# 『Hands-On Machine Learning』

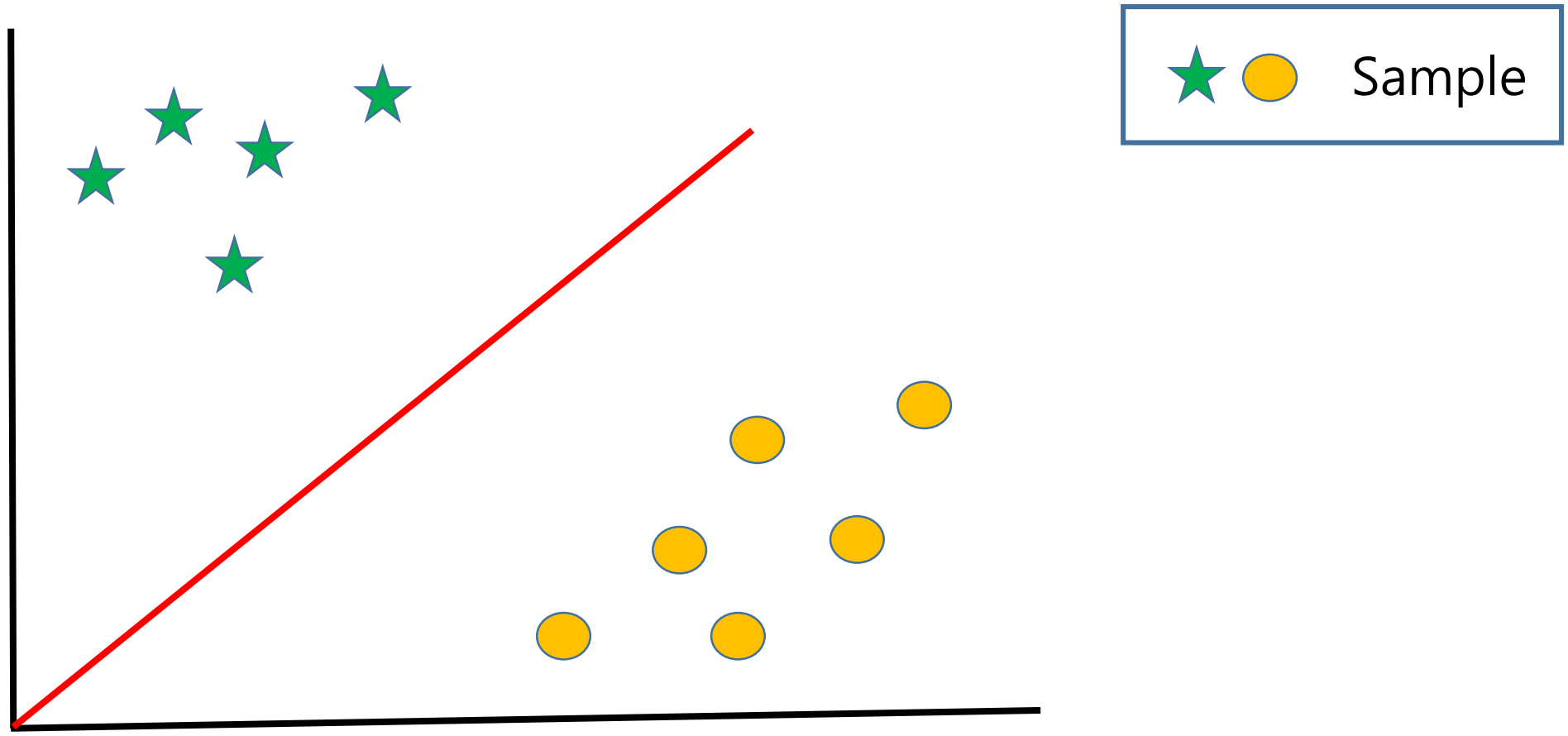
## Chapter 5. Support Vector Machine

SangHyeok Kim

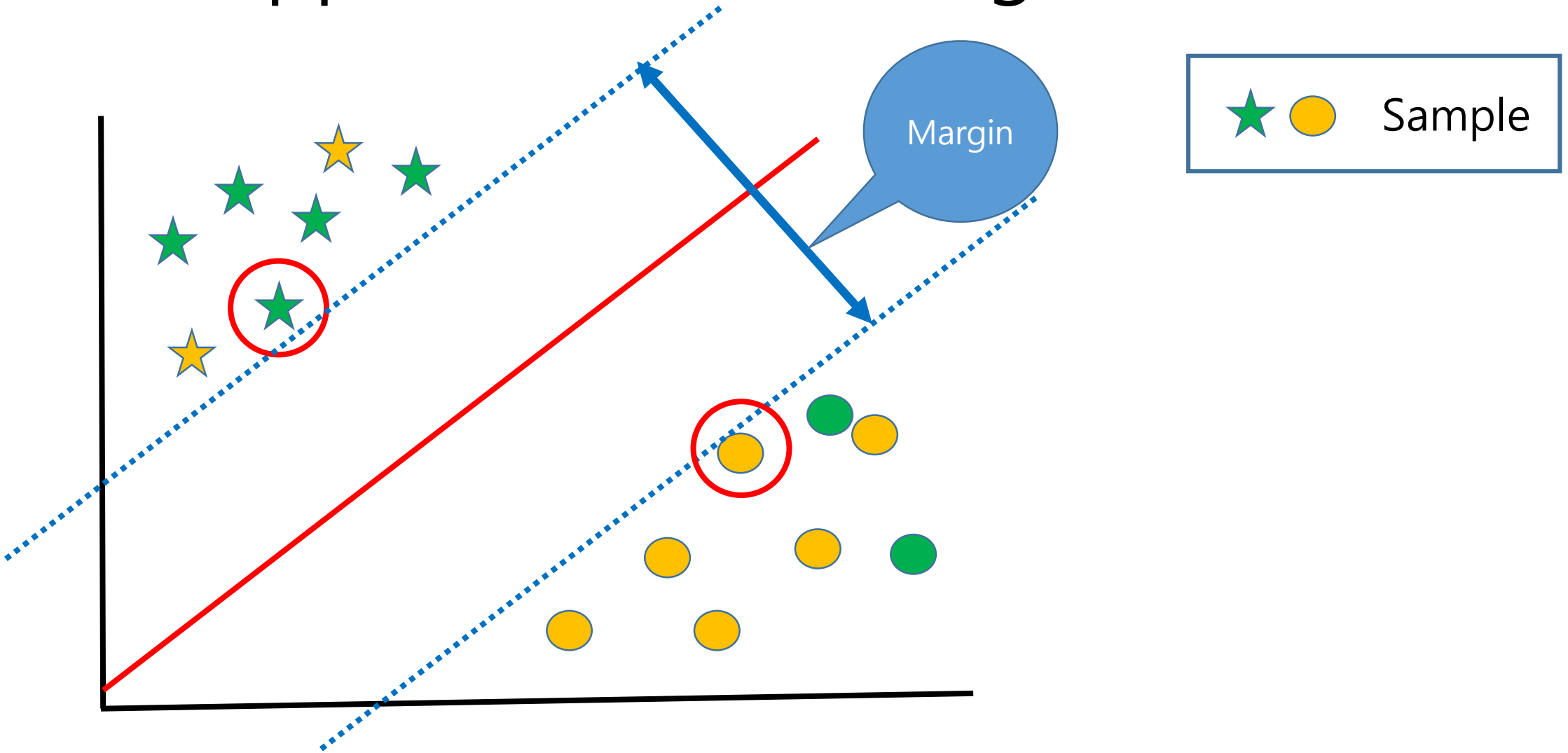
# Contents

1. 용어 설명 (SVM, SV, Margin)
2. SVM의 종류
3. SVM의 특징 및 장단점
4. Q&A

# 1.1 SVM(Support Vector Machine)??



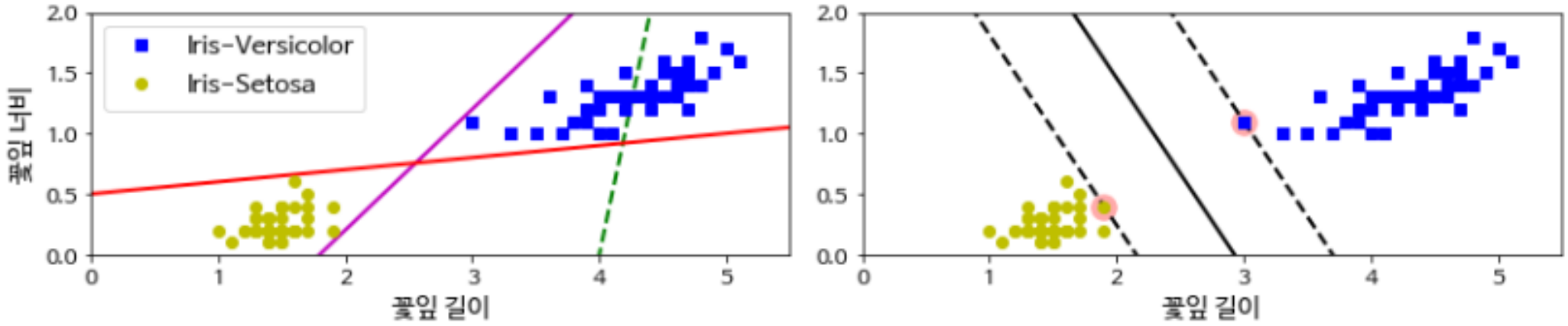
# 1.2 Support Vector?? Margin??



# 2. SVM 분류 방식

- 선형분류, 비선형 분류(다항, 유사도, 가우시안)
- 소프트 마진 분류, 하드 마진
- SVM회귀

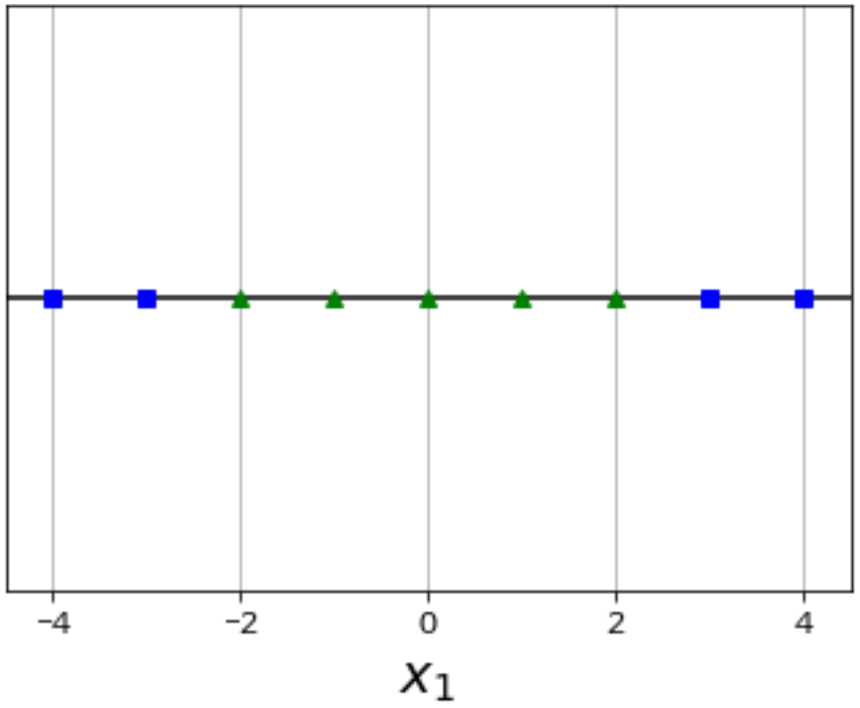
# 2.1-1) 선형 SVM, Large Margin Classification



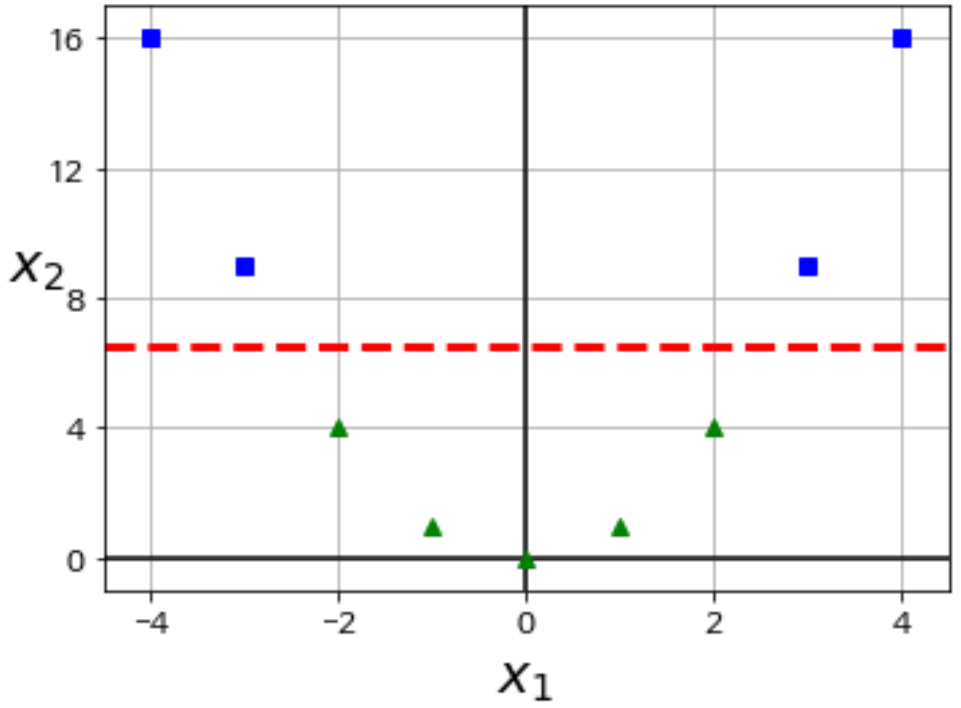
# 2.1-1) 선형 SVM의 민감성 – 특성 스케일


```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
X_scaled = scaler.fit_transform(Xs)
svm_clf.fit(X_scaled, ys)
```

# 2.1-2) 비선형 SVM




특성 X1로만 분류




특성 X1, X2를 사용한 분류

# 2.1-2) 비선형 SVM에서의 특성 추가 1

1. 다항식 특성을 추가(커널트릭)

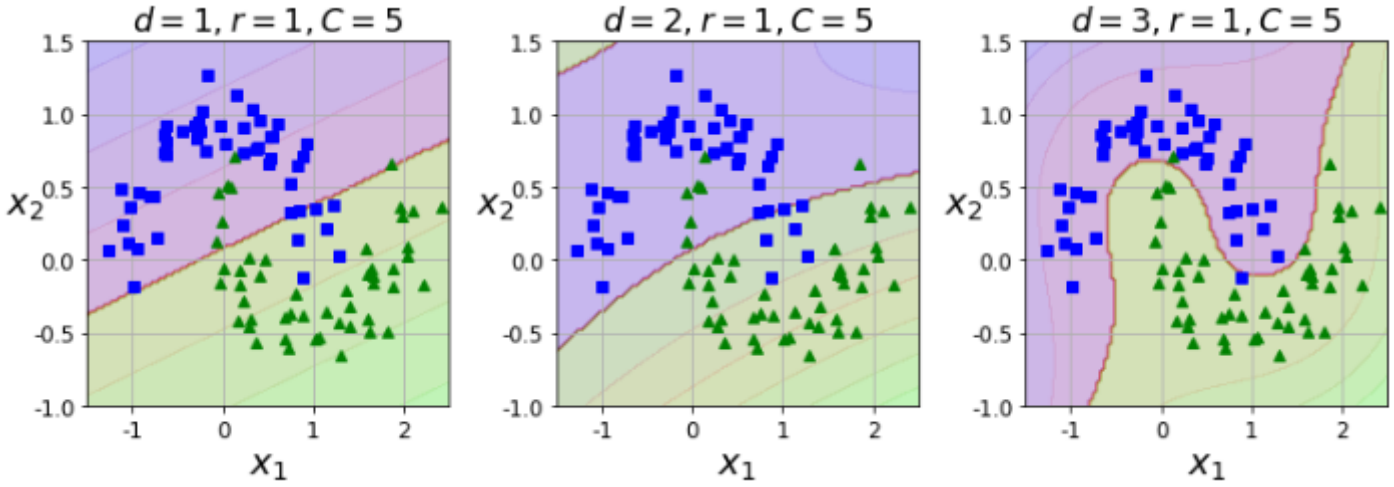
    1. 낮은 차수의 다항식은 복잡한 데이터셋 표현이 힘듦.
    2. 높은 차수의 다항식은 모델을 느리게 만듦.
    3. '커널트릭'을 이용해 특성을 추가하지 않으면서 특성을 많이 추가한 것과 같은 결과를 얻음.

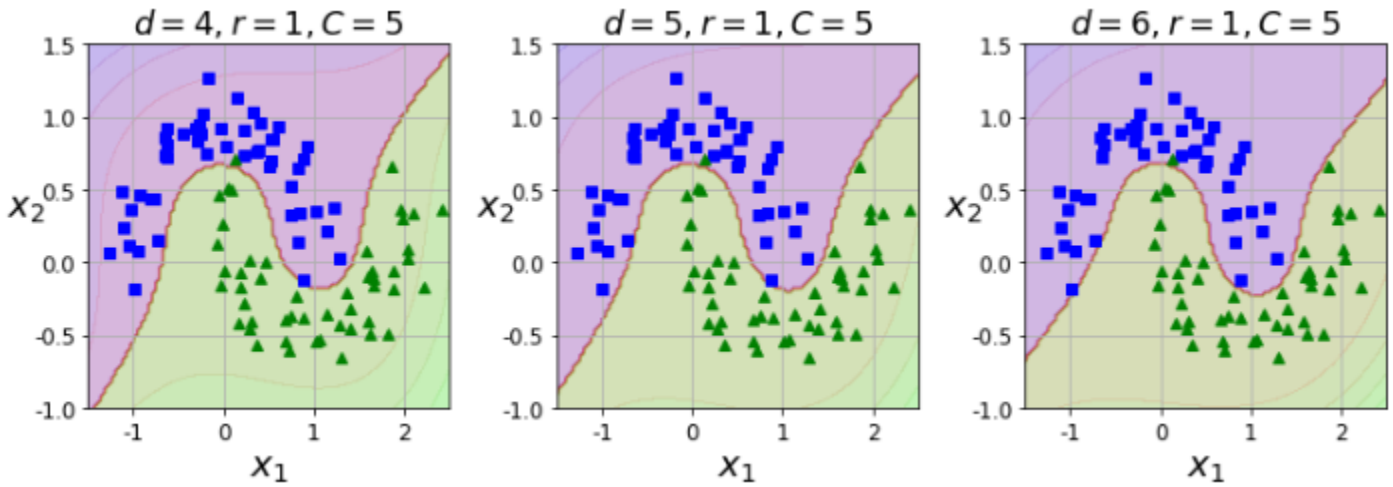
```
from sklearn.svm import SVC

poly_kernel_svm_clf = Pipeline([
        ("scaler", StandardScaler()),
        ("svm_clf", SVC(kernel="poly", degree=3, coef0=1, C=5))
    ])
poly_kernel_svm_clf.fit(X, y)
```

'degree', 'coef0', 'C' 값을 각각 변화시켜 적당한 값을 찾는다.

# 2.1-2) 다항식 특성에 따른 비교(degree)

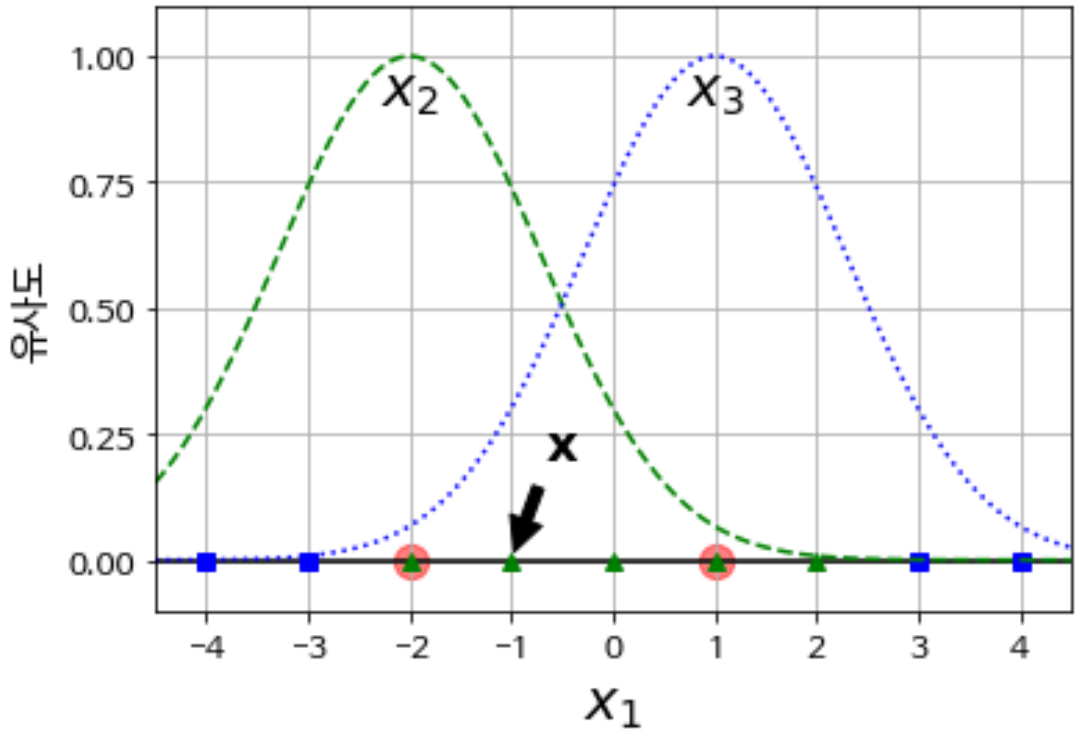


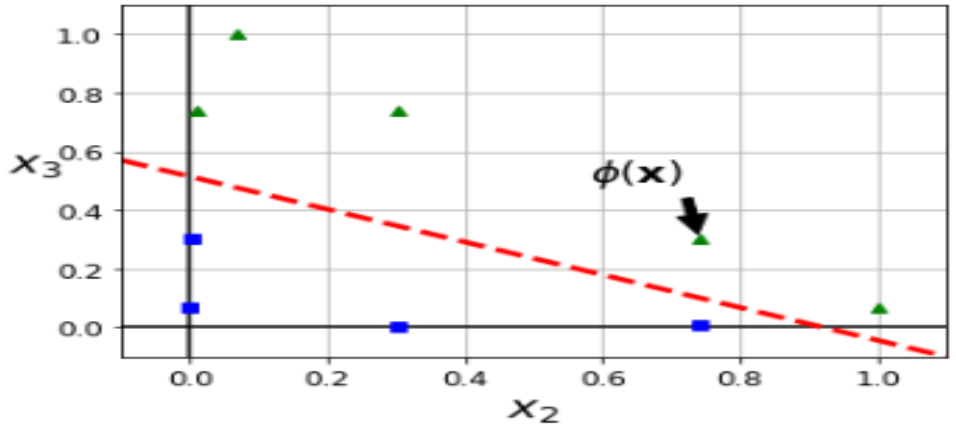
# 2.1-2) 비선형 SVM에서의 특성 추가 2

2. 유사도 특성을 추가

: 각 샘플이 특정 landmark와 얼마나 닮았는지 측정.




- 랜드마크 x1 = -2, x1 = 1을 추가하고 RBF(Radial Basis Funcion)을 유사도 함수로 정의.
- X(X1=-1)을 랜드마크에서 얼마나 떨어져 있는지를 바탕으로 데이터셋을 변경 -> 선형적인 구분이 가능해진다.



# 2.1-2) 비선형 SVM에서의 특성 추가 3

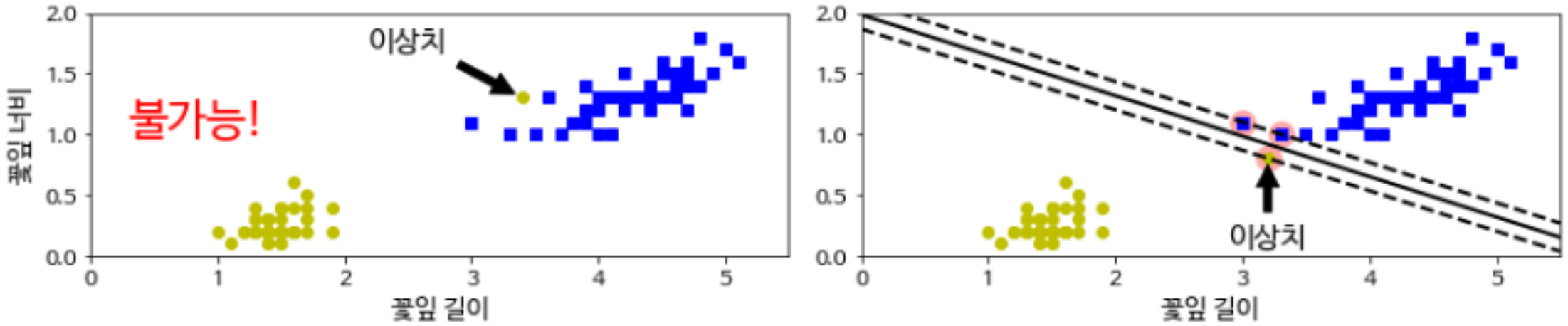
3. 가우시안 RBF

: 유사도 특성 방식도 추가 특성을 모두 계산하려면 연산이 많이 필요하기에 유사도 특성을 많이 추가하는 것과 비슷한 결과를 실제 특성을 추가하지 않고 얻는 방법.

```
rbf_kernel_svm_clf = Pipeline([
        ("scaler", StandardScaler()),
        ("svm_clf", SVC(kernel="rbf", gamma=gamma, C=C))
    ])
rbf_kernel_svm_clf.fit(X, y)
svm_clfs.append(rbf_kernel_svm_clf)
```

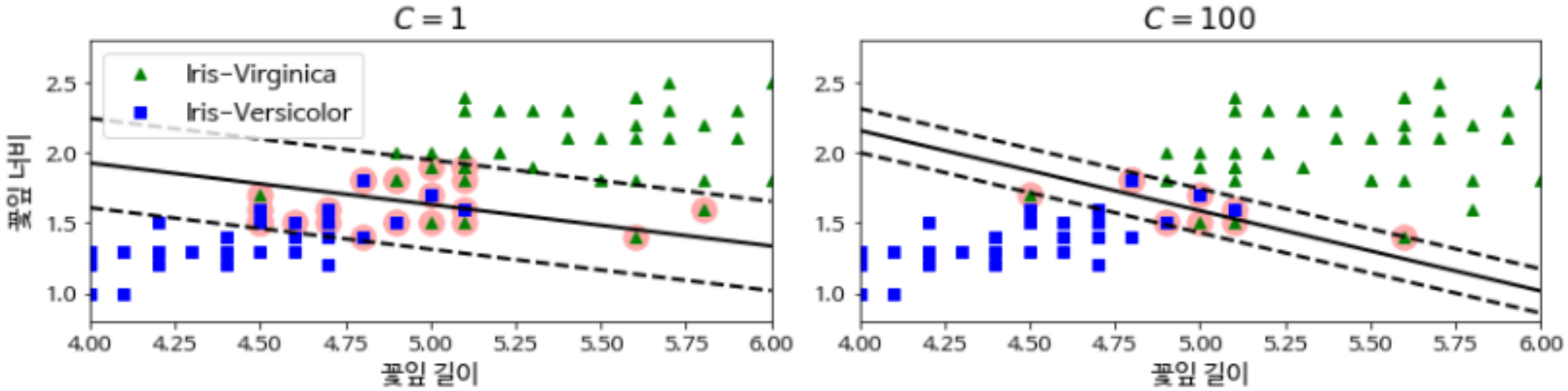
# 2.2-1) 하드 마진

- 하드 마진 분류 : 모든 샘플이 선형으로 올바르게 분류된 경우.
  - 데이터가 선형적으로 구분될 수 있어야 제대로 작동하며 이상치에 민감하다.



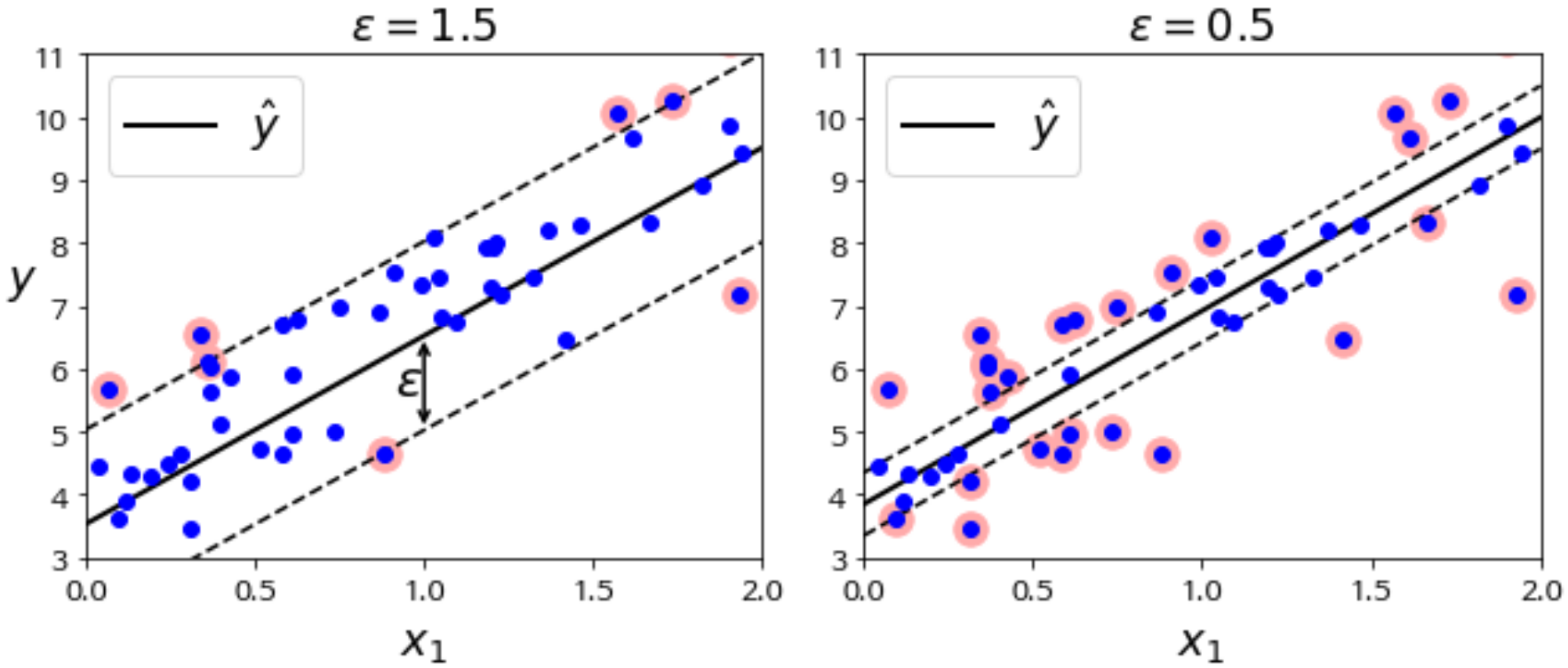
# 2.2-2) 소프트 마진

- 소프트 마진 분류 : 하드 마진의 문제점을 해결하기 위해 margin의 폭과 margin violation 사이의 적절한 균형을 줘서 분류하는 것.



# 2.3 SVM 회귀

- 마진 밖의 샘플들을 제한한 상황에서 마진 안에 최대한 많은 샘플이 들어가도록 학습하는 것.(마진의 폭을 조절)



# 3. SVM의 특징 및 장단점

- 장점
  - 다양한 데이터셋에서 잘 작동.
  - 데이터의 특성이 몇 개 안되더라도 복잡한 경계를 만들 수 있음.
  - 저차원과 고차원(특성이 적을때 많을때) 모두 잘 작동.

- 단점
  - 샘플이 많아지면 잘 맞지 않고 속도도 나오지 않음.
  - 데이터의 전처리와 매개변수 설정에 신경을 많이 써야함.
  - 분석이 어려움.

# Q&A