자동 문서 분류

국민대학교 소프트웨어학부
강 승 식

자동 문서 분류

- 미리 정의된 범주에 문서를 자동으로 할당하는 기법

- 필요성
  - 문서 생산량의 기하급수적 증가
  - 수작업 문서분류의 한계 극복

- 기계학습 기법
  - 레이블 있는 학습문서 집합으로부터 추출된 정보에 기초하여 레이블이 없는 문서를 미리 정해진 범주로 분류

문서분류 방법

- k-Nearest Neighbor (kNN)
- Roccio
- Naïve Bayes (NB)
- Support Vector Machines (SVM)

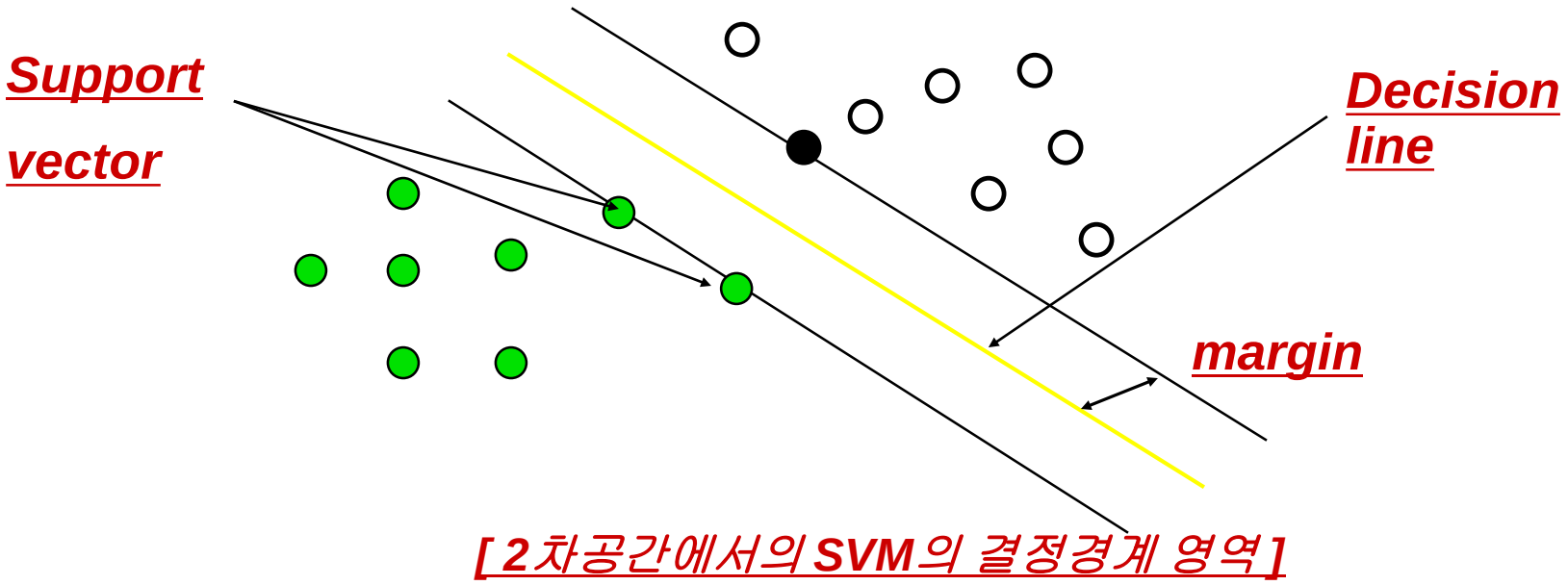
- Neural Network
- Linear Least Squares Fit (LLSF)

최근린법(k-Nearest Neighbors)

- 새로운 문서에 대한 범주를 결정할 때
- 학습문서에서 그 문서와 가장 가까운 k개의 문서들을 추출하여,
- k개 문서가 속하는 범주를 이용하여 새로운 문서의 범주를 할당

SVM: Support Vector Machine

- 두개의 클래스의 구성 데이터들을 가장 잘 분리할 수 있는 결정 경계면(decision surface)을 찾는 것



[2차공간에서의 SVM의 결정경계 영역]