R을 활용한 머신러닝

7장

SVM

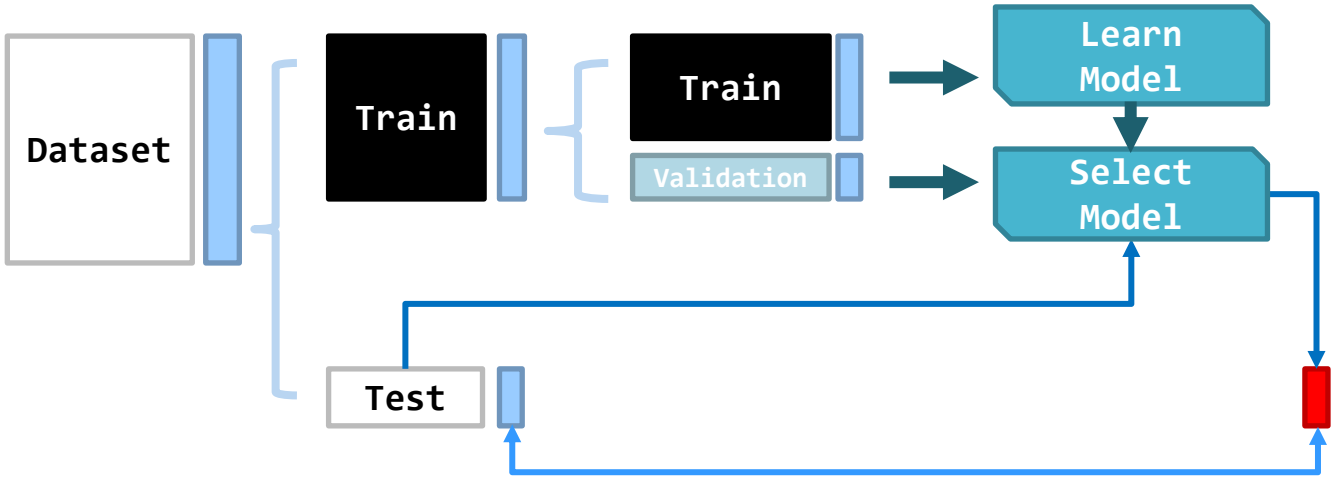
서포트벡터머신

Hyunseok Shin

Bio Information Technology Lab.

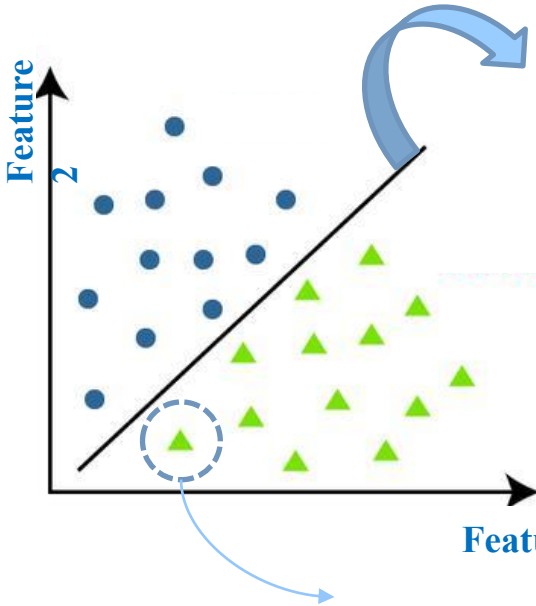
복습

- 모델 개발 과정


Dataset
Train
Train
Validation
Learn Model
Select Model
Test

복습

- 관련 용어
  - 모델(model) = 관계에 대한 가정
  - 학습(Learning) = 가정한 관계를 구체적으로 찾음
  - 분류(Classification) = 어떤 레이블을 가져야 하는지 결정
  - 특징(Feature)과 레이블(Label)



모델 = 하나의 직선이 클래스를 구분 = 관계

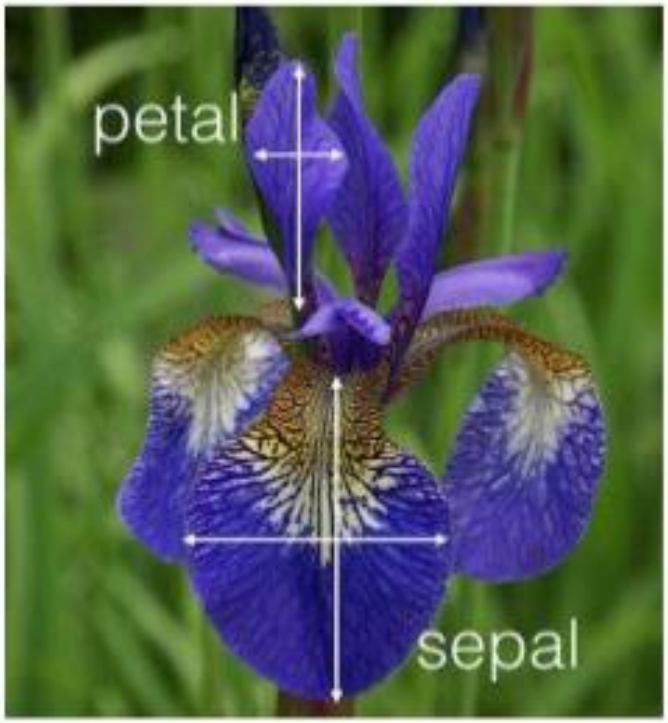
$\Leftrightarrow \quad y = f(x)$

새 데이터에 대해 예측(prediction) 가능

단, 경험 안에서 예측 가능(데이터 = 경험)

복습

- 관련 용어
  - 모델(model) = 관계에 대한 가정
  - 학습(Learning) = 가정한 관계를 구체적으로 찾음
  - 분류(Classification) = 어떤 레이블을 가져야 하는지 결정
  - 특징(Feature)과 레이블(Label)



Training / test data

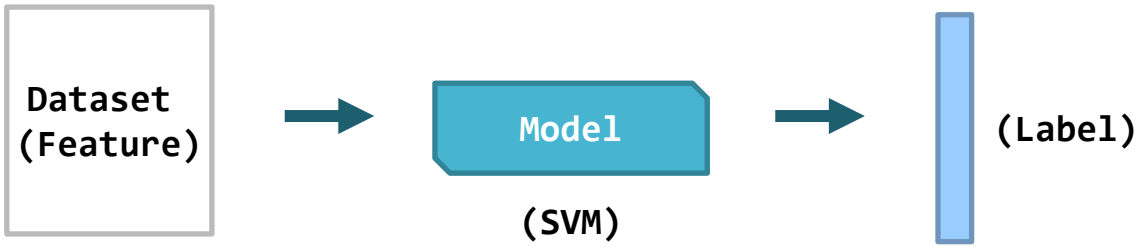
Features				Labels
Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	Iris setosa
4.9	3.0	1.4	0.2	Iris setosa
7.0	3.2	4.7	1.4	Iris versicolor
6.4	3.2	4.5	1.5	Iris versicolor
6.3	3.3	6.0	2.5	Iris virginica
5.8	3.3	6.0	2.5	Iris virginica

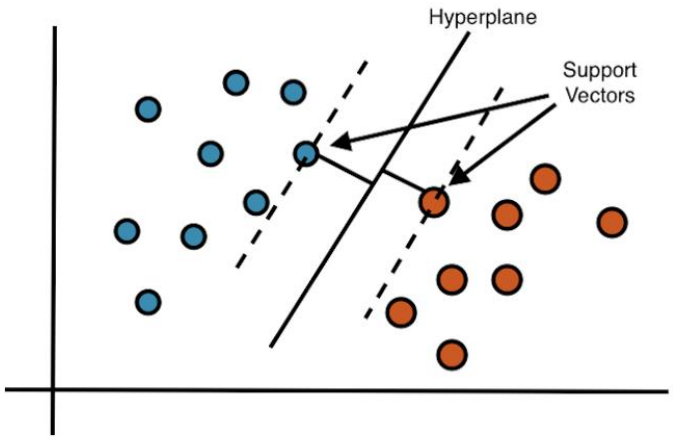
SVM

Part 1. SVM의 이해

SVM

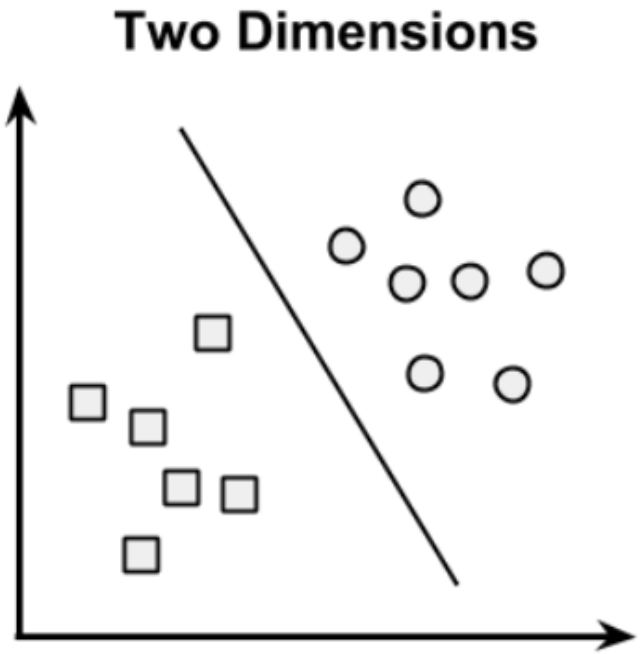
- Support Vector Machine
  - Support Vector
  - Hyperplane
  - Classification

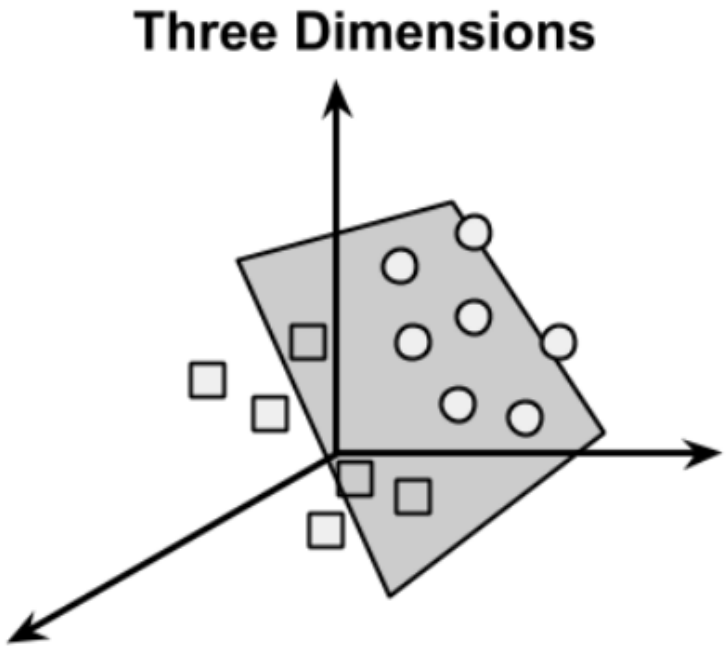




Classification with Hyperplane

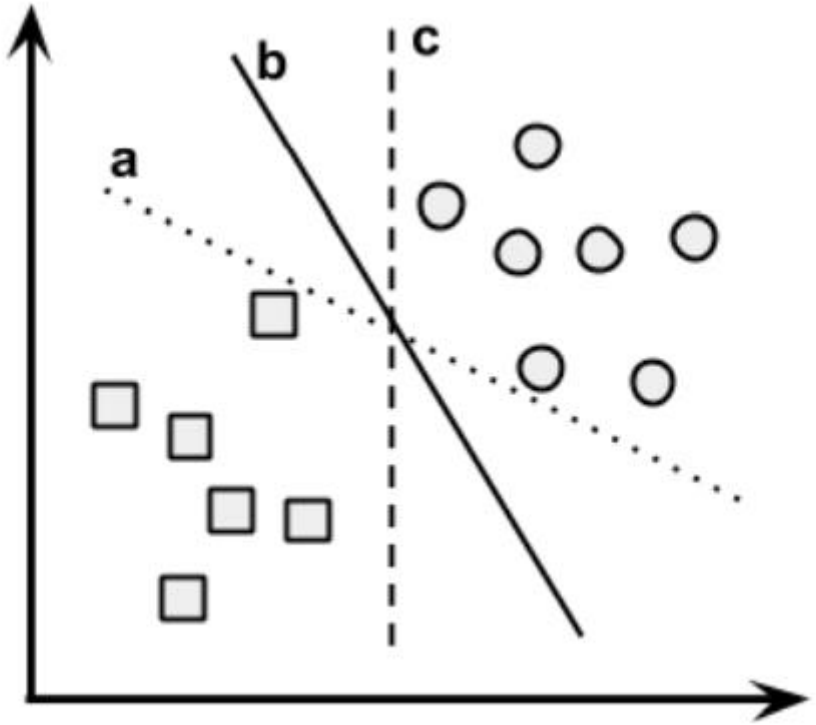
- Linearly separable





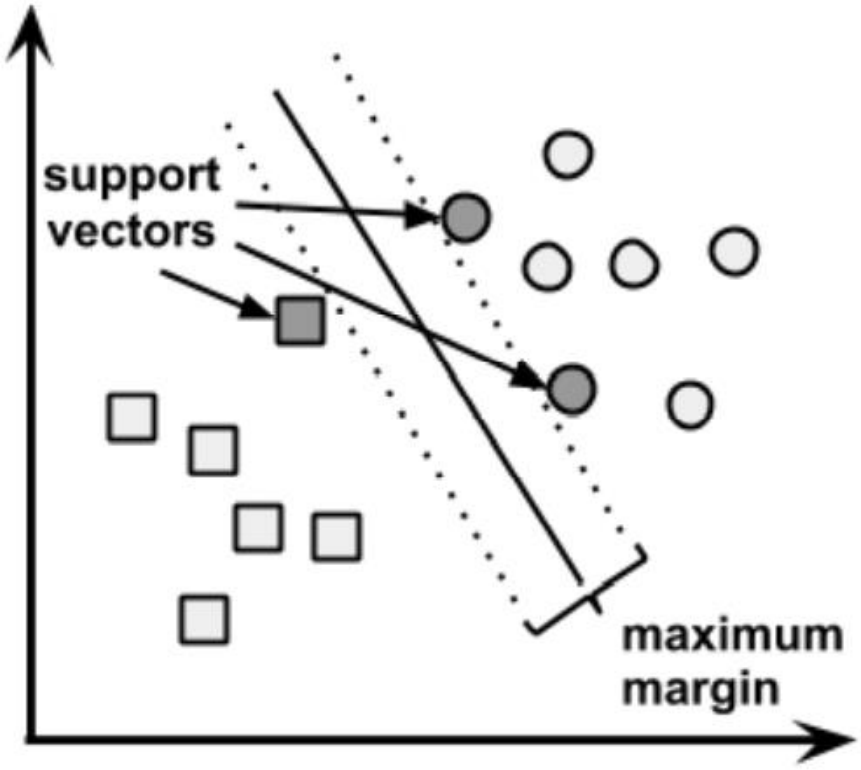
Classification with Hyperplane

- Three such possibilities are labeled a, b, and c. How does the algorithm choose?

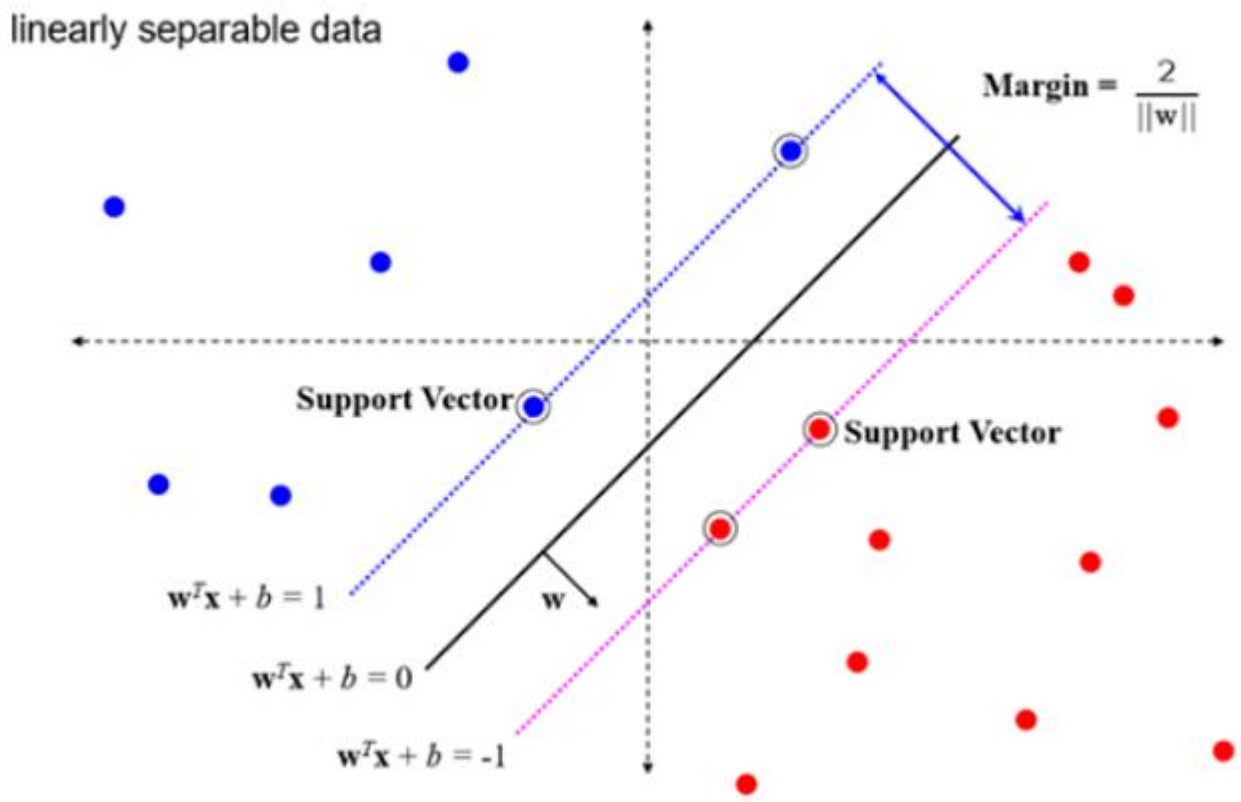


Classification with Hyperplane

- Three such possibilities are labeled a, b, and c. How does the algorithm choose?



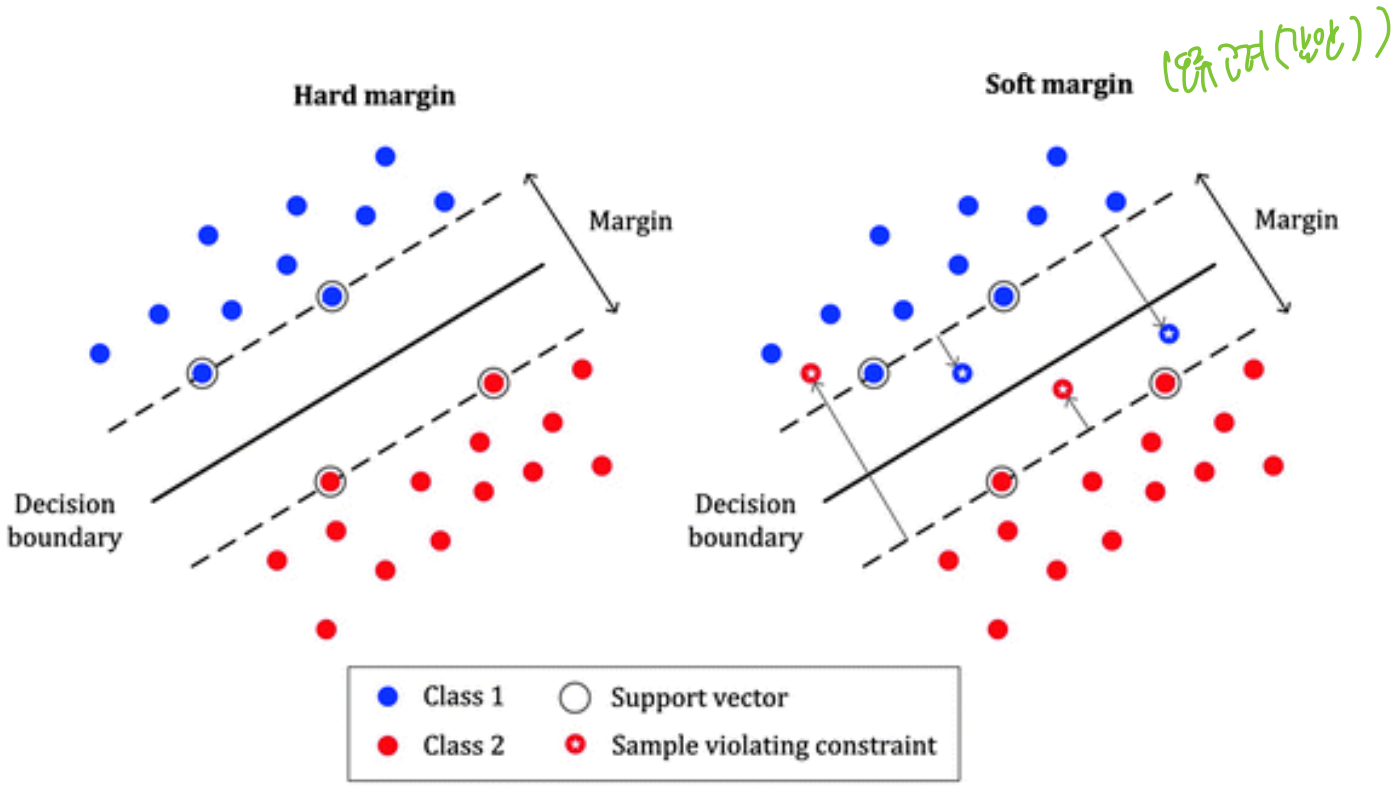
Classification with Hyperplane



Part 2. 데이터가 선형적으로 분리 가능하지 않을 때

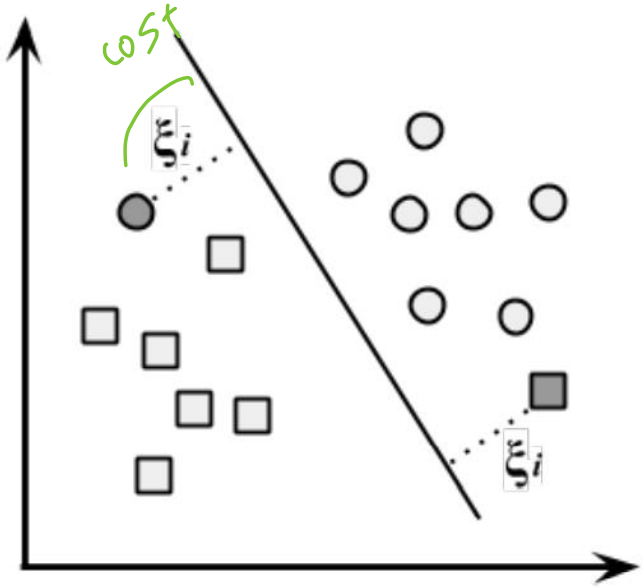
Soft Margin

- 최대 마진 vs 총비용 최소화



Soft Margin

- 총비용 최소화(vs 최대 마진)
  - Slack variable
  - 파라미터 C 의 역할은? C가 커질 때와 작아질 때의 의미는?



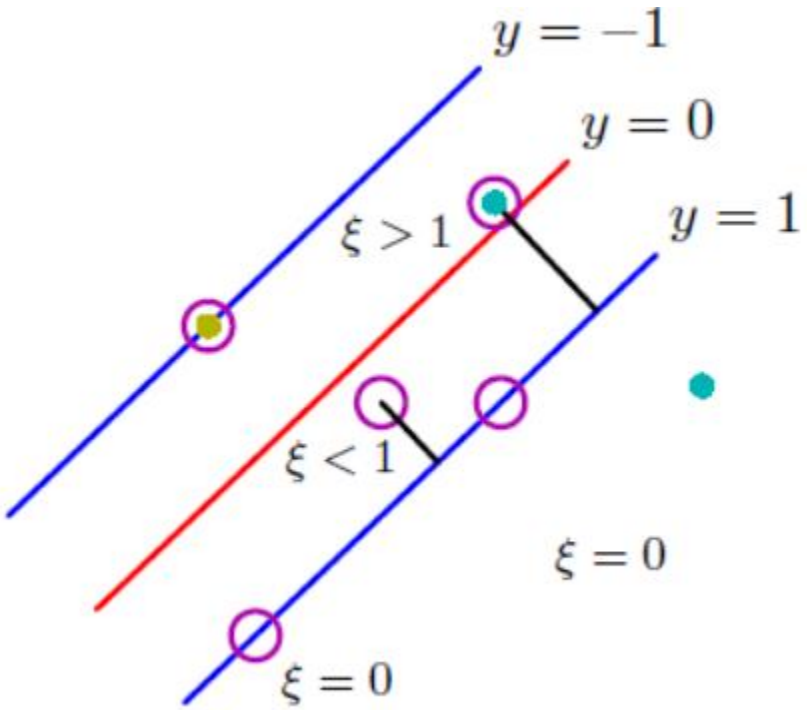
cost ↘

$$\min \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$

$$s.t.\ y_i\left(\vec{w}\cdot\vec{x}_i - b\right) \geq 1-\xi_i, \forall\vec{x}_i, \xi_i \geq 0$$

Soft Margin

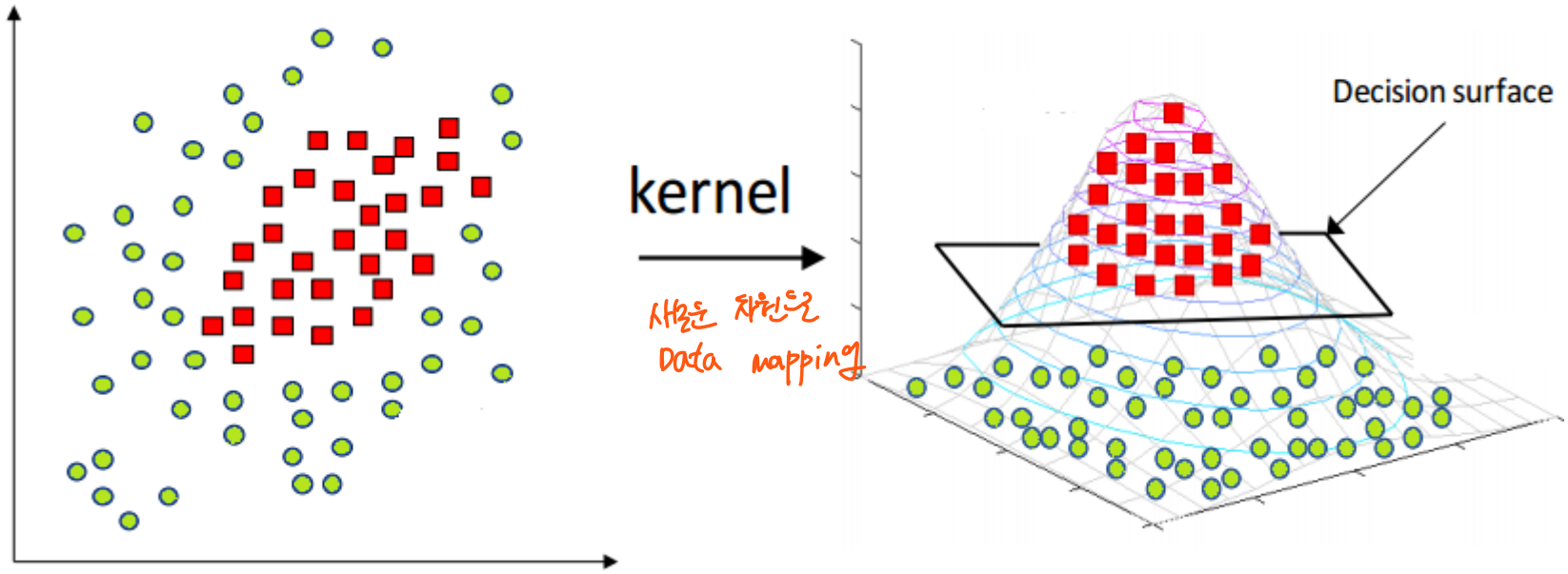
- 최대 마진 vs 총비용 최소화
  - 비용 파라미터 C : 초평면의 잘못 분류된 점에 대한 패널티 조정
  - ξ > 1, ξ < 1, ξ = 0 일때 각각의 의미는?



Part 3. 비선형 공간을 위한 커널의 사용

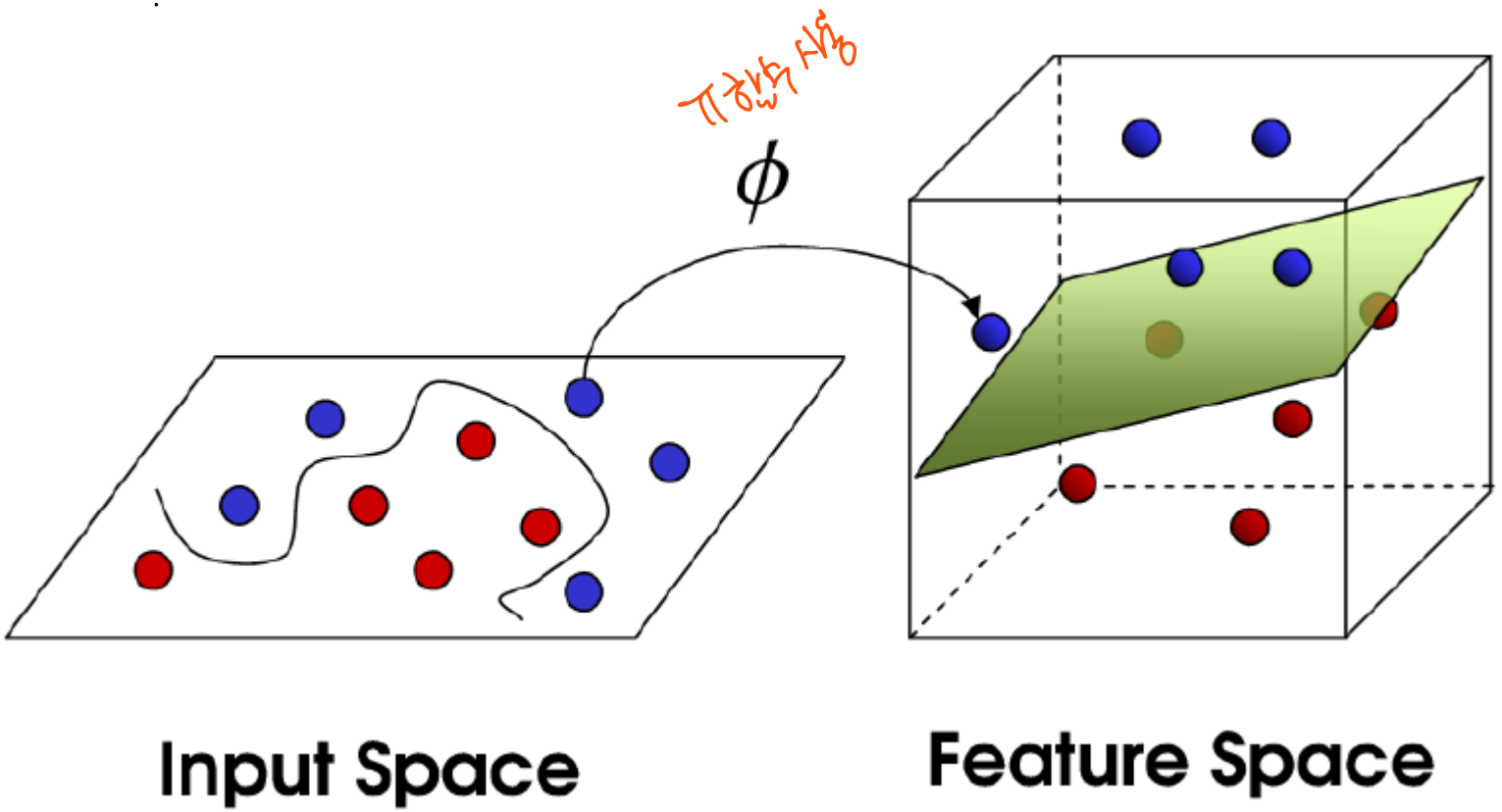
Kernel

- 고차원 공간으로 매핑
  - 새로운 차원을 추가



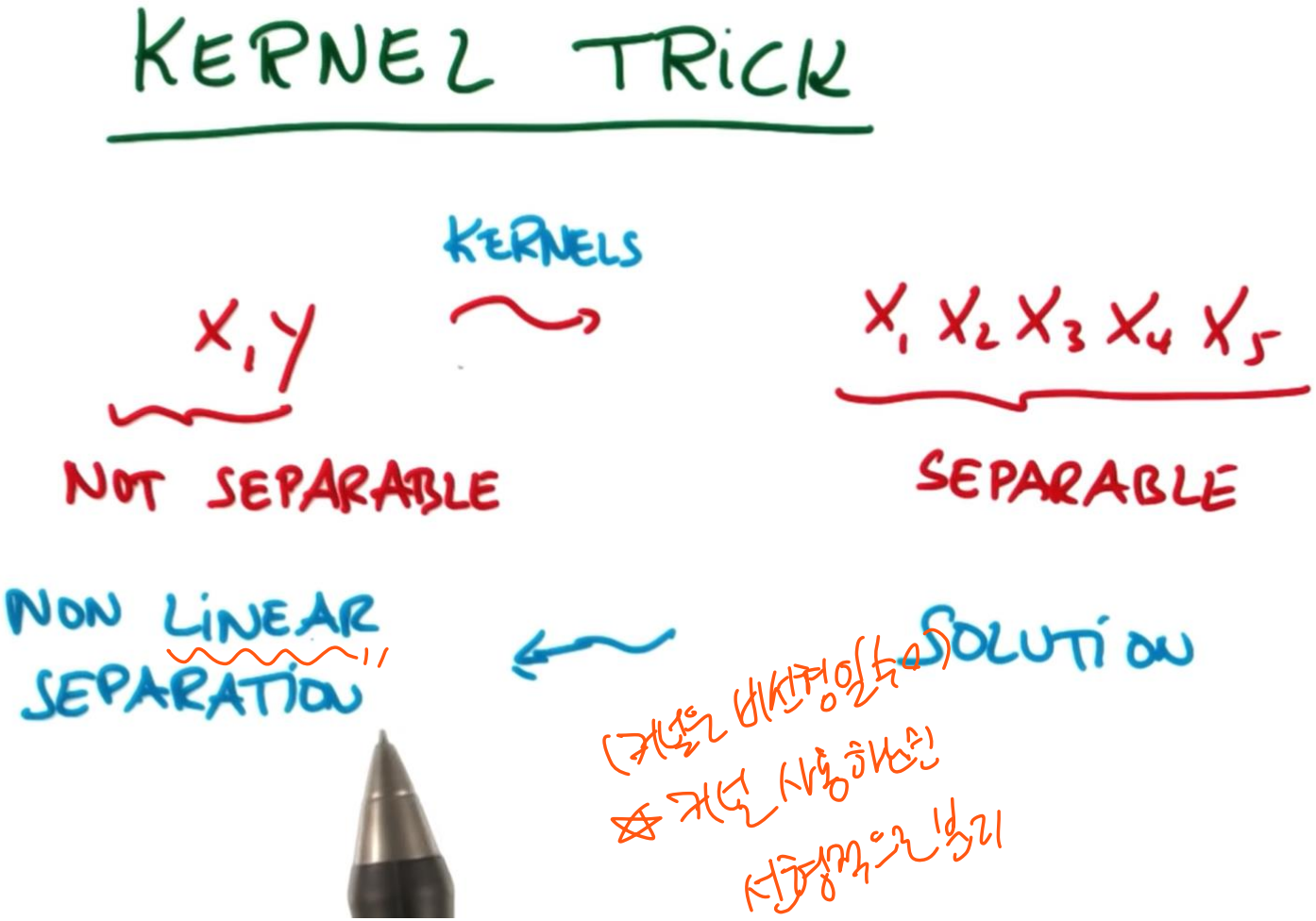
Kernel

- Mapping function



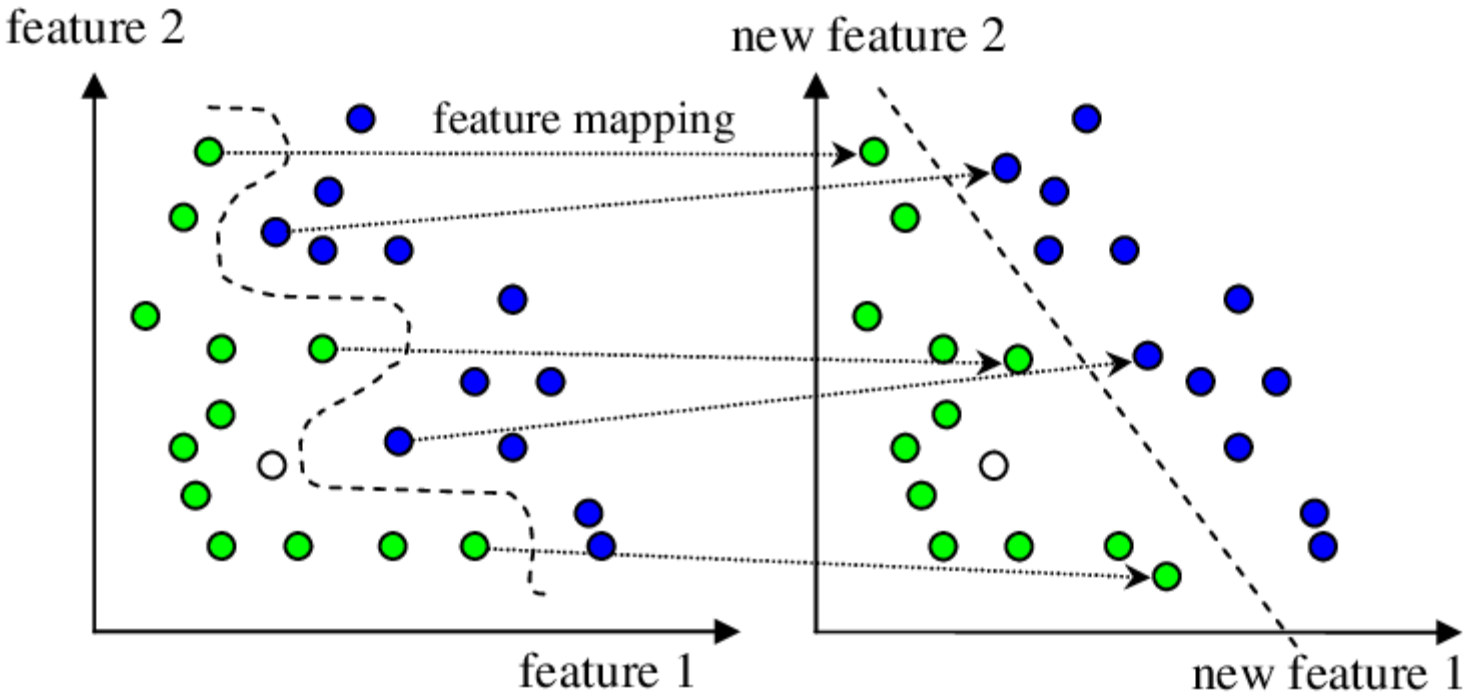
Kernel

- Kernel Trick



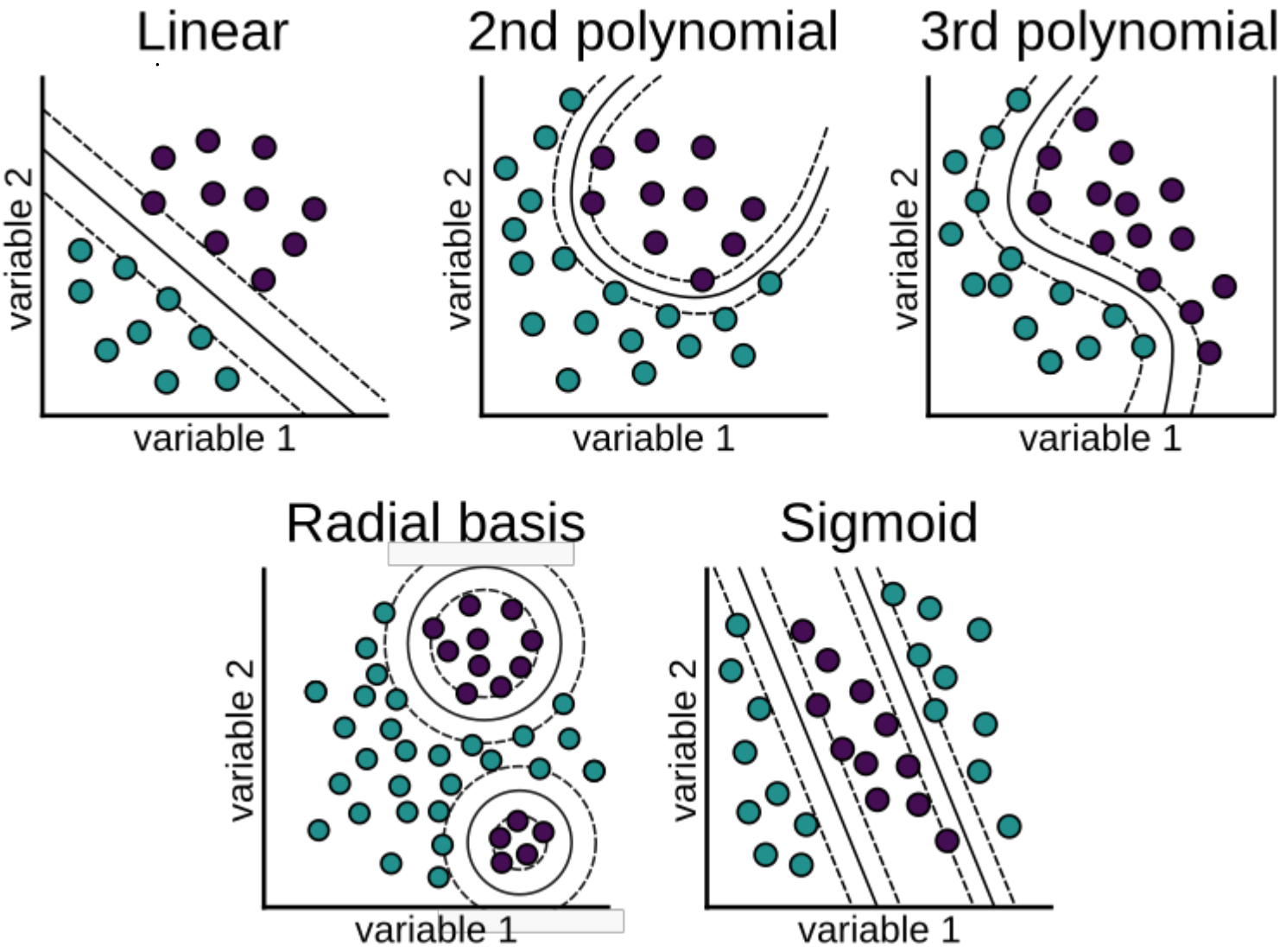
Kernel

- Kernel Trick



Kernel

- Various Kernels



Part 4. Code

R Packages

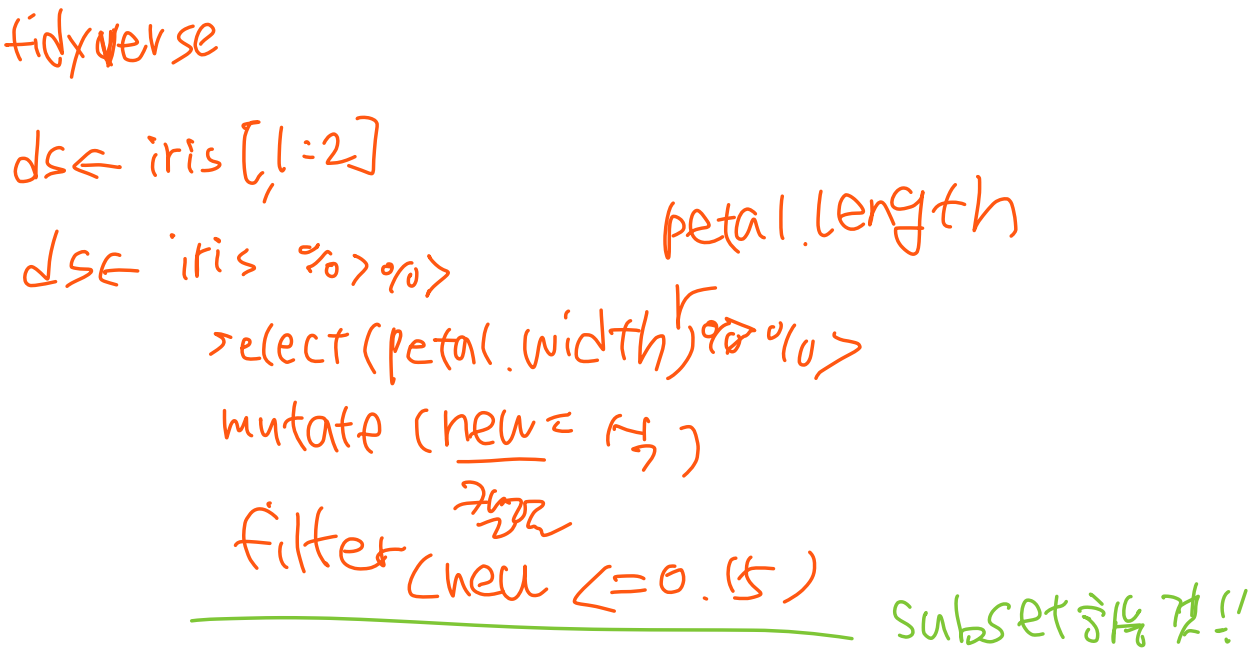
- 'kernlab' Package
  - https://cran.r-project.org/web/packages/kernlab/kernlab.pdf

```
library(kernlab)

m <- ksvm(target ~ predictor,
          data=mydata,
          kernel = 'rbfdot',
          C=1,
          scale = TRUE)

p <- predict(m, test, type = 'response')

acc <- mean(p == test_class)
```

- 'e1071' Package
  - https://cran.r-project.org/web/packages/e1071/e1071.pdf

#Homework

- iris 데이터를 이용하여 품종 예측 모델을 만드시오. 분류모델은 SVM을 사용하고 커널(kernel)과 비용(C) 파라미터를 바꿔가며 가장 좋은 정확도를 보이는 모델을 찾으시오. 또한 kernlab과 e1701 패키지를 각각 사용하여 결과를 비교해 보시오.
  - 훈련 데이터와 테스트 데이터 비율은 7:3
  - set.seed(100)



이번 시간 학습한 개념 정리

- SVM
- Linearly classification vs. Non-linearly classification
- Hard vs. Ṡoft margin
- Kernel

오늘도 한걸음

수고했습니다