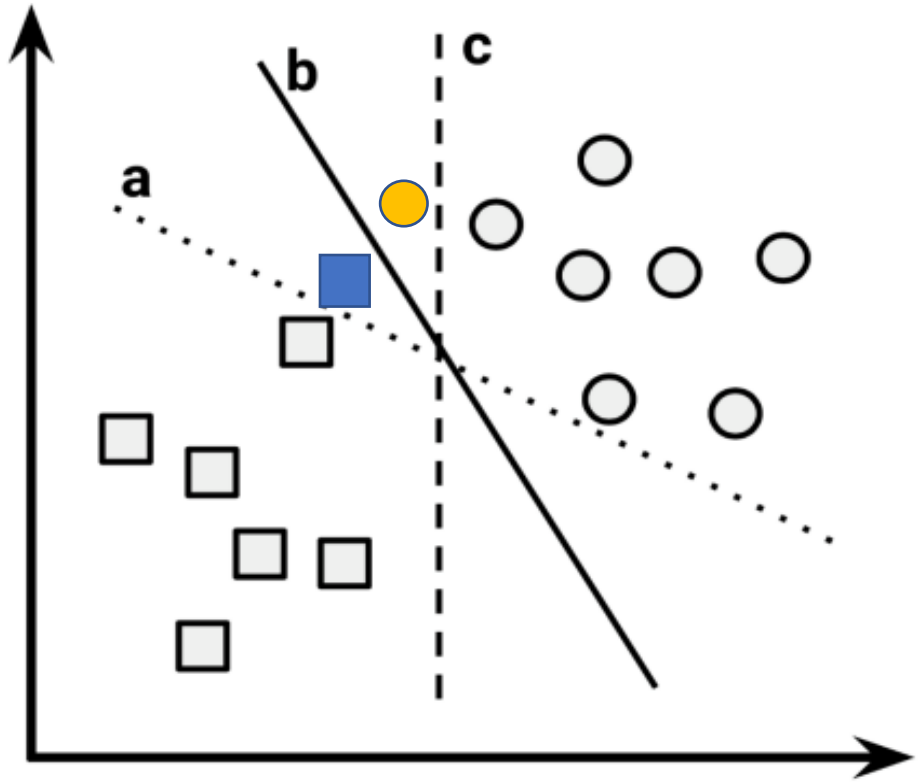
Support Vector Machine

숙명여자대학교 경영학부 오중산

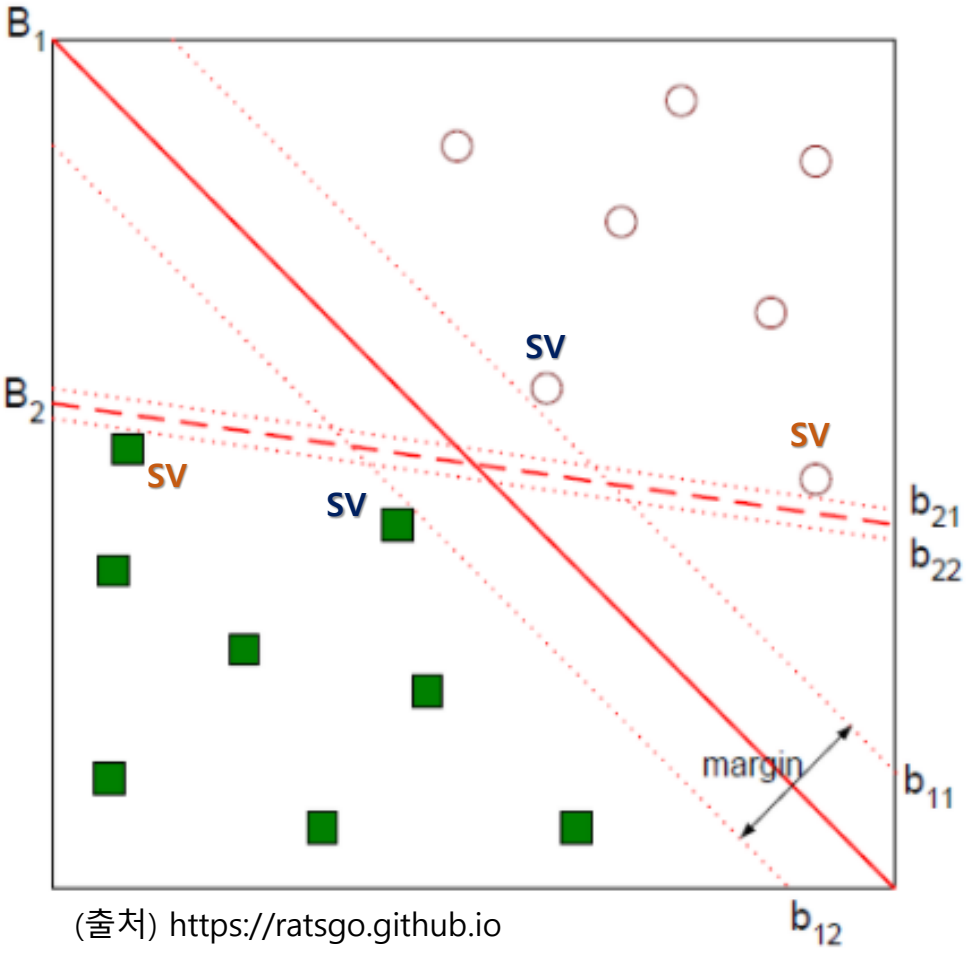
Support Vector Machine(SVM) 소개

- SVM이란?
  - 경계선/경계영역을 기준으로 사례를 선형으로 구분하는 머신러닝(지도학습) 기법
  - 사례 구분 경계를 초평면(hyperplane)이라고 함
- 세 개의 초평면에 대한 고찰
  - 초평면이 a와 c일 때의 문제는?



SVM 소개

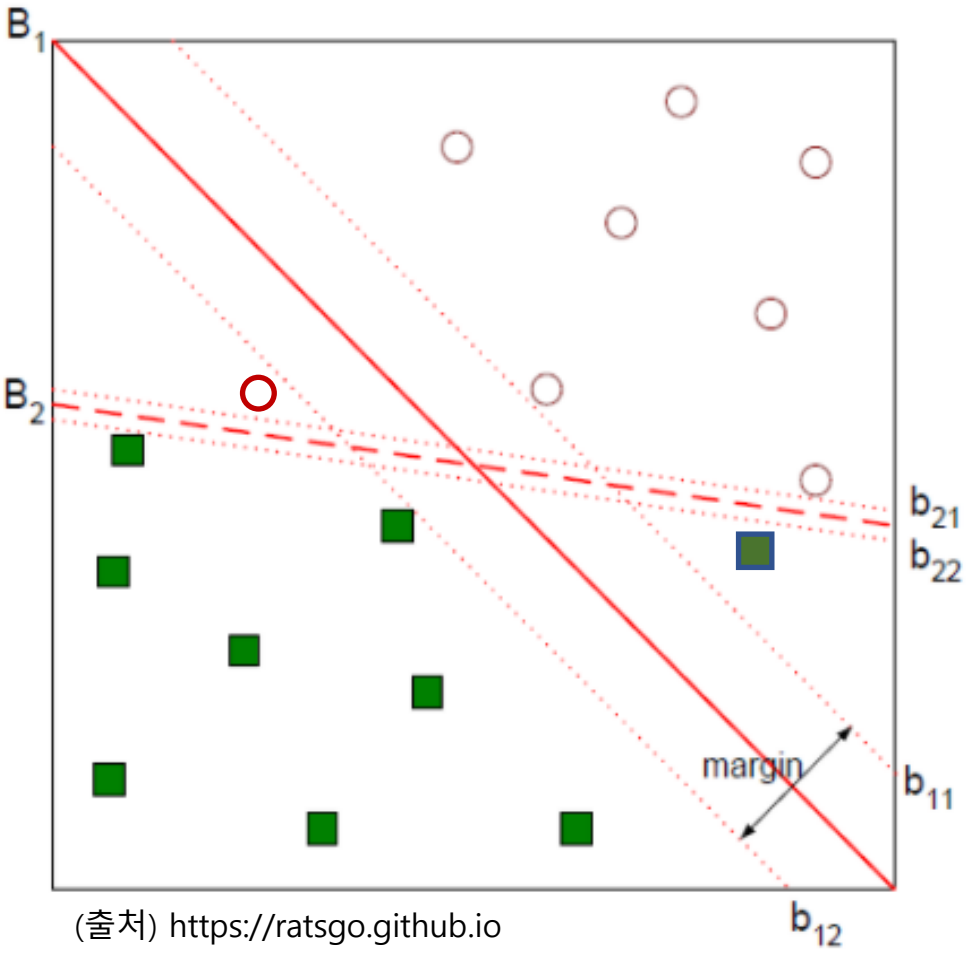
- 마진(margin)과 SV
  - ◆ 마진이란?
    - b_{i1}(plus plane)과 b_{i2}(minus plane) 간의 거리(i = 1, 2)
  - ◆ SVM은 마진의 최대화를 목표로 함
    - 마진을 최대로하는 초평면을 찾아서 사례 구분 정확도 제고가 목표
    - B_2가 아닌 B_1 선택
  - ◆ SV란?
    - SV는 b_{i1} 혹은 b_{i2} 위에 존재하여 마진 결정에 영향을 미치는 사례



(출처) https://ratsgo.github.io

SVM 소개

- SVM에서의 과대적합과 과소적합
  - 마진을 너무 줄여서 이상치가 포함되면 과대적합 발생
    - 이상치 오분류를 일정 정도 용인하여 과대적합 문제에 대응
  - 마진을 너무 극대화해서 이상치를 배제하면 과소적합 발생
    - 이상치를 적절하게 포함하여 과소적합에 대응



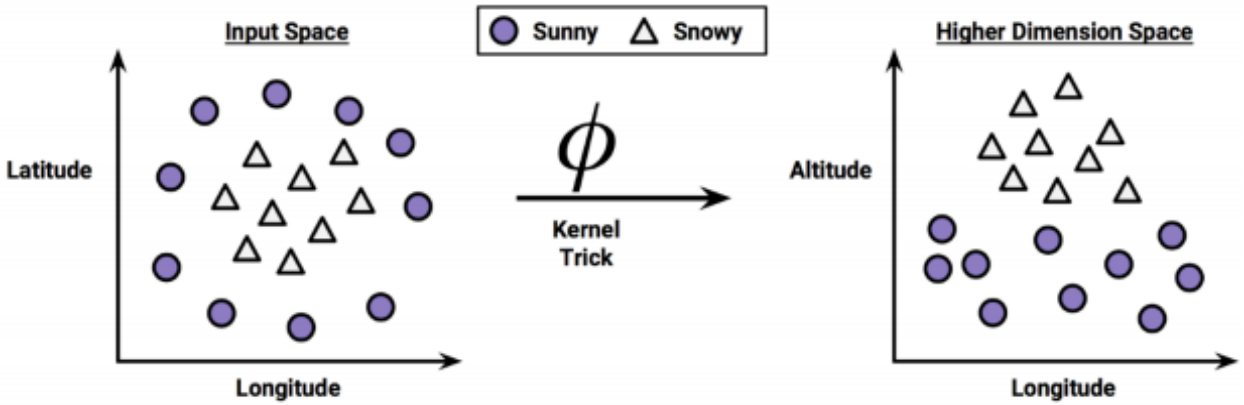
(출처) https://ratsgo.github.io

SVM 소개

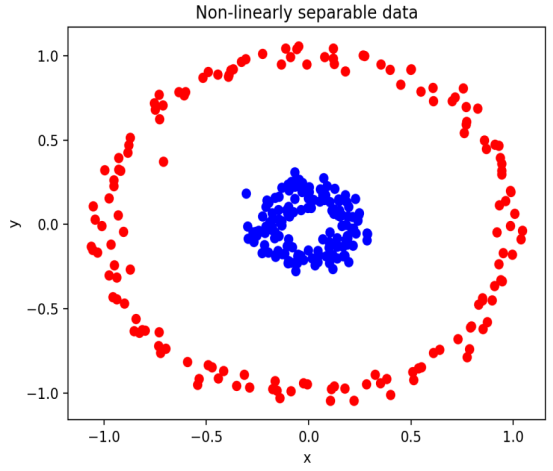
- SVM에서의 과대적합과 과소적합
  - 사례 오분류 비용(c)을 적절히 설정하여 과대적합과 과소적합 조절
    - 오분류 비용은 어떤 사례를 잘못 분류했을 때 발생하는 비용
    - 오분류 비용을 크게 설정하면 오분류를 줄이기 위해 마진이 작아져 과대적합 발생
    - 오분류 비용을 작게 설정하면 오분류를 용인해도 부담이 없어서 마진이 커져 과소적합 발생

SVM 소개

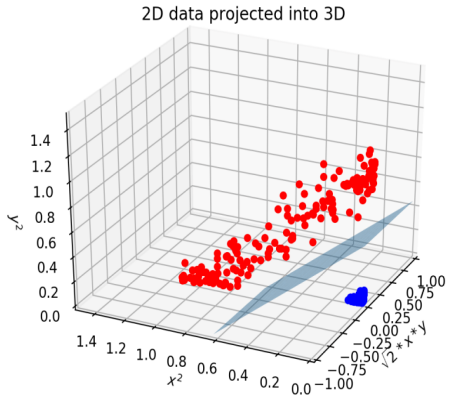
- 커널트릭(Kernel trick)
  - ◆ 선형 초평면으로는 사례를 정확히 구분하는 것이 쉽지 않음
  - ◆ 매핑함수(Φ, 파이)를 이용한 커널트릭
    - 매핑함수를 이용하여 사례를 원공간(input space)에서 고차원공간(feature space)으로 매핑하면 선형 초평면으로 구분이 가능함



[선형 초평면으로 데이터 분리 어려움]



[다항커널트릭을 이용한 선형 초평면 형성]



SVM 실습: 암진단

- STEP1: 데이터 프레임 준비 및 전처리
  - KNN과 동일함!
  - cancer.csv 파일을 불러와서 데이터프레임(cancer) 구성
  - str과 summary를 활용한 데이터 검토
  - class(DV) 척도를 범주형으로 변경
  - 불필요한 변수 id 제거
  - IV에 대한 표준화 및 새로운 데이터 프레임(cancer_svm) 구성

SVM 실습: 암진단

- STEP2: train 데이터셋과 test 데이터셋 구성
  - KNN과 동일함!
  - cancer_svm을 7:3의 비율로 두 개 데이터 프레임으로 구분
    - sample 함수를 쓸 때 결과가 달라지지 않도록 set.seed() 함수를 먼저 사용해야 함
      - sample함수 기본 형식: sample(추출 대상, 추출 개수, replace = F/T, prob = c(p1, p2))
    - 전자(70% 비율)는 train 데이터 프레임(cancer_svm_train)으로 구성
    - 후자(30% 비율)는 test 데이터 프레임(cancer_svm_test)으로 구성

SVM 실습: 암진단

- STEP3: train 데이터 프레임을 이용한 훈련
  - ◆ e1071 패키지의 tune.svm() 함수 사용
    - ▪ 네 가지 커널트릭
      - ❖ 선형(linear), 다항(polynomial), 방사기저(radial base function: RBF), 시그모이드(sigmoid)
    - ▪ 선형 커널트릭: linear.svm <- tune.svm(class~., data = cancer_svm_train, kernel = “linear”, cost = c())
      - ❖ 사전에 set.seed 함수 실행해야 함
      - ❖ 기본적으로 10-fold cross validation으로 실행

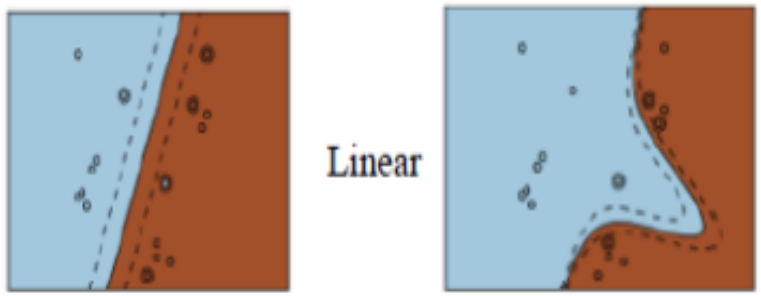
SVM 실습: 암진단

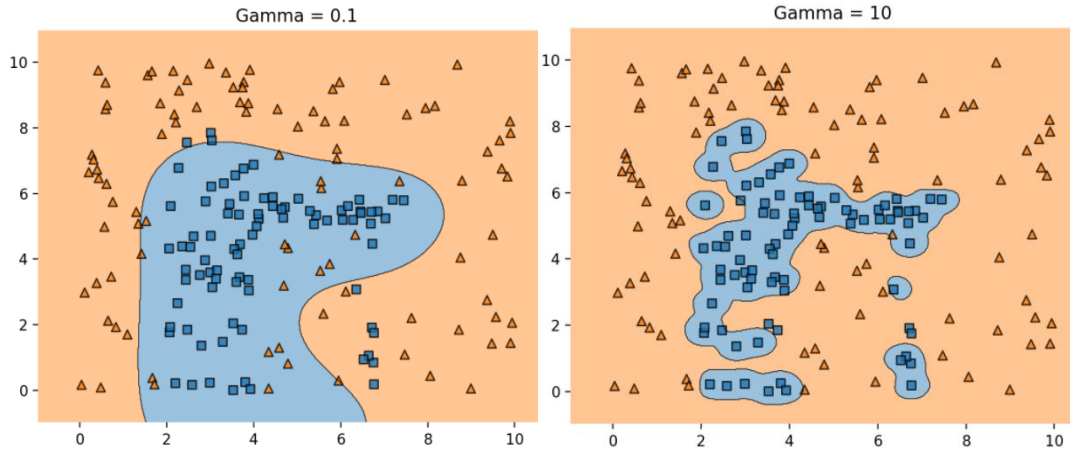
- STEP3: train 데이터 프레임을 이용한 훈련
  - ◆ summary(linear.svm) 실행
    - 최적의 cost와 이때 정확도 확인
  - ◆ linear.svm$best.model 실행
    - 최적의 cost와 SV 개수 확인
- STEP4: test 데이터를 이용한 성능 평가
  - ◆ predict 함수 사용하여 test 데이터에 적용하고, 결과 저장
    - 기본 형식: test model <- predict(linear.svm$best.model, new data = cancer_svm_test)
    - confusionmatrix 함수 사용하여 Accuracy와 Kappa 값 기준 성능 평가 및 과적합문제 검토

SVM 실습: 암진단

- STEP4: test 데이터를 이용한 성능 평가
  - ◆ 다양한 커널트릭으로 STEP3와 STEP4 반복
    - poly.svm <- tune.svm(class~., data = cancer_svm_train,

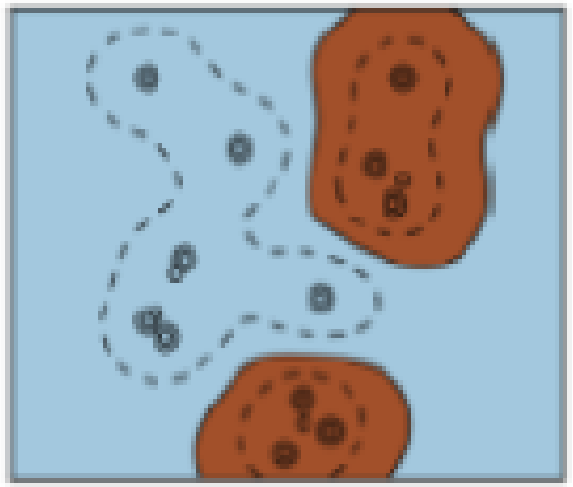
      kernel = "polynomial", degree = c(), gamma = seq(), coef0 = seq(), cost = c())





- ❖ degree는 다항식의 차수로 보통 2~5
- ❖ gamma가 커지면 초평면이 복잡해져 과대적합 우려
- ❖ coef0는 커널 계수로 gamma와 동일하게 설정
- ❖ 적절한 파라미터를 찾는 것이 관건

SVM 실습: 암진단



RBF

- STEP4: test 데이터를 이용한 성능 평가
  - ◆ 다양한 커널트릭으로 STEP3와 STEP4 반복
    - rbf.svm <- tune.svm(class~., data = , kernel = “radial”, gamma = seq(), cost = seq())
      - ❖ 방사형커널은 가우시안커널이라고도 하며 가장 널리 사용되는데, 사례를 무한 차원으로 매핑
    - sigmoid.svm <- tune.svm(class~., data = , kernel = “sigmoid”, gamma = seq(), coef0 = seq())
      - ❖ 시그모이드함수 대체재인 쌍곡선 탄젠트 함수(tanh) 활용
  - ◆ 최적의 모델 선정
    - Accuracy와 Kappa가 가장 큰 커널트릭에 기반한 모델 선택

SVM 실습: 암진단

- STEP5: 성능 개선
  - STEP4에서 선정된 커널트릭에 대해 파라미터(매개변수) 조정
- STEP6: 예측
  - 성능 개선된 모델을 이용하여 새로운 사례에 대해 예측
  - 예측 결과를 KNN을 이용한 예측 결과와 비교