP / (TP + FN)**
  True 클래스 중 모델이 잘 예측한 클래스의 비율?
  재현율은 민감도(Sensitivity) 혹은 TPR(True Positive Rate)라고도 불립니다. 재현율이 상대적으로 더 중요한 지표인 경우는 실제 Positive 양성인 데이터 예측을 Negative로 잘못 판단하게 되면 업무상 큰 불이익이 발생하는 경우이다.

- **F1 Score = 2 X (정밀도 X 재현율) / (정밀도 + 재현율)**
  정밀도와 재현율은 상호 보완할 수 있는 수준에서 적용되어야 한다. 정밀도와 재현율을 결합한 지표를 F1 Score라고 하며, F1 Score는 정밀도와 재현율이 어느 한쪽으로 치우치지 않는 수치를 나타낼 때 상대적으로 높은 값을 나타낸다.

05-3 혼동 행렬 (Confusion Matrix)

		True condition	
	Total population	**Condition positive**	**Condition negative**
Predicted Condition	**Predicted condition positive**	**True positive**	**False positive**
	Predicted condition negative	**False negative**	**True negative**

- 정보검색에서 자주 사용하는 정밀도$_{\text{precision}}$와 재현률$_{\text{recall}}$

$$\text{정밀도} = \frac{TP}{TP + FP}$$

$$\text{재현률} = \frac{TP}{TP + FN}$$

- 의료 분야에서 주로 사용하는 특이도 (Specificity)와 민감도(Sensitivity)

$$\text{특이도} = \frac{FP}{FP + TN}$$

$$\text{민감도} = \frac{TP}{TP + FN}$$

05-4 정밀도와 재현률 또는 특이도와 민감도

암 판정 예: 정밀 검사를 통한 최종 암 판정을 그라운드 트루스, 의사의 초진을 모델의 예측으로 간주

	1	2	3	4	5	6	7	8	9	10
최종 판정(그라운드 트루스)	N	N	P	N	P	P	N	N	N	P
초진(모델의 예측)	N	P	P	N	N	P	N	N	P	P
	TN	FP	TP	TN	FN	TP	TN	TN	FP	TP

$$\text{정밀도} = \frac{TP}{TP+FP}$$

$$\text{재현률} = \frac{TP}{TP+FN}$$

$$\text{특이도} = \frac{FP}{FP+TN}$$

$$\text{민감도} = \frac{TP}{TP+FN}$$

혼동 행렬

		그라운드 트루스	
		긍정	부정
예측	긍정	TP=3	FP=2
	부정	FN=1	TN=4

$$\text{정밀도} = \frac{3}{3+2} = 0.6, \quad \text{재현률} = \frac{3}{3+1} = 0.75$$

$$\text{특이도} = \frac{2}{2+4} = 0.333, \quad \text{민감도} = \frac{3}{3+1} = 0.75$$

요약

- **Support Vector Machines**
  - ✓선형/비선형 모델
  - ✓svm() 함수를 이용한 모델 생성/예측
  - ✓결과 분석
- **성능 평가**
  - ✓교차 검증
  - ✓혼동 행렬

- 기말 프로젝트:
  - 주어진 데이터에 대한 분류 모델 생성 및 성능 평가
    - ✓Logistic Regression
    - ✓Decision Trees
    - ✓Random Forests
    - ✓Support Vector Machines
  - 실습 영상에 프로젝트 관련 자세한 설명 있음
  - 제출 기간: 12월 16일(수요일)까지 (기일 엄수)
- 15주차 수업: 보충 수업 없음 (14주차 금주가 마지막 수업임)
- 기말시험: 기말 프로젝트로 대체함.

한학기동안 수고 많았어요.