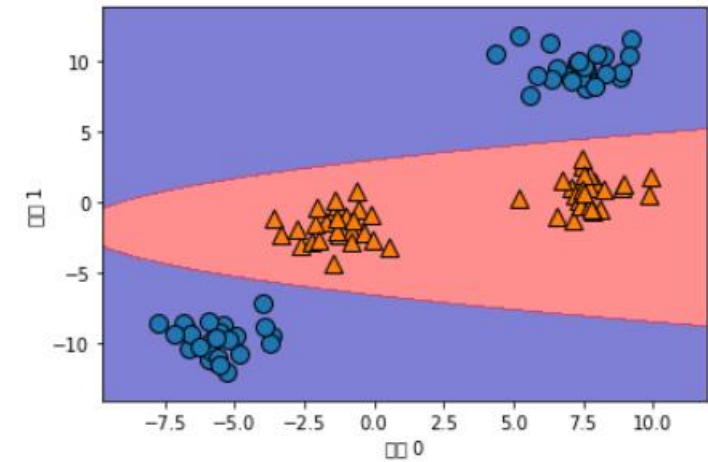
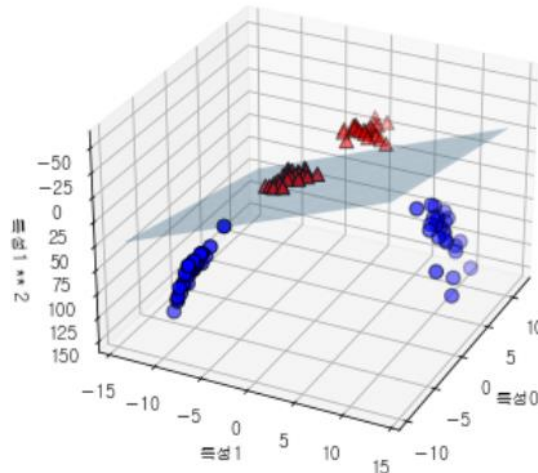
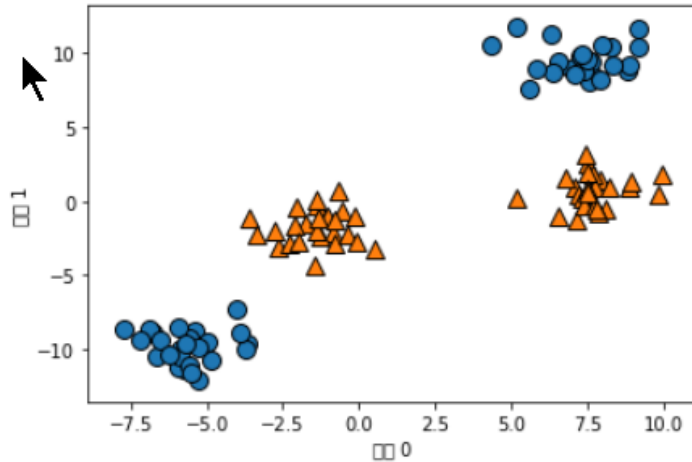


## 3.2 kernel SVM(Support Vector Machine)

- 직선과 초평면은 비선형성이 있는 데이터셋을 다루지 못함(cell 83, 84).
- 비선형성이 있는 데이터셋을 선형모델로 변환하는 방법 :
  - 특성들끼리의 곱/거듭제곱 등으로 특성을 확장
- 확장된 데이터셋에서는 선형 구분이 가능해짐.
- 확장된 데이터셋에서의 구분경계는 축소된 데이터셋에서는 비선형임(cell 84 ~ 86).

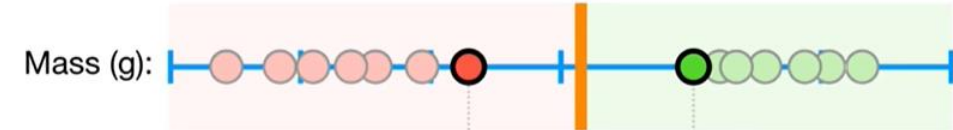


## 3.2 SVM(Support Vector Machine)

- Support Vector : 경계를 결정짓는데 직접 관여하는 데이터포인트
- margin을 최대화하는 샘플.

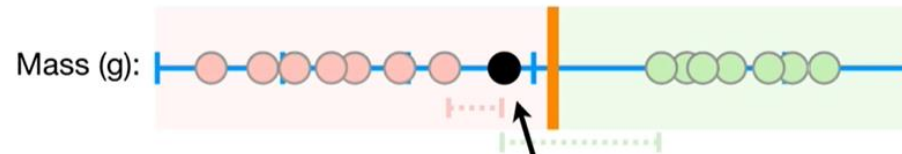
Mass (g):

...and the **green dots** represent mice are **obese**.



...the distances between the observations and the threshold are the same and both reflect the **margin**.

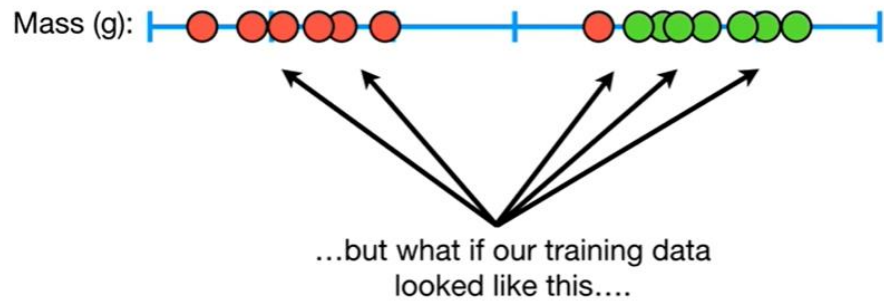
새로운 샘플 판정 : not obese



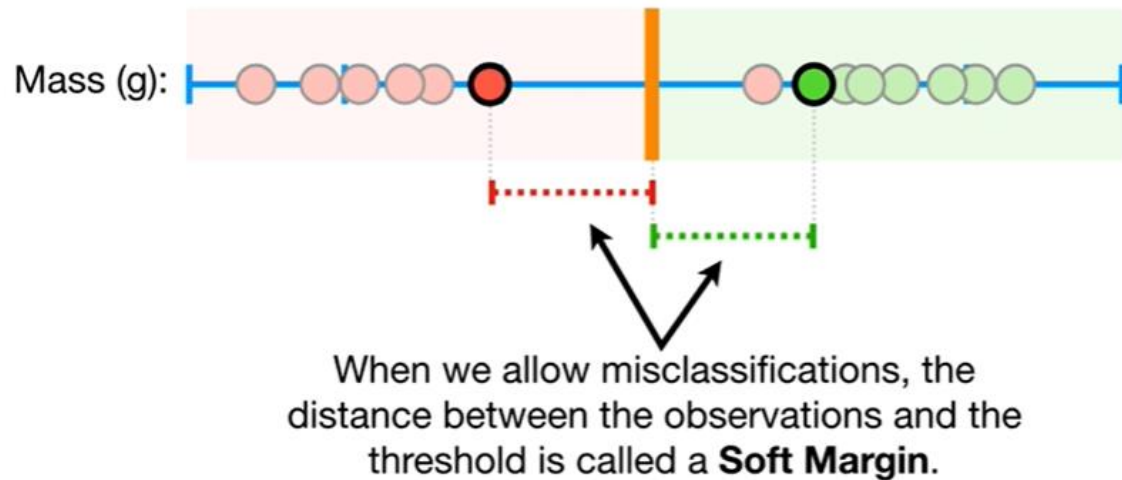
So it makes sense to classify this new observation as **not obese**.

## 3.2 SVM(Support Vector Machine)

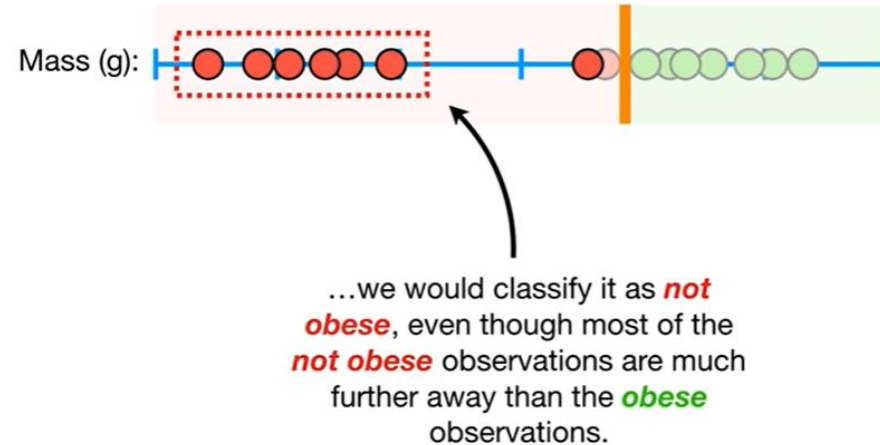
Outlier 문제



오분류를 허용하여 경계선 설정 (C 인수로 조절 가능)

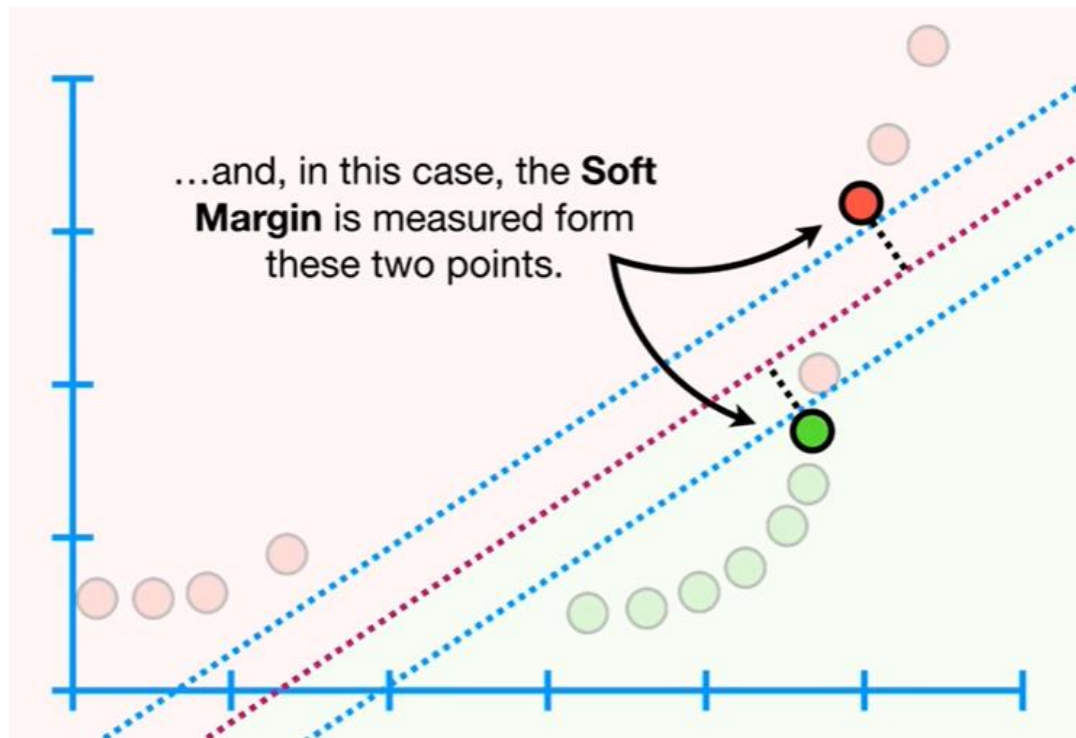


Outlier 샘플로 인해 잘못된 경계선 설정

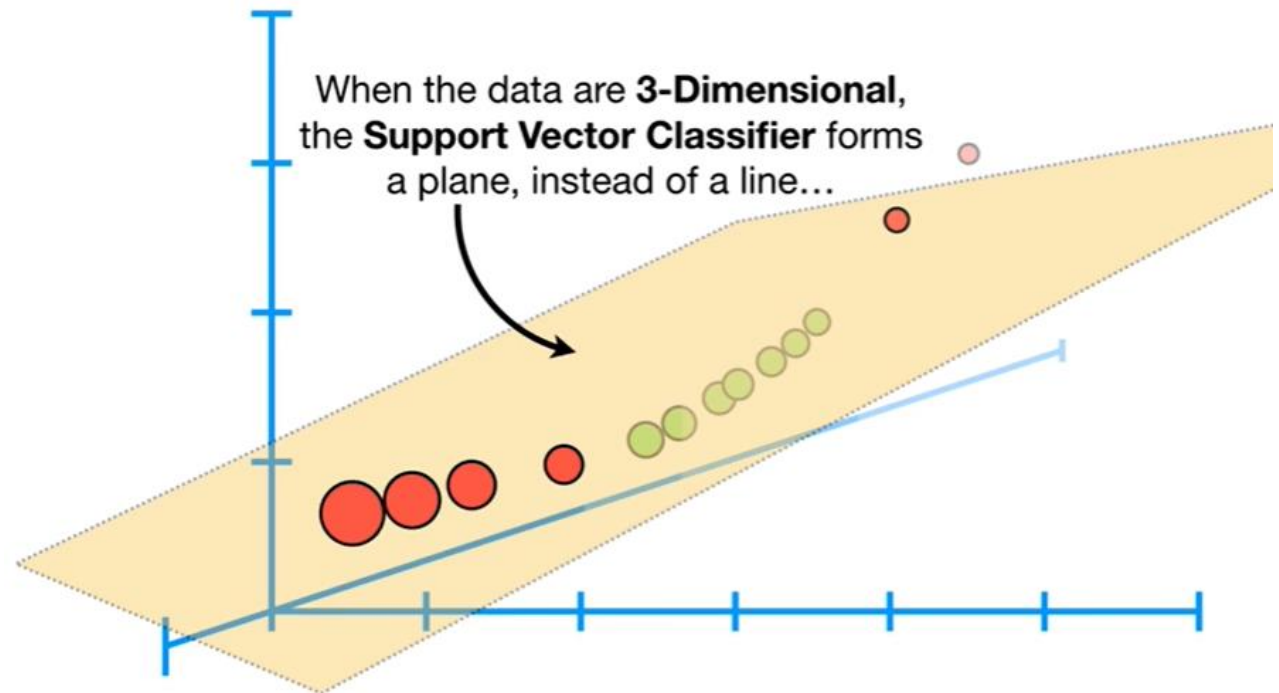


## 3.2 SVM(Support Vector Machine)

특성이 2차원인 경우 마진을 최대로 하는 경계선 설정



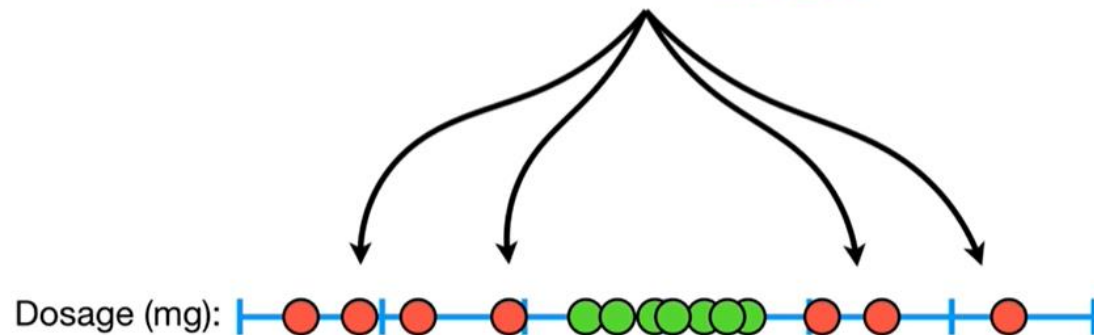
특성이 3차원인 경우 마진을 최대로 하는 경계면 설정



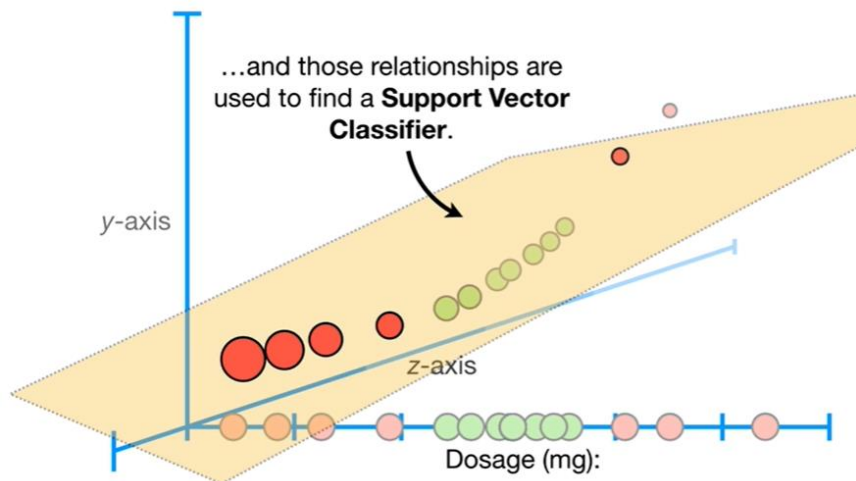
## 3.2 kernelSVM(Support Vector Machine)

선형으로 분류할 수 없는 데이터  
→ Support Vector Machines

...and the **red dots** represent patients that were **not cured**...

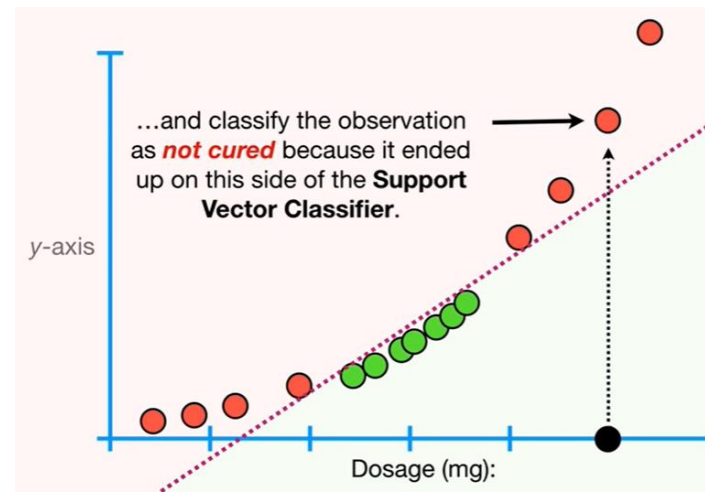


Polynomial kernel( $d=3$ )을 적용하여 1차원 데이터를 3차원 데이터로 변환한 경계선



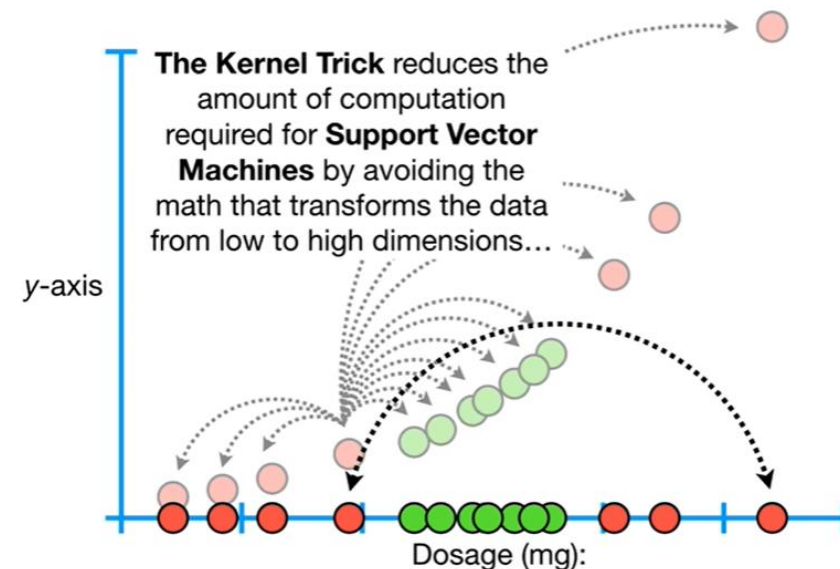
Polynomial kernel( $d=2$ )을 적용하여 1차원 데이터를 2차원 데이터로 변환

**y-axis value = Dosage<sup>2</sup>**



## 3.2 kernel SVM(Support Vector Machine)

- 특성추가 방법의 문제점 : 추가할 특성을 알기 어려움. 기본특성이 많을 경우 연산비용이 많이 소요됨.
  - 차수가 낮은 경우 : 복잡한 유형의 데이터를 잘 표현하지 못함
  - 차수가 높은 경우 : 특성이 많아지므로 분석 수행에 많은 시간 소요
- 커널기법
  - 단순한 초평면으로 정의할 수 없는 입력데이터를 처리할 수 있도록 더 복잡한 모델을 만드는 기법
  - 새로운 특성데이터를 명시적으로 만들지 않고 확장된 특성에 대한 데이터포인트들의 거리만을 계산.
  - 사용할 수 있는 커널 : Mercer's Theorem을 만족하는 함수
  - polynomial kernel(지정된 차수까지의 다항식 계산), 가우시안(rbf) 커널(무한 차수의 다항식 계산), linear 커널, sigmoid 커널.
  - SVM 발표(Vladimir Vapnik, Corinna Cortes, 1995).
  - kernel SVM 발표(2000)



## 3.2 kernelSVM(Support Vector Machine)

---

- 학습 및 매개변수 튜닝(cell 88)
  - gamma: 하나의 훈련샘플이 미치는 영향의 범위.
  - C: 규제 매개변수. 각 포인트의 중요도 제한.
- 암 데이터셋에 적용(cell 89 ~ 93)
  - 기본값( $C = 1$ ,  $\text{gamma} = 1/\text{특성갯수}$ )일 때 0.90, 0.94 달성 → 과소적합.
  - 데이터 전처리 후 다시 학습 → 0.984, 0.972
- 장단점
  - 우수한 성능 보임.
  - 데이터크기가 커지면(10만개 이상) 속도, 자원소요 측면에서 어려움 있음.
  - 데이터 전처리와 매개변수 설정에 민감함.
  - 모델의 해석이 어려움.