
AN
INTRODUCTION
TO
MACHINE
LEARNING
WITH R
DAY 7

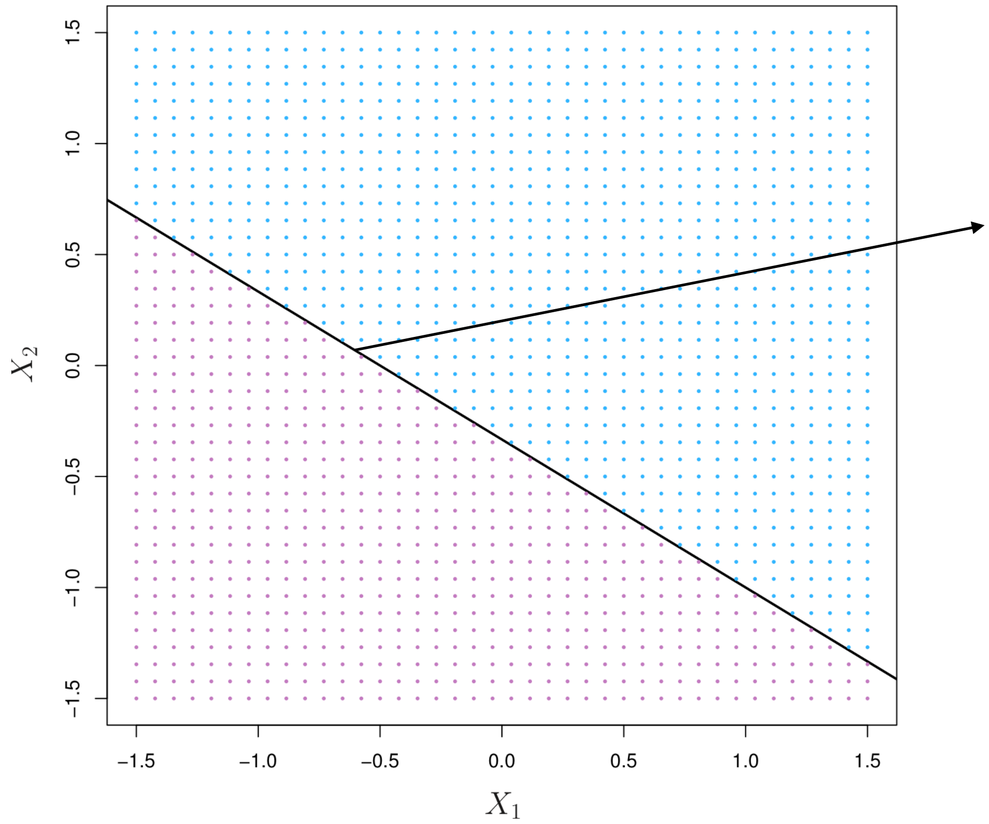
DAY 7

Support Vector Machine

Support Vector Machine

1962
Maximal Margin Classifier

Hyperplane



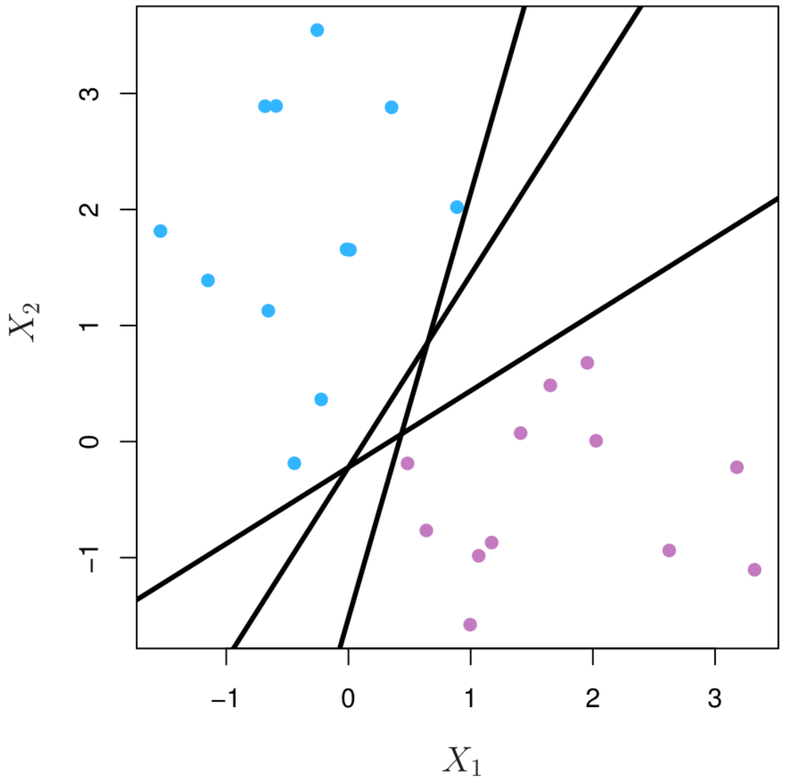
초평면(Hyperplane)
원점(Origin)이 어디인지 알 수 없는 평면
2차원 공간의 초평면은 선이다.

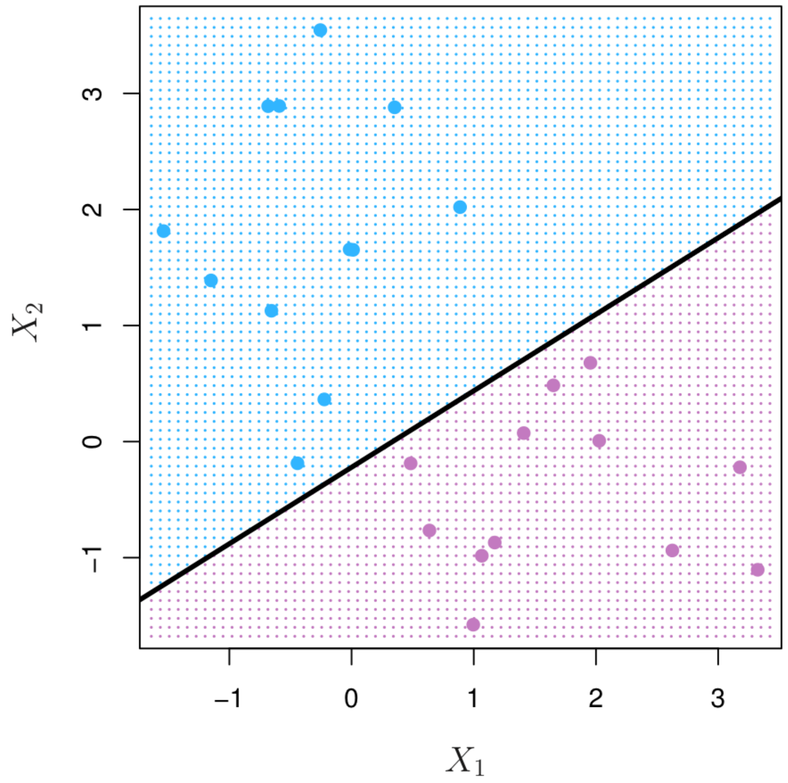
이 초평면은 기존의 공간을 반으로 나눈다.

Support Vector Machine

1962
Maximal Margin Classifier

Separation Hyperplane

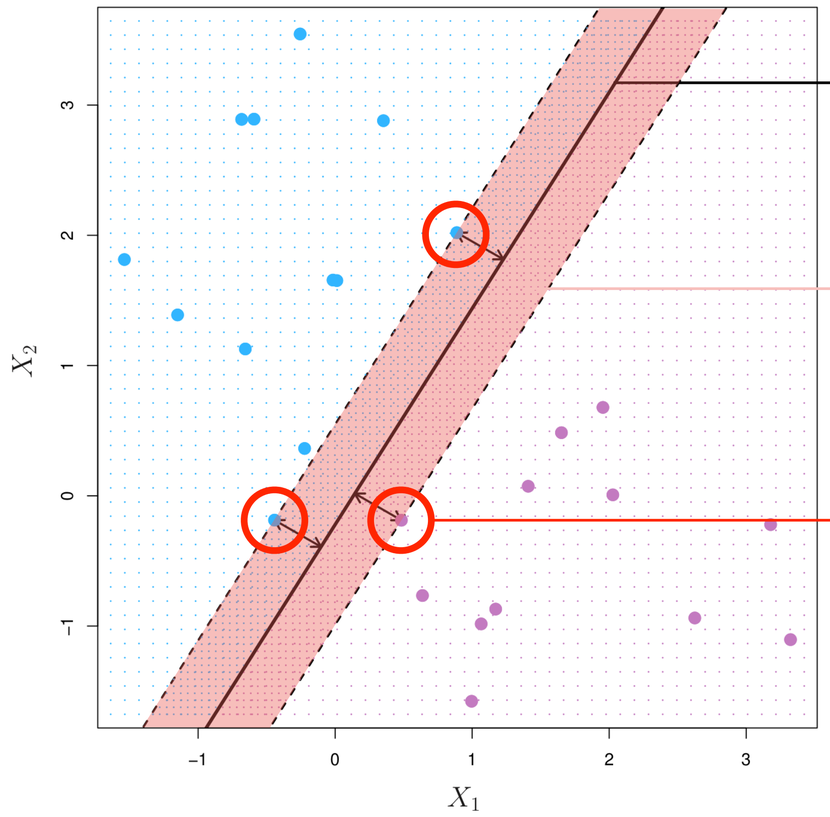




Support Vector Machine

1962
Maximal Margin Classifier

Maximal Margin Classifier



Maximal Margin Hyperplane
서포트 벡터(Support Vector)들로부터
가장 멀리 떨어진 분리 초평면

Margin
서포트 벡터와 최대 마진 초평면까지의 가장 짧은 거리

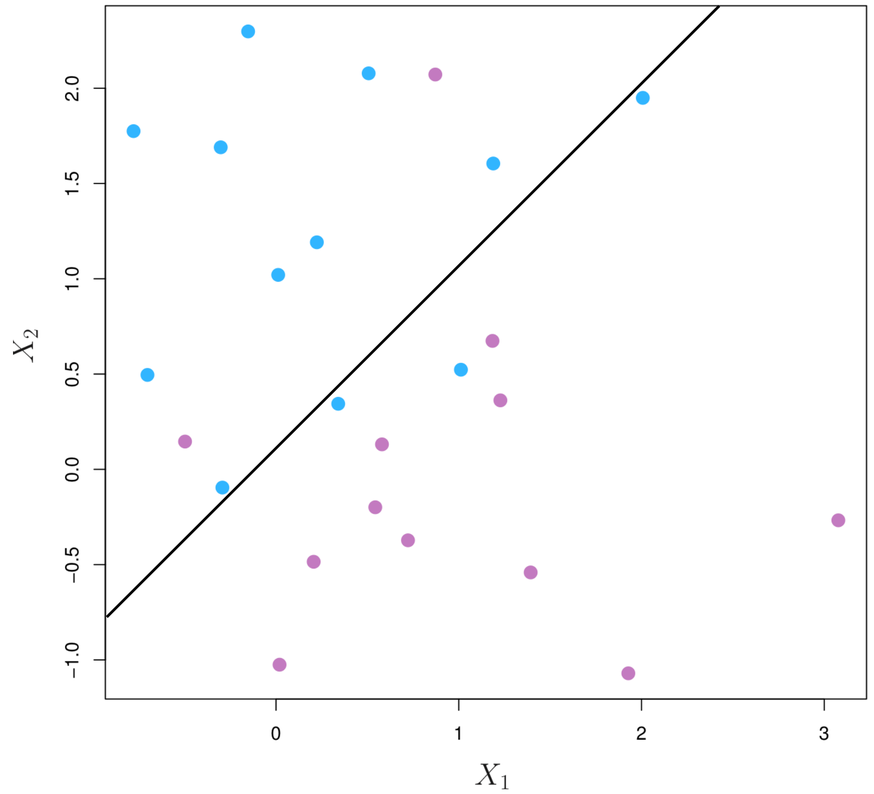
MAXIMIZE

Support Vector
최대 마진 초평면과 가장 가까운 데이터 포인트
이 점의 존재가 최대 마진 초평면의 위치를 **서포트**한다.
계산 비용을 줄이는 효과를 야기한다.

Support Vector Machine

1962
Maximal Margin Classifier

Maximal Margin Classifier, Too Ideal

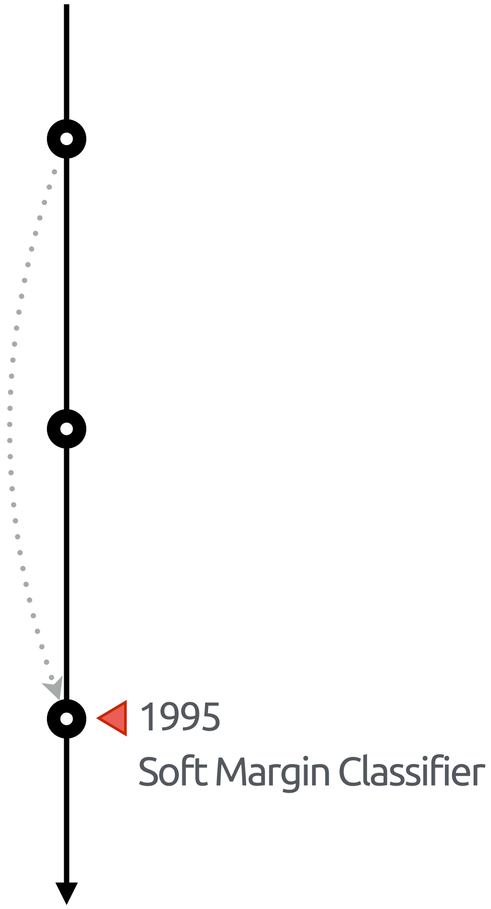


Maximal Margin Hyperplane를
이용해서 모든 데이터를 나눌 수 없다.

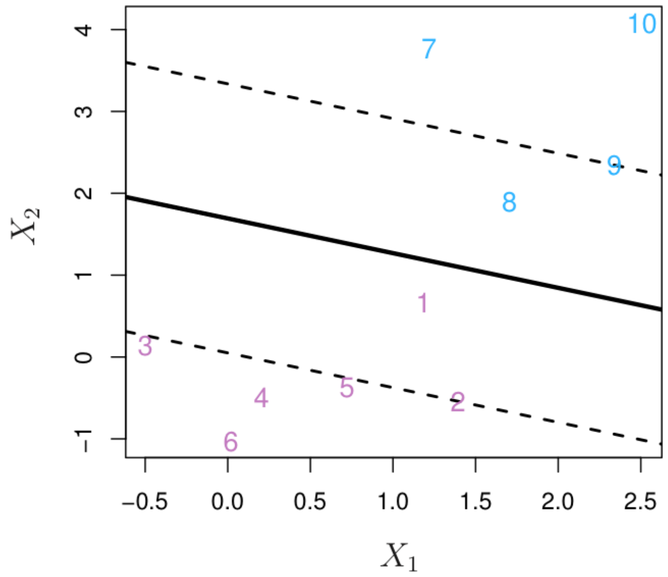
모든 데이터에 대해서 완벽하게 분리할 수 있는
분리 초평면이 존재할 때에만
최대 마진 분류기가 존재한다.

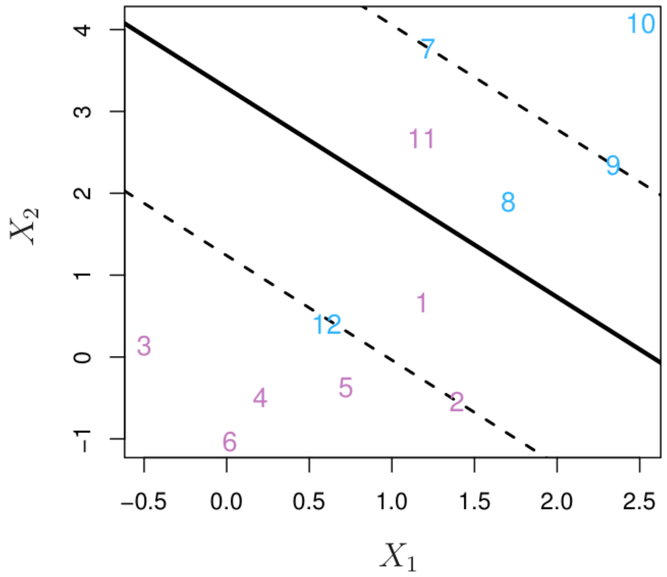
조금만 타협을 보면 되지 않을까?

Support Vector Machine



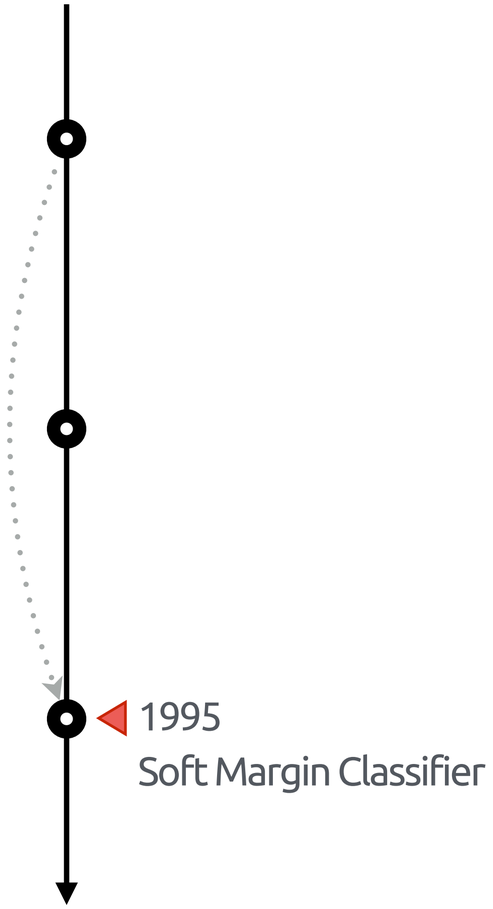
Let's Make a Deal



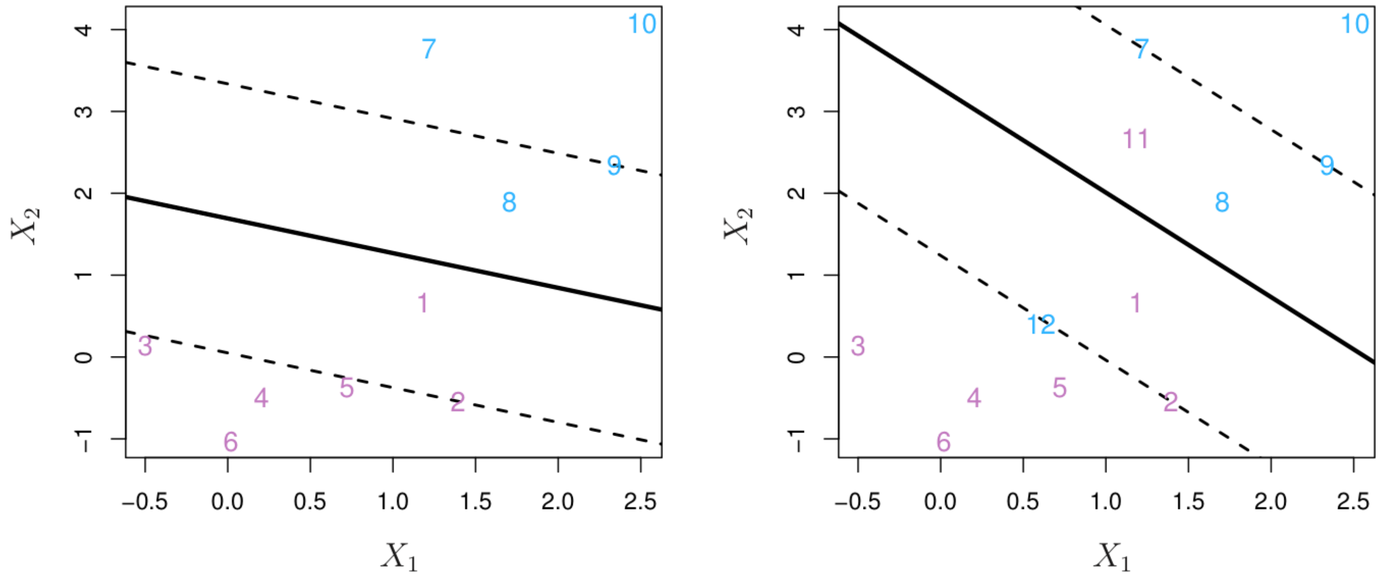


오분류된 데이터 일부를 허용할 수 있는 모델을 생각해보자.
오분류된 데이터를 얼마나 허용할 지를 적당하게 결정한다면 될 것 같다.

Support Vector Machine



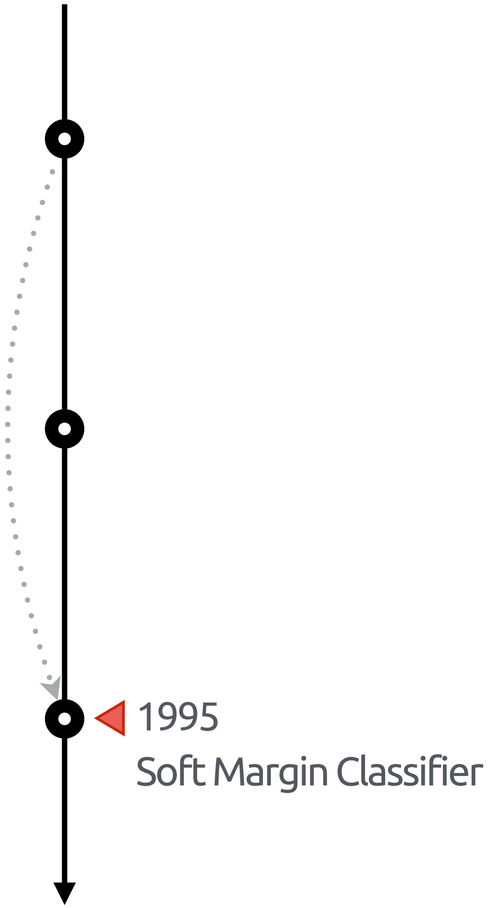
Slack Variable



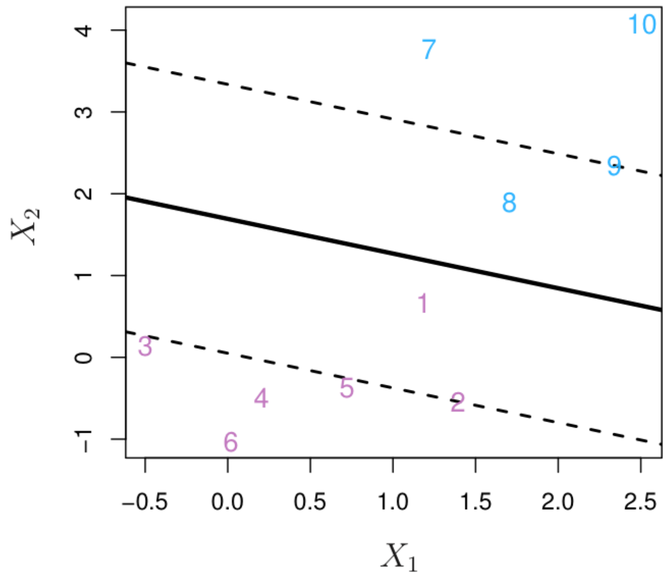
슬랙 변수(Slack Variable)
데이터 포인트의 위치를 나타낸다.

Tuning Parameter
슬랙 변수들의 합의 한계를 설정해주는 값

Support Vector Machine



Slack Variable



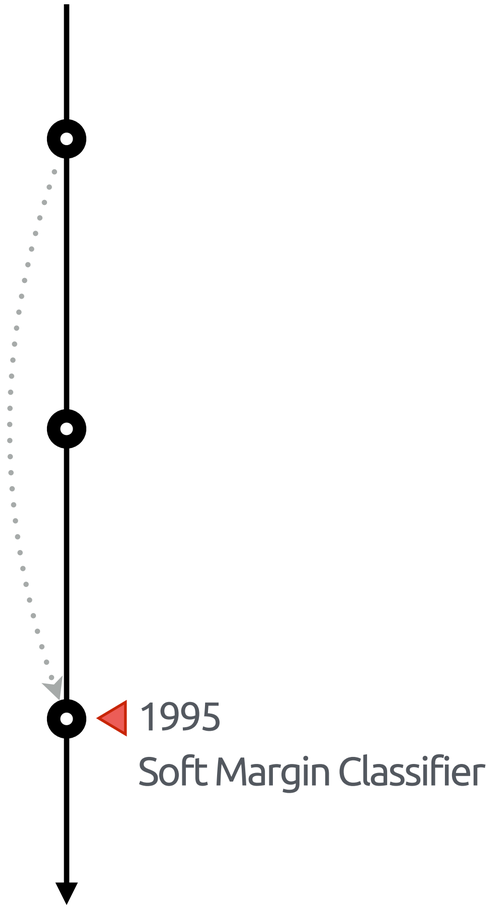
슬랙 변수(Slack Variable)

e = 0 : 올바르게 분류되었다.

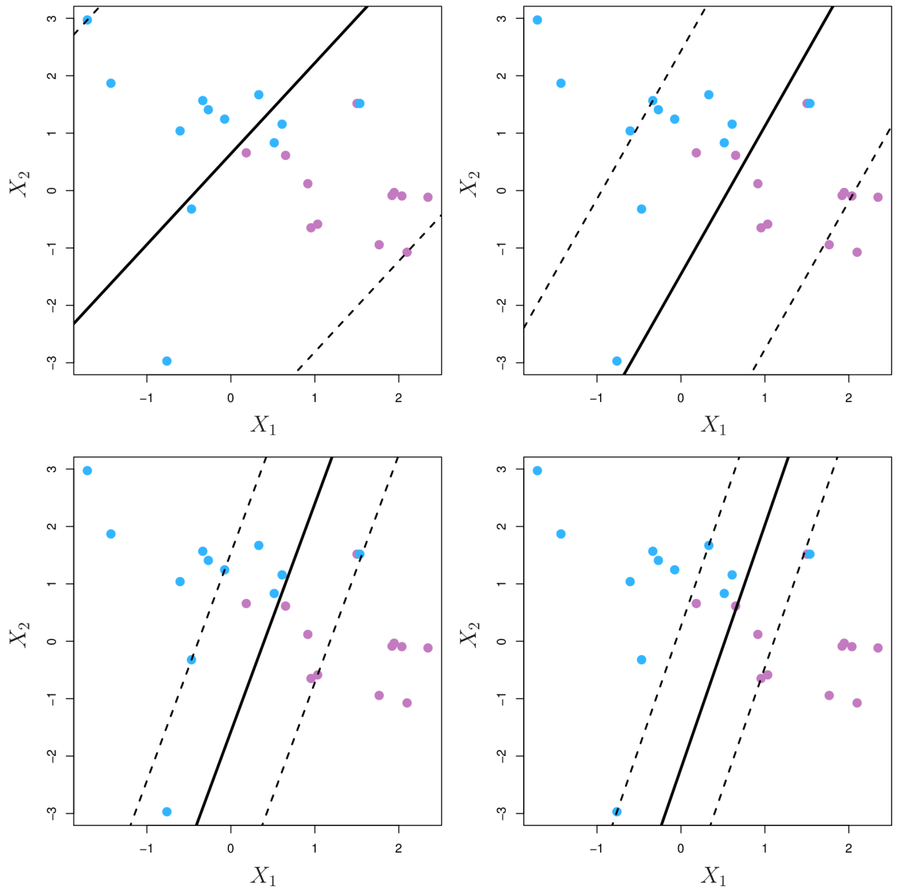
1 > e > 0 : 올바르게 분류 되었으나
마진을 넘어서는 범위에 있다.

e > 1 : 올바르게 분류되지 않았다.

Support Vector Machine



Soft Margin Classifier (Support Vector Classifier)



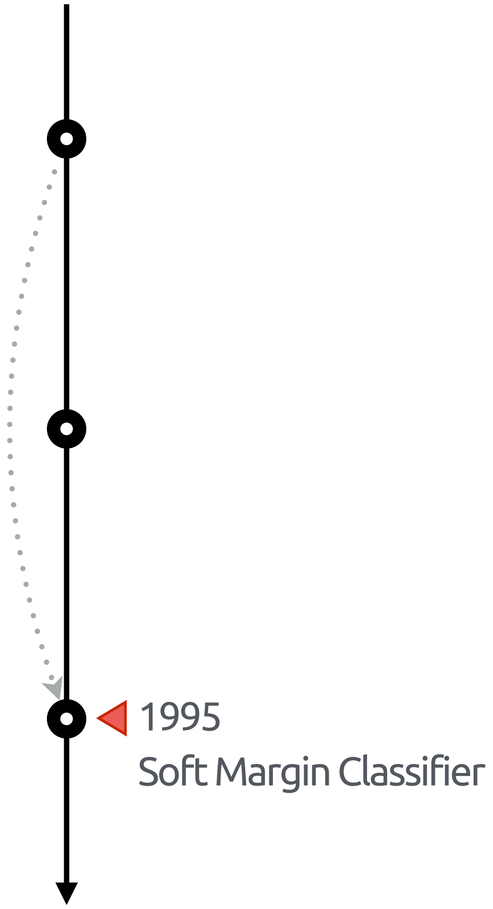
Tuning Parameter

설정에 따라서 모델이 많이 바뀐다.

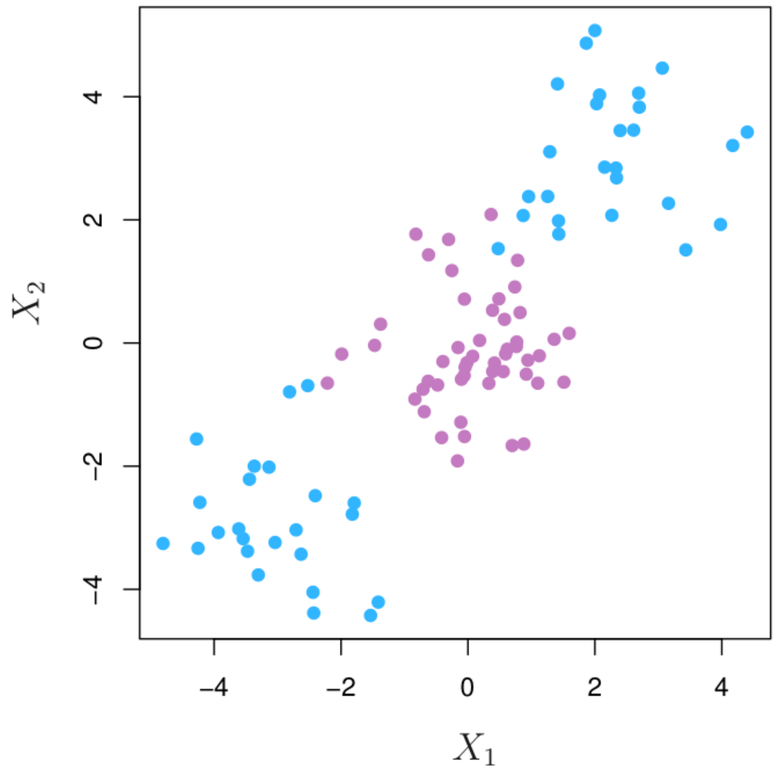
C 값이 크면? : 과소적합(underfitting)

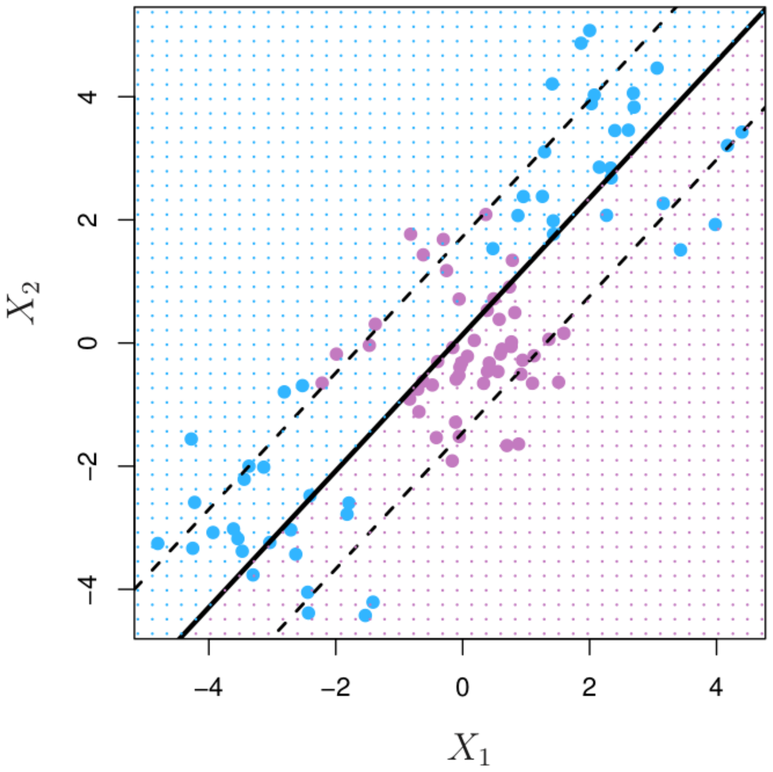
C값이 작으면 : 과적합(overfitting)

Support Vector Machine

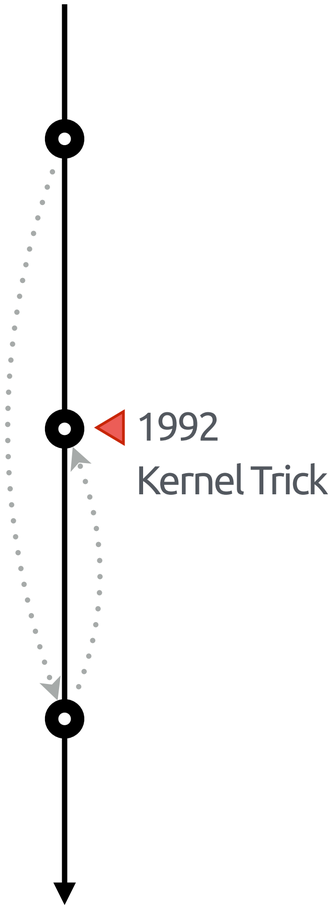


Oh My Days…

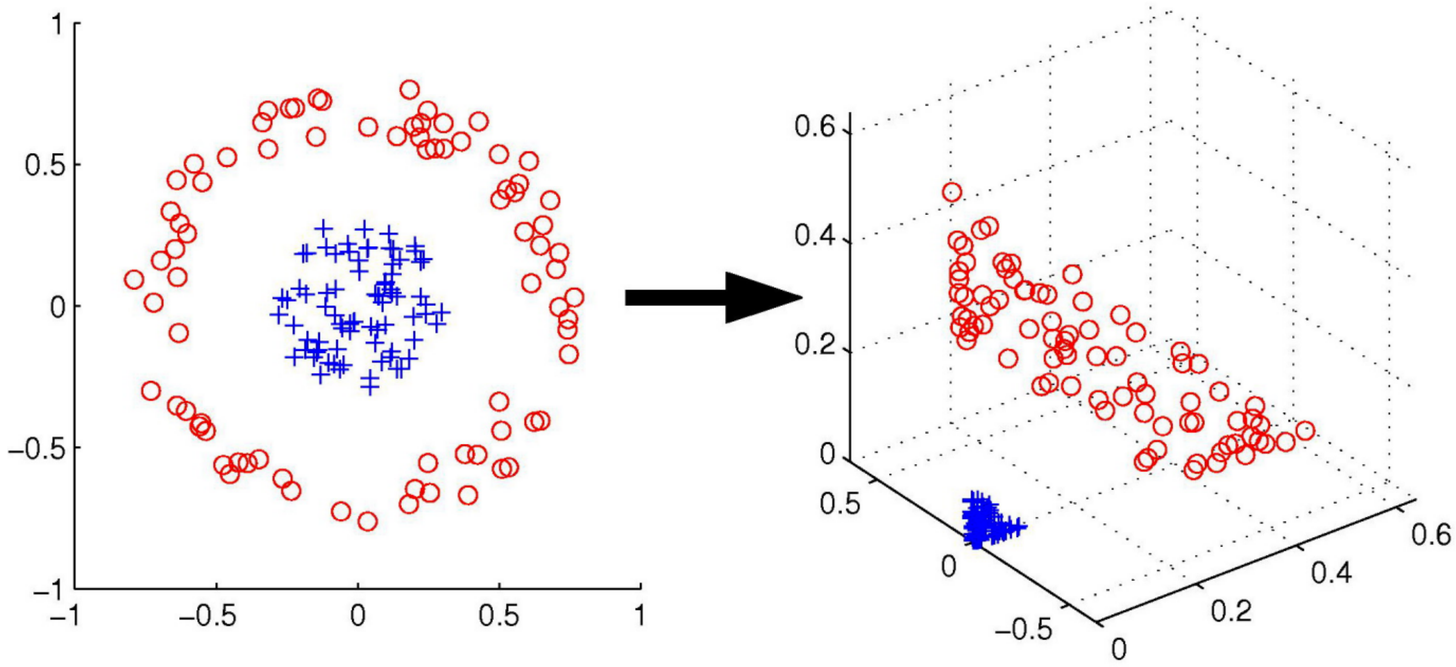




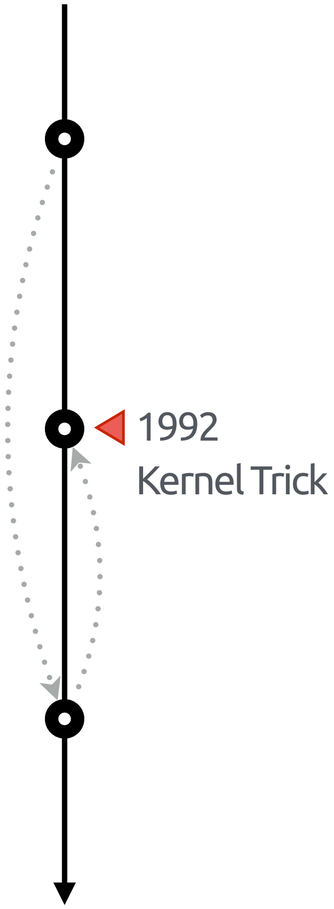
Support Vector Machine



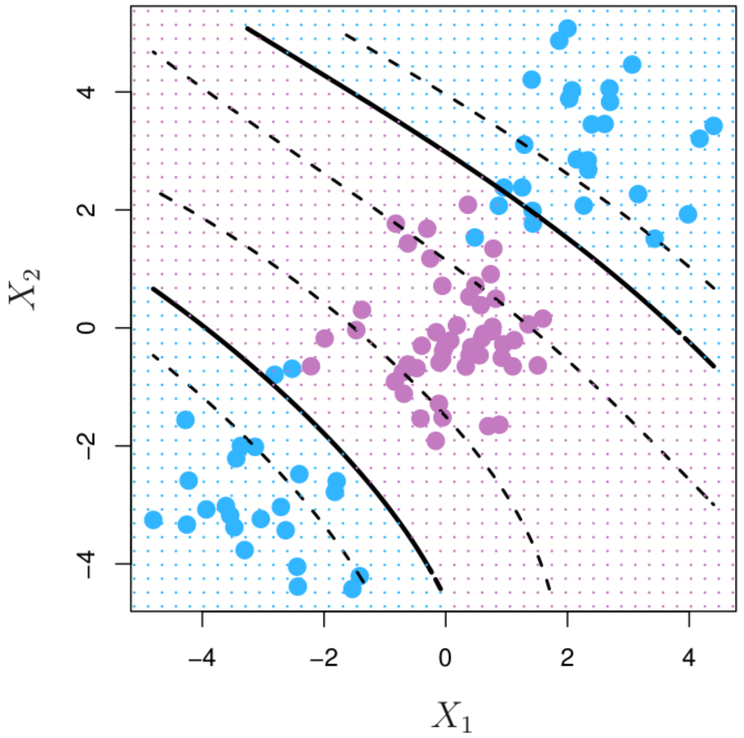
Kernel Trick



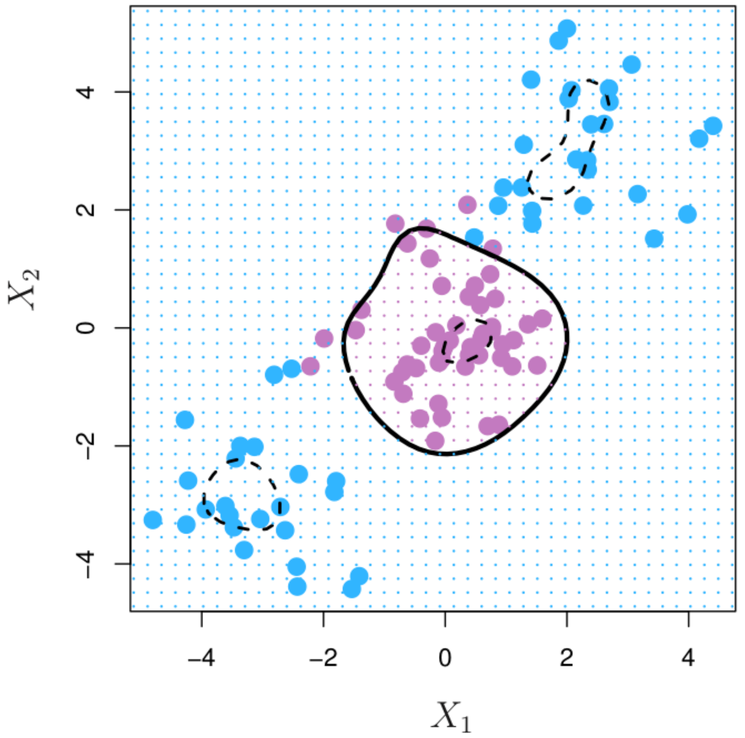
Support Vector Machine



With Kernel Trick (Support Vector Machine)

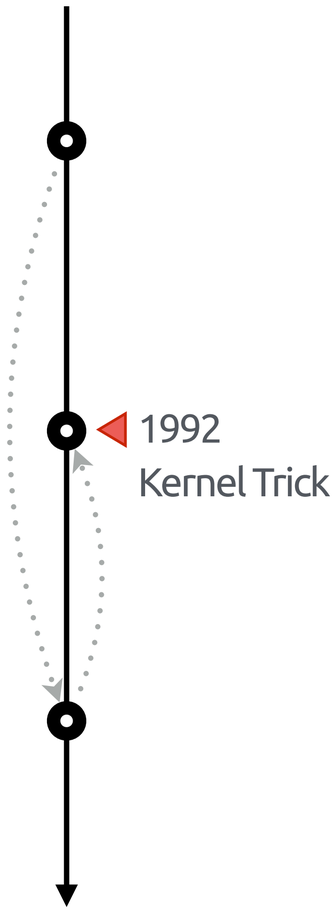


Polynomial Kernel



Radial Kernel

Support Vector Machine



Type of Kernel

Linear Function $K_l(\mathbf{x_i}, \mathbf{x_j}) = \langle \mathbf{x_i}, \mathbf{x_j} \rangle$

Polynomial Kernel $K_p(\mathbf{x_i}, \mathbf{x_j}) = (\gamma \cdot \langle \mathbf{x_i}, \mathbf{x_j} \rangle + r)^d$

Gaussian Radial Basis Kernel $K_r(\mathbf{x_i}, \mathbf{x_j}) = e^{-\gamma \cdot |\mathbf{x_i} - \mathbf{x_j}|^2}, \quad \text{where } \gamma > 0$

Sigmoid Kernel $K_s(\mathbf{x_i}, \mathbf{x_j}) = \tanh(\gamma \cdot \langle \mathbf{x_i}, \mathbf{x_j} \rangle + r)$

실습

서포트 벡터 머신

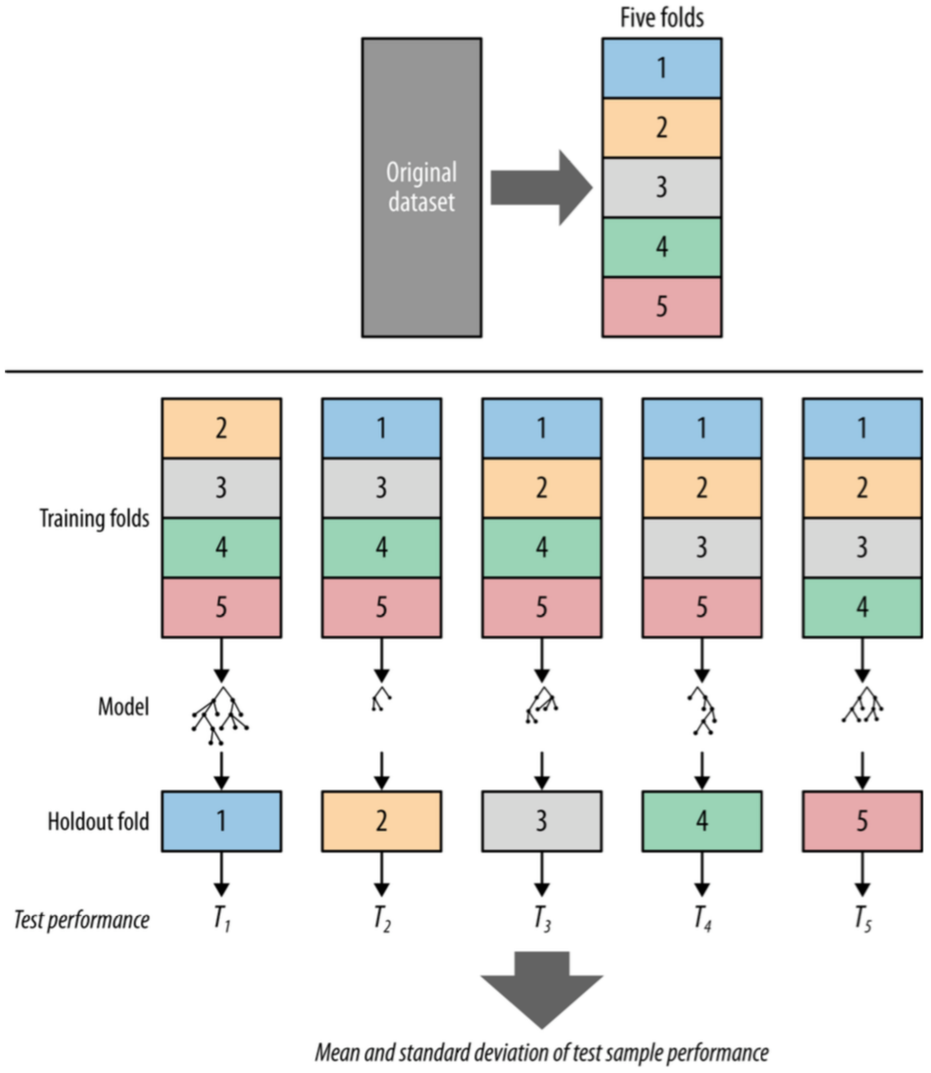
TOY DATA :)

실습

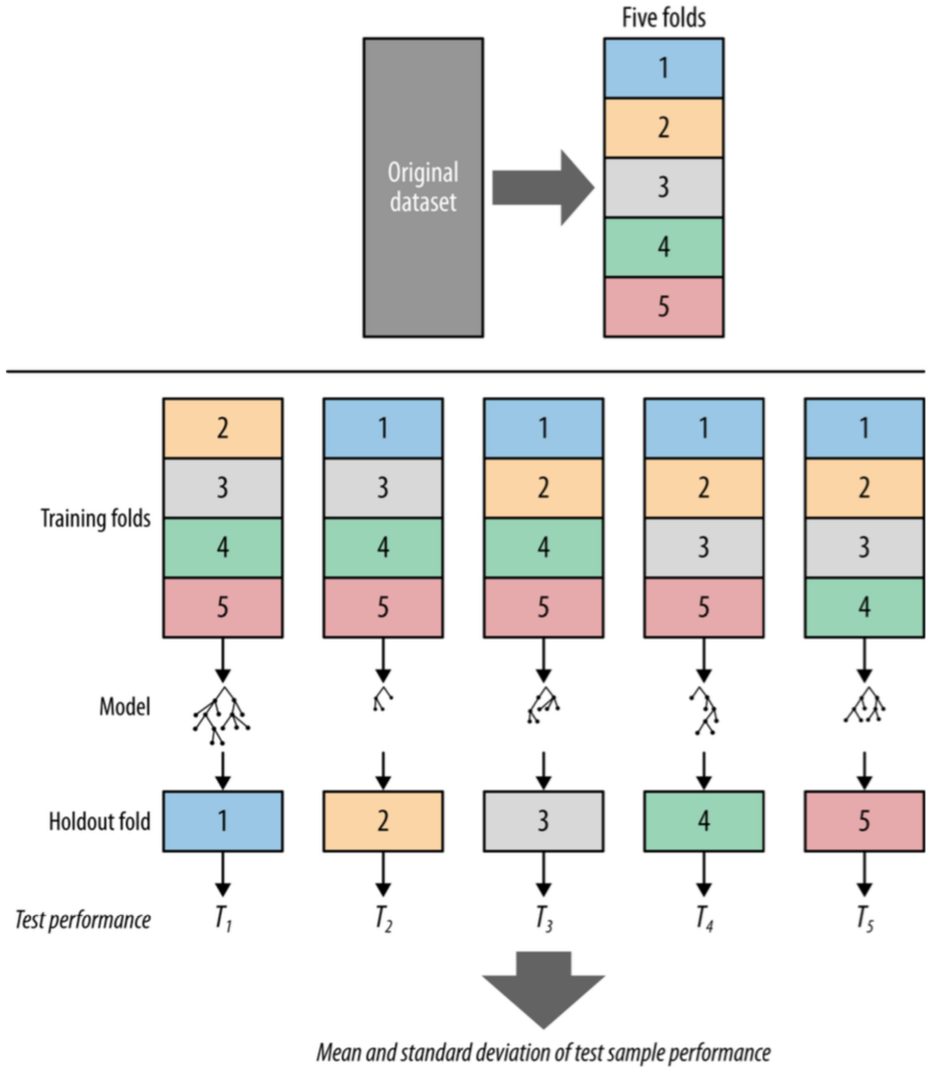
서포트 벡터 머신

DIABETES DIAGNOSTIC

Cross Validation



Cross Validation



과적합(overfitting) 방지

모델의 일반화 성능 향상

다양한 파라미터를 설정할 때,
굉장히 유용하게 사용할 수 있음

과제

자동차 연비 예측

서포트 벡터 머신을 이용하여
자동차 연비를 예측해보자.

과제

자동차 연비 예측

서포트 벡터 머신을 이용하여
자동차 연비를 예측해보자.

트레이닝 데이터 : `train_car`
테스트 데이터 : `test_car`

1. Accuracy가 95%가 넘는 모델을 구축할 것
2. 반드시 **`tune.svm()`** 함수를 사용하여,
   Cross Validation을 통한 최적의 모델을 구할 것
3. 예측 결과물은 **`confusionMatrix()`**를 이용하여
   모델의 퍼포먼스를 확인할 것
4. 해당 모델의 Parameter를 명시할 것
5. 사용한 커널 트릭이 무엇인지 명시할 것


THX :)