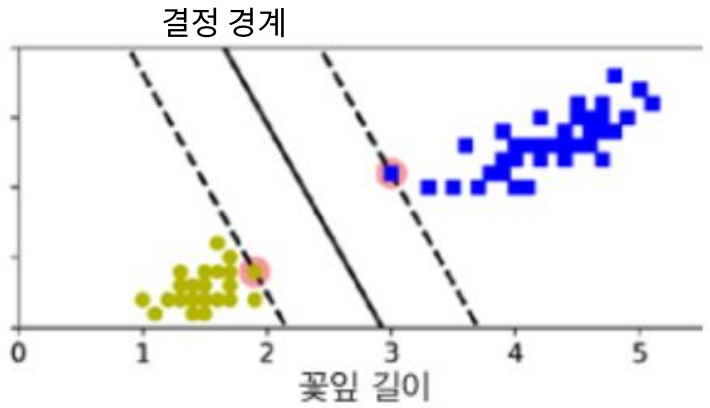
서포트 벡터 머신 (SVM) - 분류

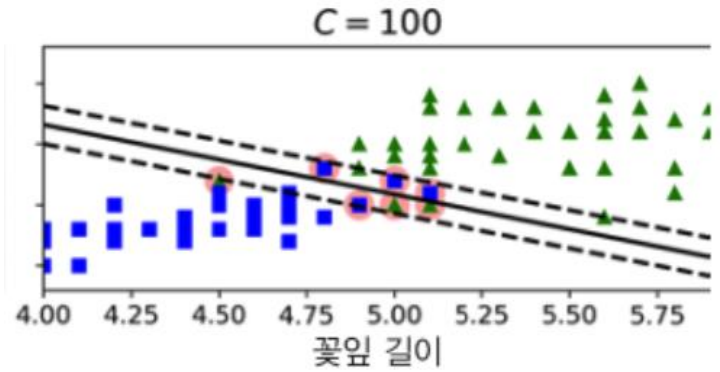
- 선형/비선형
- 분류/회귀/이상치 탐색

- **서포트 벡터** (Support Vector) : 경계에 위치한 샘플

  ➔ 결정 경계에 영향을 끼치는 샘플



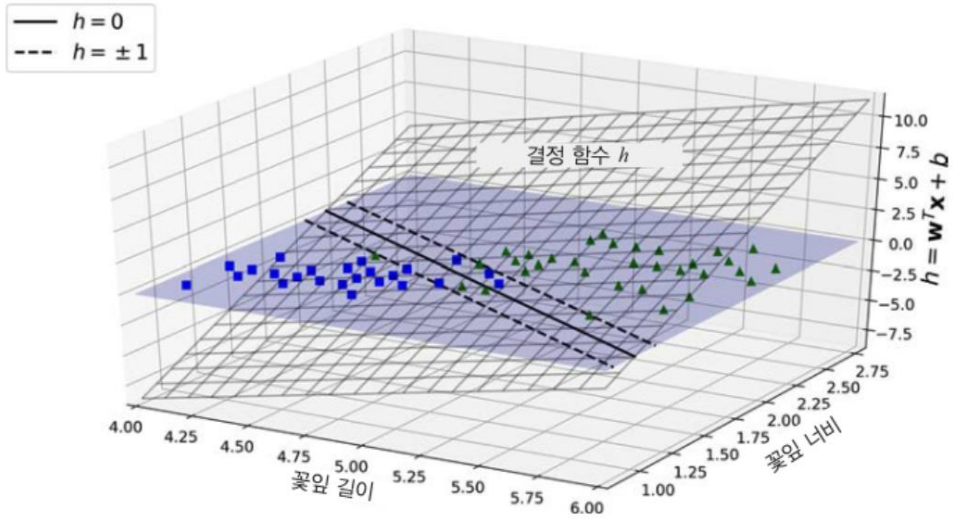
- 복잡한 데이터의 경우, 완벽한 분류가 불가능할 수 있음



- 마진 오류 : 경계 내부 또는 반대쪽에 위치한 샘플로 인해 발생
- 마진 오류를 줄이기 위해 결정 경계를 좁히면, 과대적합
- 일반화를 위해 결정 경계를 넓히면, 마진 오류 증가
- sklearn.svm 에서는 “C”라는 하이퍼파라미터를 통해 이 둘의 균형을 설정

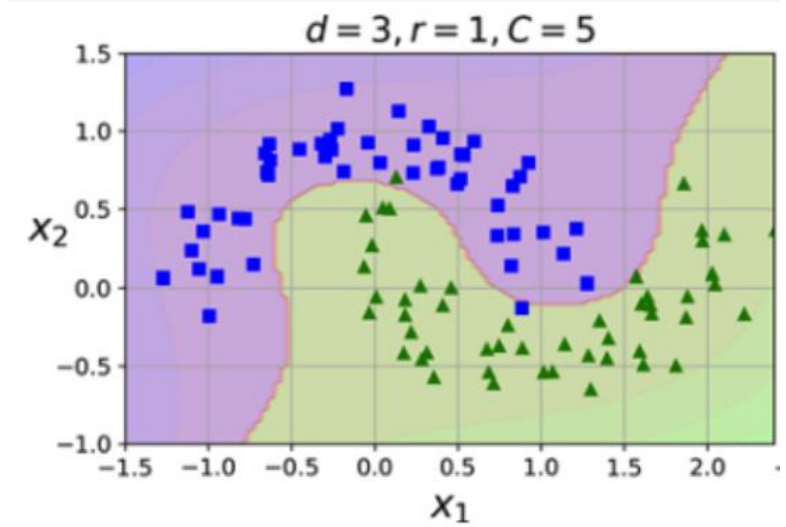
▪ 선형 SVM

- 결정 경계(함수)가 선형 식!
- sklearn.svm.LinearSVC



▪ 비선형 SVM

- 비선형 SVM은 커널 SVM으로 불리기도 함
- 커널(Kernel) : ~~'linear'~~, 'poly', 'rbf', 'sigmoid'
- sklearn.svm.SVC



- **규제 파라미터 C**
  - 작은 C → 넓은 마진(결정 경계의 폭) → 일반화에 유리
  - 큰 C → 좁은 마진 → 과대적합에 유리

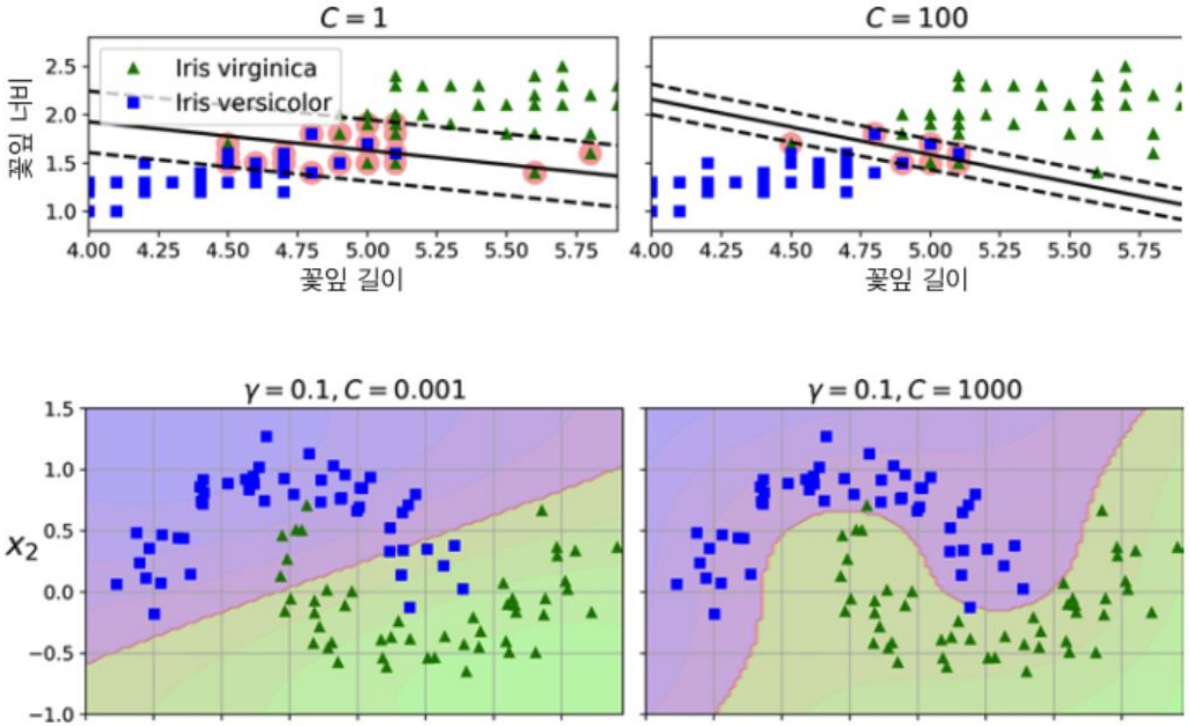
식 5-4 소프트 마진 선형 SVM 분류기의 목적 함수[20]

$$\underset{w,b,\zeta}{\text{minimize}} \ \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{m}\zeta^{(i)}$$

경계를 얼마나 위반하는지

- C 가 클수록, 적은 샘플이 경계를 위반해도 그 영향이 커짐

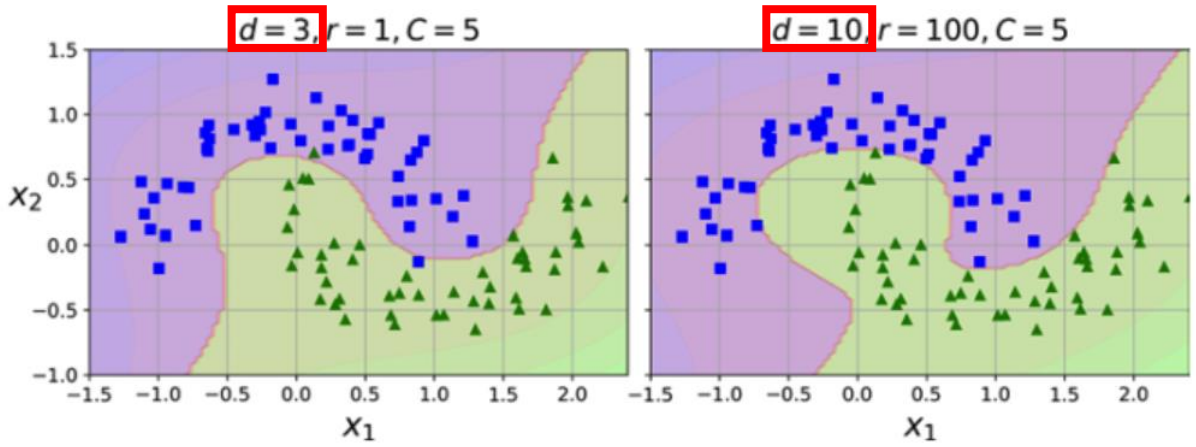
  ➔ 많은 샘플이 경계를 위반하지 않도록 마진 폭이 좁아짐



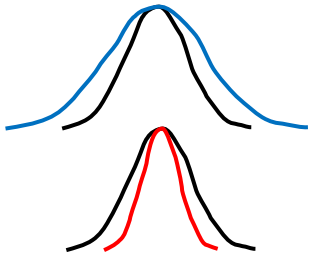
- 하이퍼파라미터인 C의 최적 값을 찾기 위한 방법은 일반적으로 “그리드 탐색“, “무작위 탐색“, “베이지안 최적화“와 같이 검증 데이터셋을 활용하는 방법이 있음

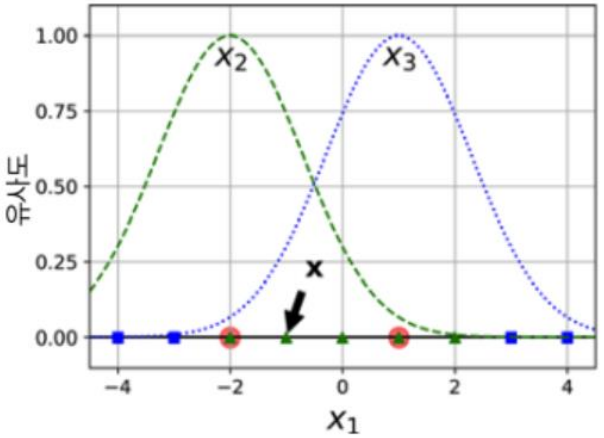
- **커널**
  - 복잡한(비선형) 특성을 학습할 때 사용됨

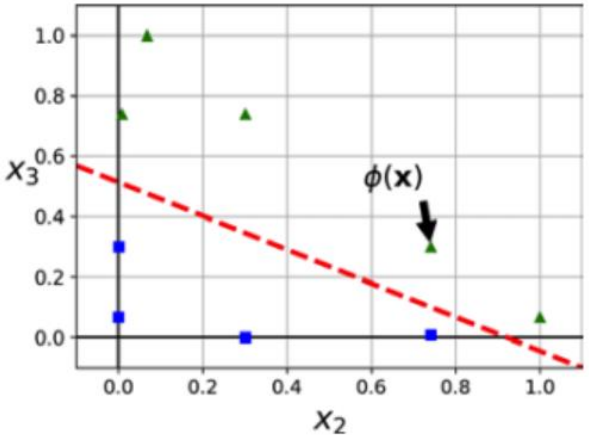
- **다항식 커널 (kernel = 'poly')**
  - 낮은 차수의 다항식 : 복잡한 특성을 표현하지 못함, 빠름
  - 높은 차수의 다항식 : 복잡한 특성을 표현 가능함, 느림
  - 과대적합 → **차수(degree) 낮춤** → 일반화
  - 과소적합 → **차수(degree) 높임** → 적합



- **가우시안 RBF 커널 (kernel = 'rbf')**
  - 유사도 특성 방식을 적용
  - 과대적합 → **gamma 작게** → 일반화
  - 과소적합 → **gamma 크게** → 적합



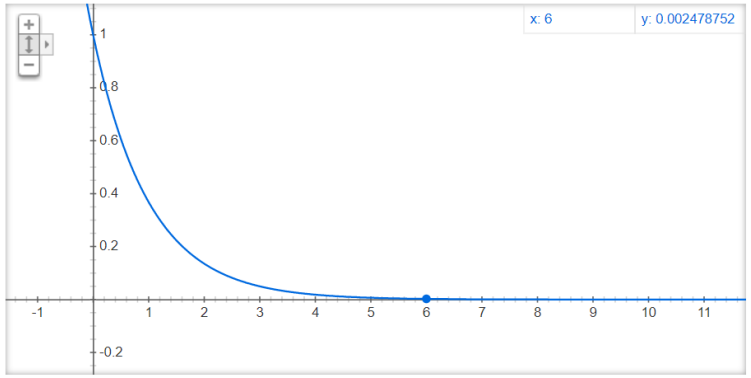


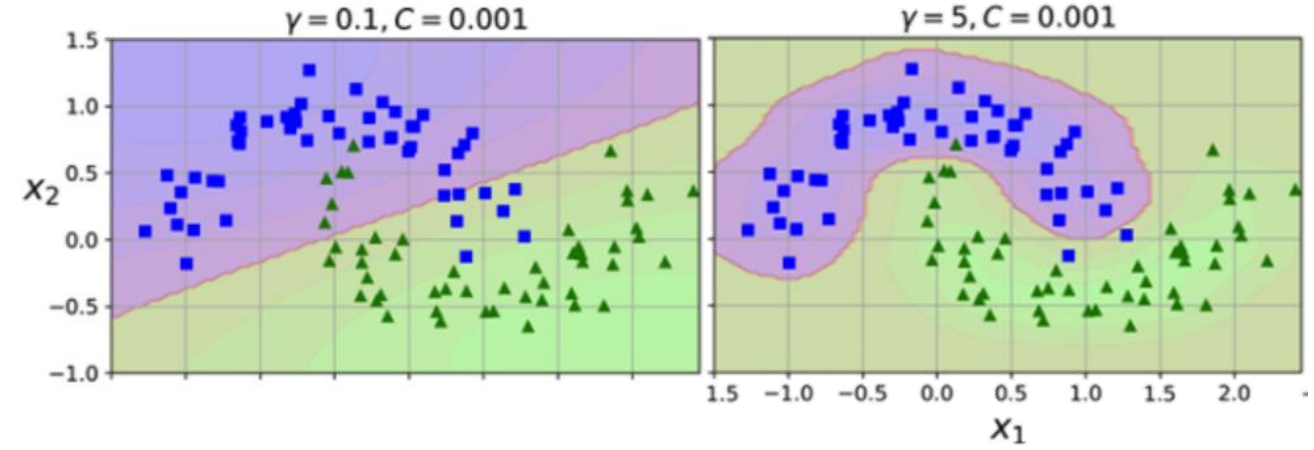


식 5-1 가우시안 RBF

$$\phi_\gamma(\mathbf{x}, \ell) = \exp\left(-\gamma \|\mathbf{x} - \ell\|^2\right)$$

exp(-x) 그래프





- **복잡도**
  - m : 훈련 샘플 수
  - n : 특성 수

파이썬 클래스	시간 복잡도	외부 메모리 학습 지원	스케일 조정의 필요성	커널 트릭
LinearSVC	$O(m \times n)$	아니오	예	아니오
SGDClassifier	$O(m \times n)$	예	예	아니오
SVC	$O(m^2 \times n) \sim O(m^3 \times n)$	아니오	예	예