SESSION # 05

K-NN Algorithm

By Team 5

CONTENTS

01 K-NN이란?

02 차원의 저주

03 SVM 분류기

04 실습/퀘스트

K-NN이란?

개념 정의
K-NN의 원리
장단점

당신은 내가 다음 대선 때 어떤 후보를 뽑을지 알 수 있는가?

내가 야당을 선호하는 지역에 살고 있다는 것을 안다면?
그런데, 지역 뿐만 아니라 나이 소득수준, 자녀 수까지 안다면?

01 K-NN이란? 개념정의

플레이 메이커, 타고난 센스, 인기 스타

크리스티아누 호날두
포르투갈

그라운드의 지휘관, 중원의 지배자, 패스 마스터

프랭크 램퍼드
잉글랜드

공격수, 득점 기계, 팀의 해결사

호나우두
브라질

팀의 에이스

마라도나
아르헨티나

날쎈 돌이, 달리기 선수

아르옌 로번
네덜란드

나의 축구 등번호는?

나는 그라운드의 장군, 중원의 왕자, 패스는 잘하는 편

K-NN 끝

K-Nearest Neighbors (K-근접이웃방법)

K-NN 알고리즘은 머신러닝 알고리즘 중의 하나
지도학습인 '분류' 와 '예측' 을 위한 알고리즘

가까운 거리에 가장 많이 분포하는 것으로 나의 데이터를 분류 하는 알고리즘

K-Nearest Neighbors (K-근접이웃방법)

주어진 x_0에서 가장 가까운 k개의 관측치들로 x_0에서 각 법주의 확률을 추정하고
가장 많이 나온 범주로 추정한다.

따라서 필요한 것은

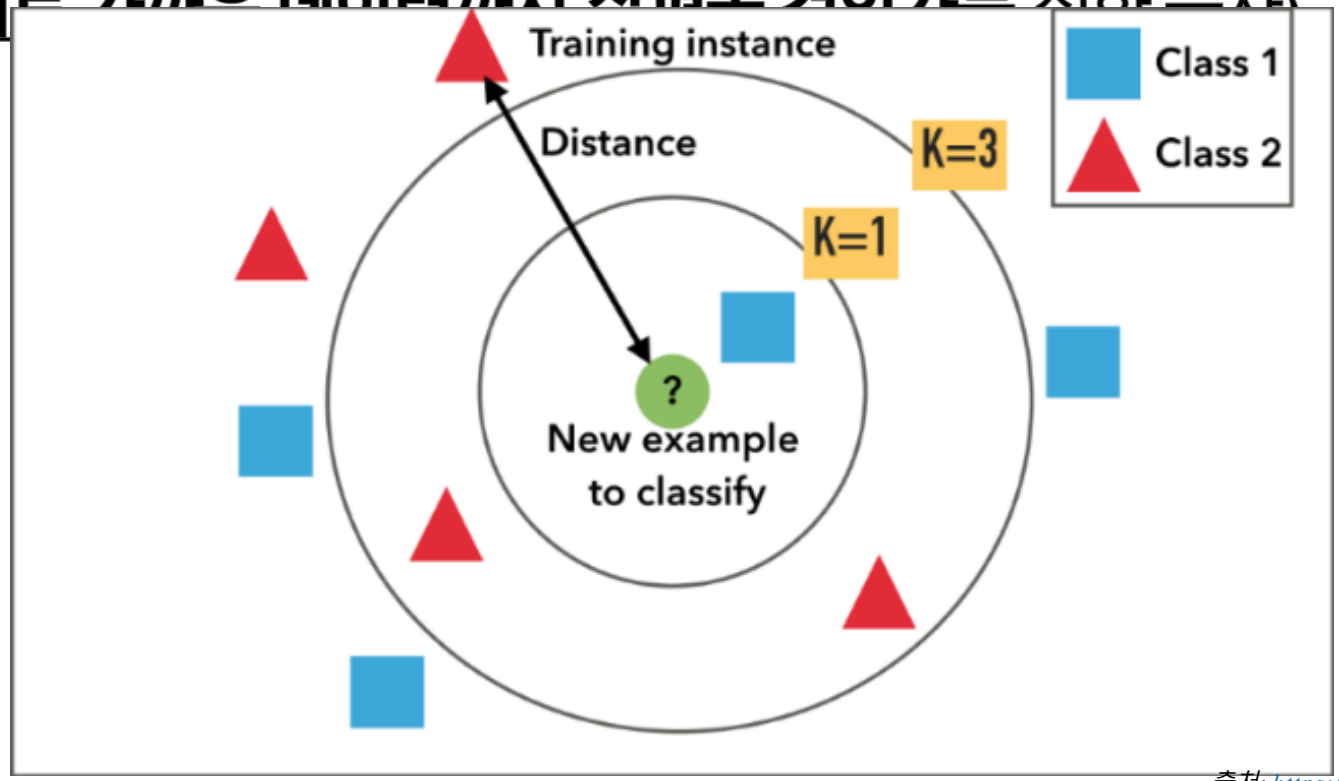
- 거리를 재는 방법
- 어느 정도 가까운 거리로 할 것인가?
- 대전제: "서로 가까운 점들은 유사하다"

출처: http://kkokkilkon.tistory.com/14

출처: http://hamait.tistory.com/843

새로 들어온 ?은 ■ 그룹의 데이터와 가장 가까우니 ?은 ■ 그룹이다.

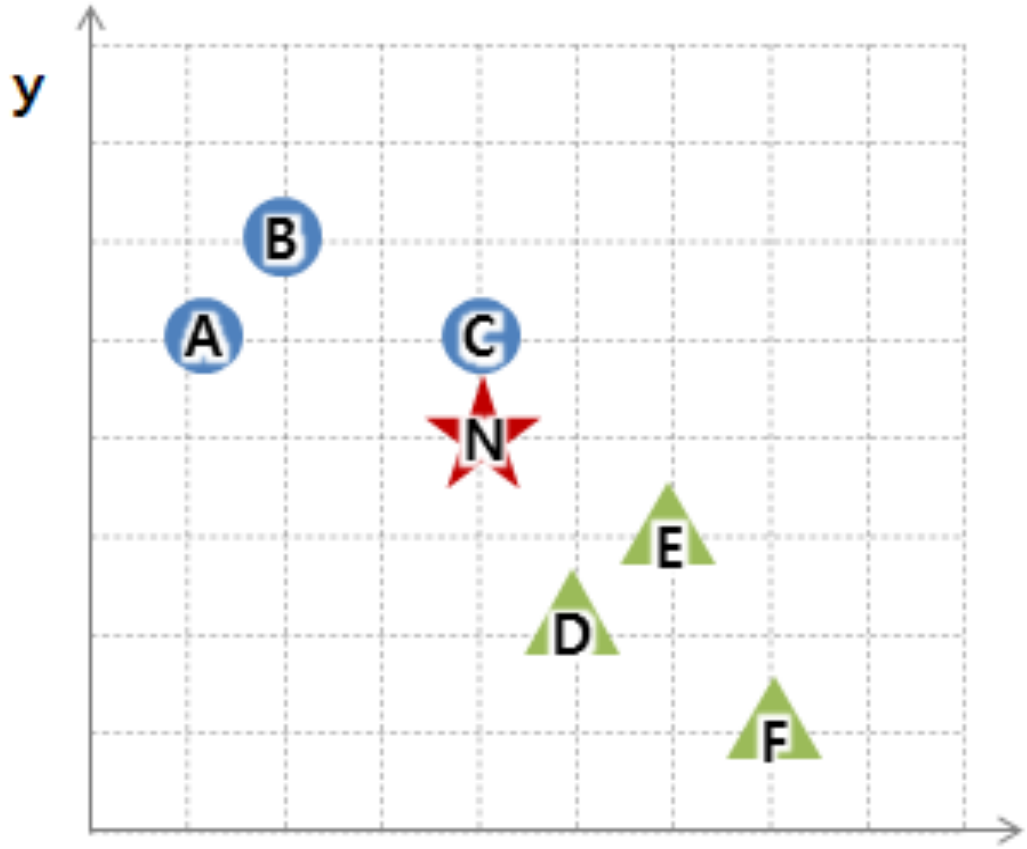
(k는 몇 번째로 가까운 데이터까지 살펴볼 것인가를 정하는 숫자)



출처: https://arifuzzamanfaisal.com/k-nearest-neighbor-regression/

K-NN의 원리(6개의 기존 데이터 A~F와 1개의 신규 데이터 N)

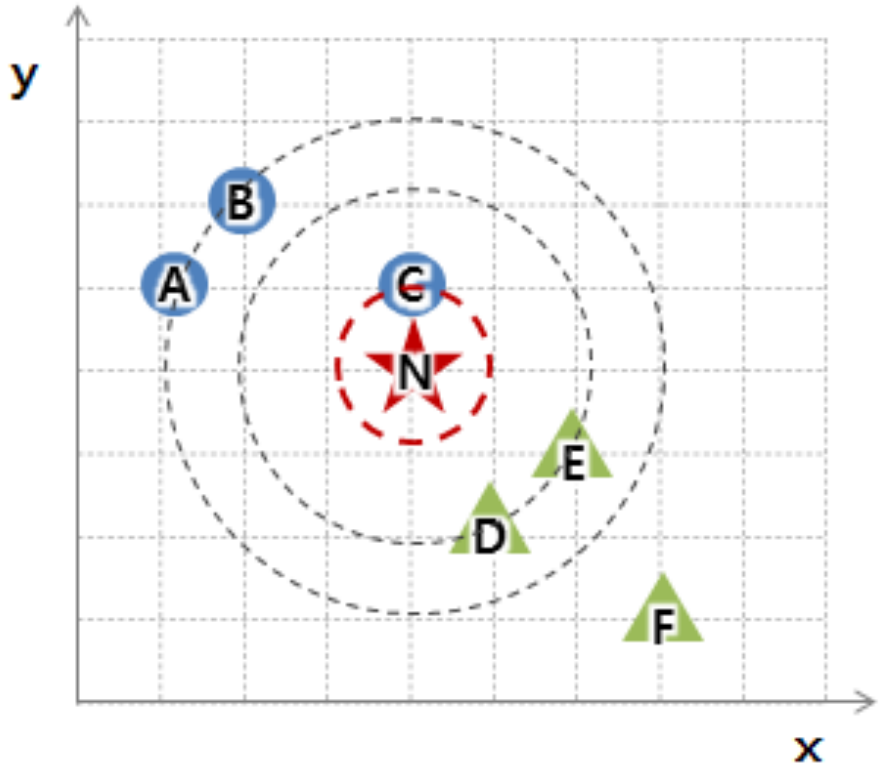
데이터	x좌표	y좌표	그룹
A	1	5	●
B	2	6	●
C	4	5	●
D	5	2	▲
E	6	3	▲
F	7	1	▲
N	4	4	?



출처: http://kkokkilkon.tistory.com/14

출처: http://kkokkilkon.tistory.com/14

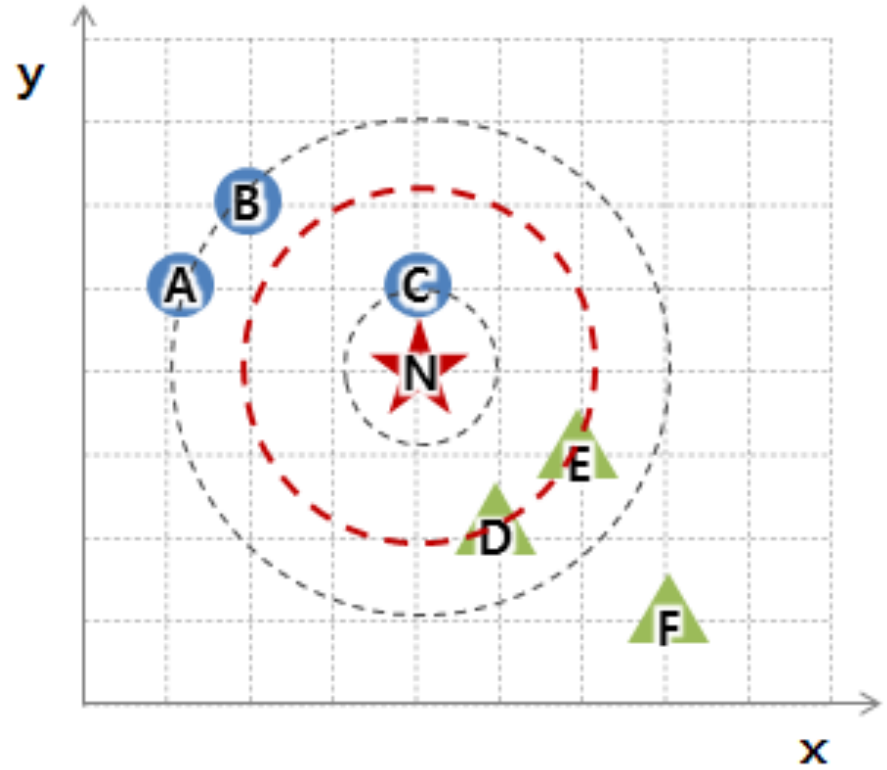
K=1 이라면,



거리가 1번째로 가까운 C만을 보고 신규 데이터를 분류. 따라서 N은 C와 같은 그룹인 ●로 분류된다.

출처: http://kkokkilkon.tistory.com/14

K=3 이라면,



거리가 3번째로 가까운 C, D, E까지 보고 신규 데이터를 분류 **이때 그룹이 갈리면 다수결의 원칙에 따른다.**
여기서는 1 : 2가 되어 N은 ▲로 분류된다.

출처: http://kkokkilkon.tistory.com/14

K-NN 특징

장점

- 개념이 단순하다.
- 비모수 방법으로써 파라미터에 대한 가정이 거의 없다.
- 일관성 있는 결과를 도출 한다.
- 이상치에 둔감하다.

단점

- 계산시간이 많이 걸린다.
- 특정 현상의 원인을 파악하는데 큰 도움이 안된다.

나. 알고리즘의 특징	
특징	설명
최고인접 다수결	기존 데이터 중 가장 유사한 K개의 데이터를 측정하여 분류
유사도(거리) 기반	유클리디언거리, 마할라노비스거리, 코사인 유사도등 활용
Lazy learning 기법	새로운 입력 값이 들어온 후 분류 시작
단순유연성	모형이 단순하며 파라미터의 가정이 거의 없음

출처: http://blog.naver.com/santalsm/220526205088

가. KNN 알고리즘의 장점	
장점	설명
학습 간단	- 모형이 단순하고 구현이 쉬움 - 모수 (parameter) 및 데이터에 대한 가정이 거의 없음
유연한 경계	- 거리의 변형, 가중치 적용이 용이함 - 유클리디언, 코사인유사도, 가중치 적용, 정규화 적용 용이
모델의 유연성	- 데이터에 대한 가정을 반영하여 변형하기에 간편 - 변형한 데이터의 training data set 기반 분류기 검증 용이
높은 정확도	- 사례기반(instance based) 으로 높은 정확성 - 훈련 데이터 클 수록 클러스터 매칭의 정확성 좋아짐

출처: http://blog.naver.com/santalsm/220526205088

나. KNN 알고리즘의 단점	
단점	설명
모수 k 선정 어려움	- K 수에 따라 알고리즘의 성능을 좌우하는 어려운 문제 - under/over fitting 의 trade off 문제 발생 요인
공간 예측 부정확	- 공간정보 예측모델에서는 특정 이벤트의 발생이 일정하지 않고, 영향변수 많아 적용이 어려움
거리계산 복잡성	- 모든 데이터와의 유사도, 거리 측정 수행 필요 - 명목변수 및 결측치를 따로 처리 필요 - 기존데이터의 실측값, instance 에 크게 의존
고 비용	- 모든 데이터를 메모리 기반 연산, 거리측정 필요 - 데이터 커질 수록 메모리 및 연산시간 증가 문제
노이즈에 약함	- 노이즈로인해 큰 K 설정을 필요로 함 - 민감하고 작은 테이터 무시되는 under fitting 문제 야기

출처: http://blog.naver.com/santalsm/220526205088

01 K-NN이란?
개념정의

KNN (K-Nearest Neighbor) 알고리즘의 동작원리 상세	
동작원리	동작원리 상세
Fingerprint 확인	- 새로운 입력값 확인 - 가까운 데이터는 같은Label(클러스터) 가능성 큼 - 기존의 모든 데이터와 새로운 Fingerprint 비교 준비
명목변수기반 그룹 분류	- 기존의 저장 되어있는 데이터 셋의label 화 -서로다른 범주 데이터를 정규화 수행 - 분류기 검사 수행,예시 데이터의 90%를 훈련데이터 10%를 테스트로 활용
거리측정	-유클리디언 거리, - 새로운 Fingerprint (0,0)와 기존 data(1,2) 간의거리 예시 - 메모리 기반 Fingerprint와 모든 데이터간의 거리계산 - 계산된 거리의 정렬 수행
K 선정	- 양의정수 값, 정렬된 거리 중가장 가까운 k개 데이터 선정 - 여러k 값을 모델링 후가장 성능이 좋은 k값선정 - 노이즈 클수록 큰k 값선정이 좋음 - 큰K는 노이즈에 좋지만, 작고 중요한 패턴을 무시 가능 - 작은k는 극단값 및노이즈를 허용하여 클러스터링 오류가능
클러스터 매칭	- 명목데이터경우, Majority voting 기반의 클러스터 매칭 수행, k개 데이터가 많이 속해있는 클러스터로 새로운 값을 분류 - 수치형 데이터 경우 k개데이터의 평균 (또는 가중평균)을 이용하여 클러스터 매칭

출처: http://blog.naver.com/santalsm/220526205088

차원의 저주

개념 정의

KNN에서 차원의 저주

결론

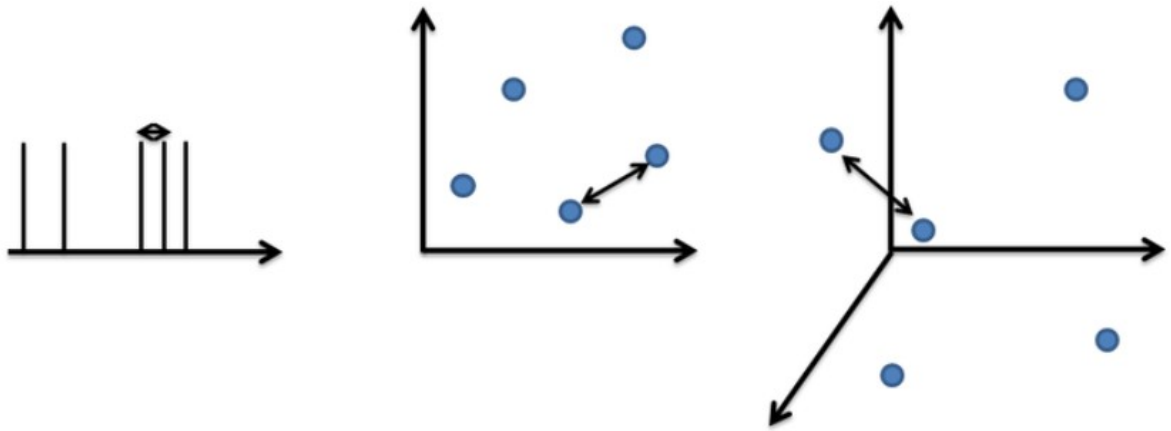
차원의 저주(Curse of Dimensionality)

차원이 증가하면 그것을 표현하기 위한 데이터 양이 기하급수적으로 증가하며
그로 인해 신뢰도가 낮아지고
러닝 타임이 길어지고 정확도가 크게 감소한다.

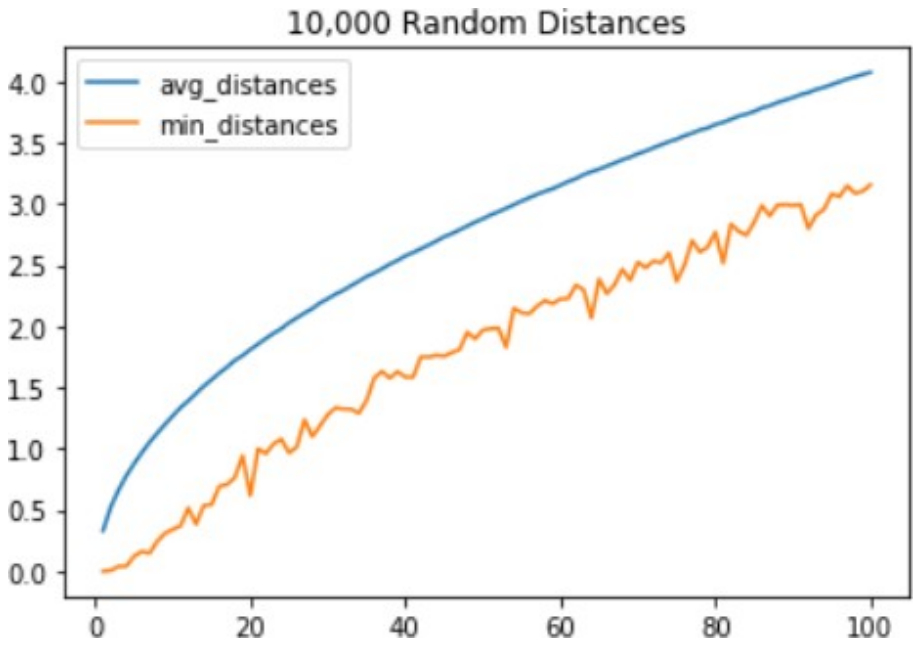
- Rechard E Bellman
- 데이터 시각화 세션 中

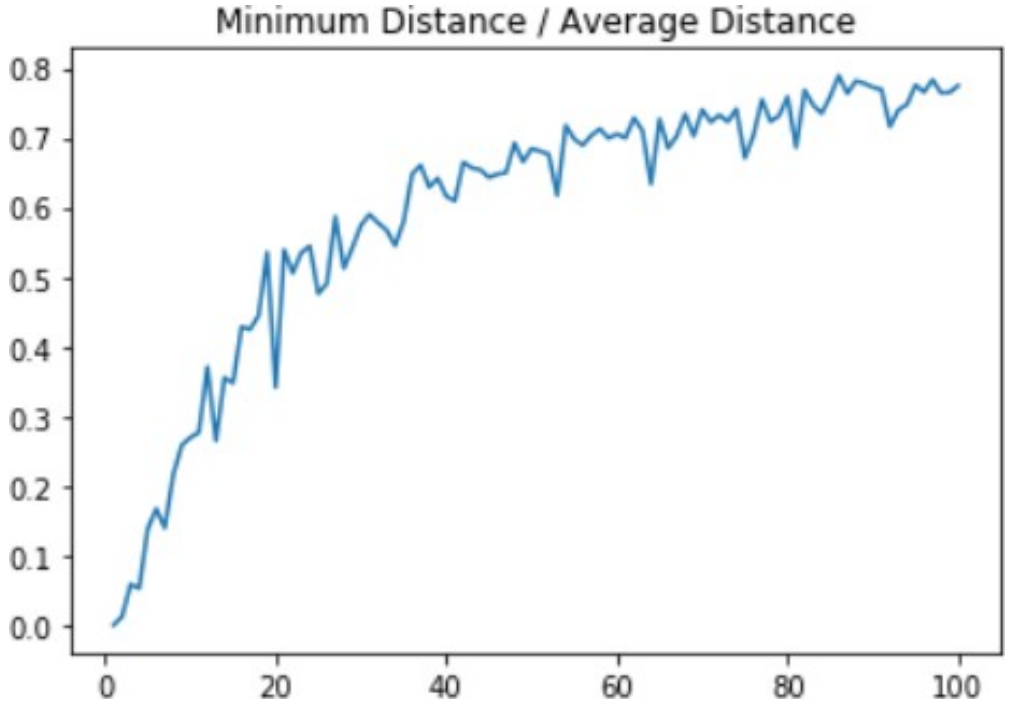
K-NN 알고리즘에서 차원의 저주

차원이 커지면, 점의 개수가 같아도 점과 점 **사이의** 최소 **거리가 점점 길어져** 평균 **길이와 비슷해 진다.**
최소 길이와 평균길이가 **비슷해 지면 K-근접이웃방법에서** '이웃'의 **의미가** 없어지게 **된다.**
즉 차원이 **커질** 수록 최소 **길이의 의미가** 없어져 이웃 점에 대한 신뢰성이 **떨어지게 된다.**

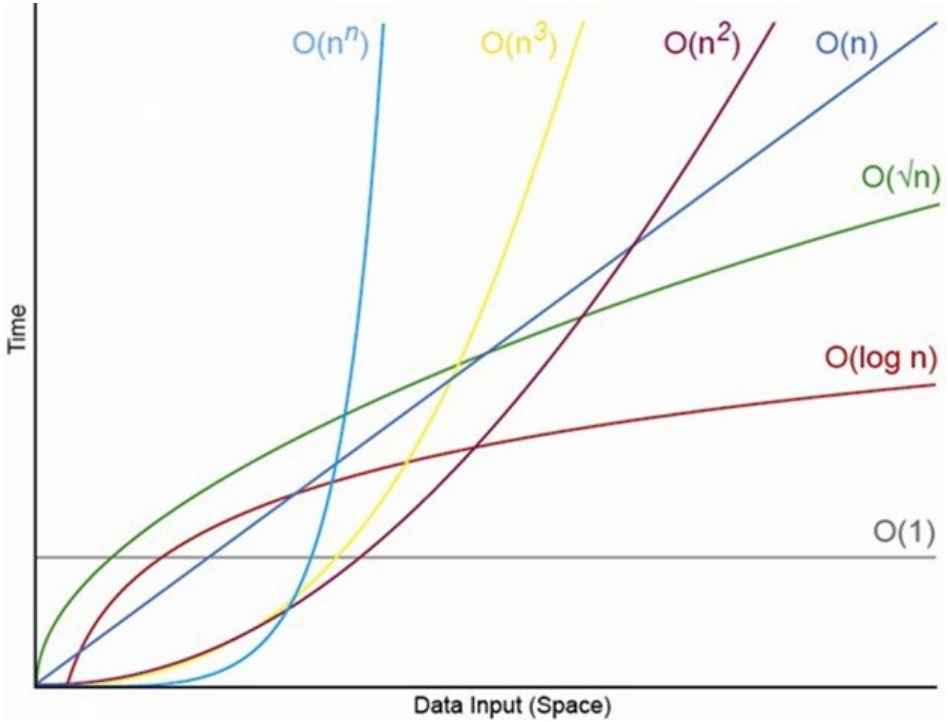


K-NN 알고리즘에서 차원의 저주





Big O Notation



Big O란 알고리즘의 성능과 복잡성을 설명해 준다.

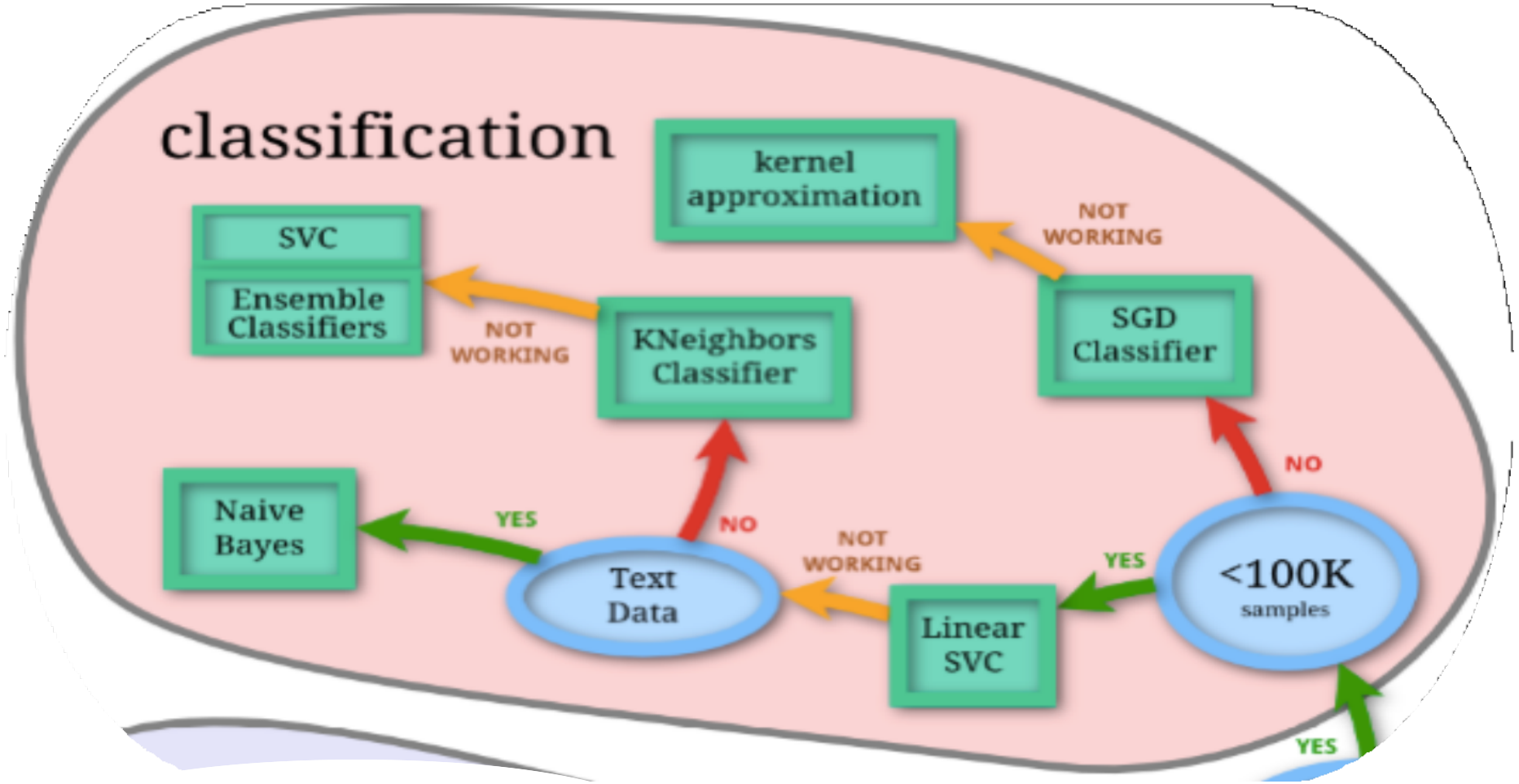
결론

흔히 Feature 또는 Dimension을 늘리는 것은 쉽다.
하지만 알고리즘 (특히 K-NN 알고리즘와 같은 lazy learning 알고리즘)에는 전혀 도움이 안된다!

SVM 분류기

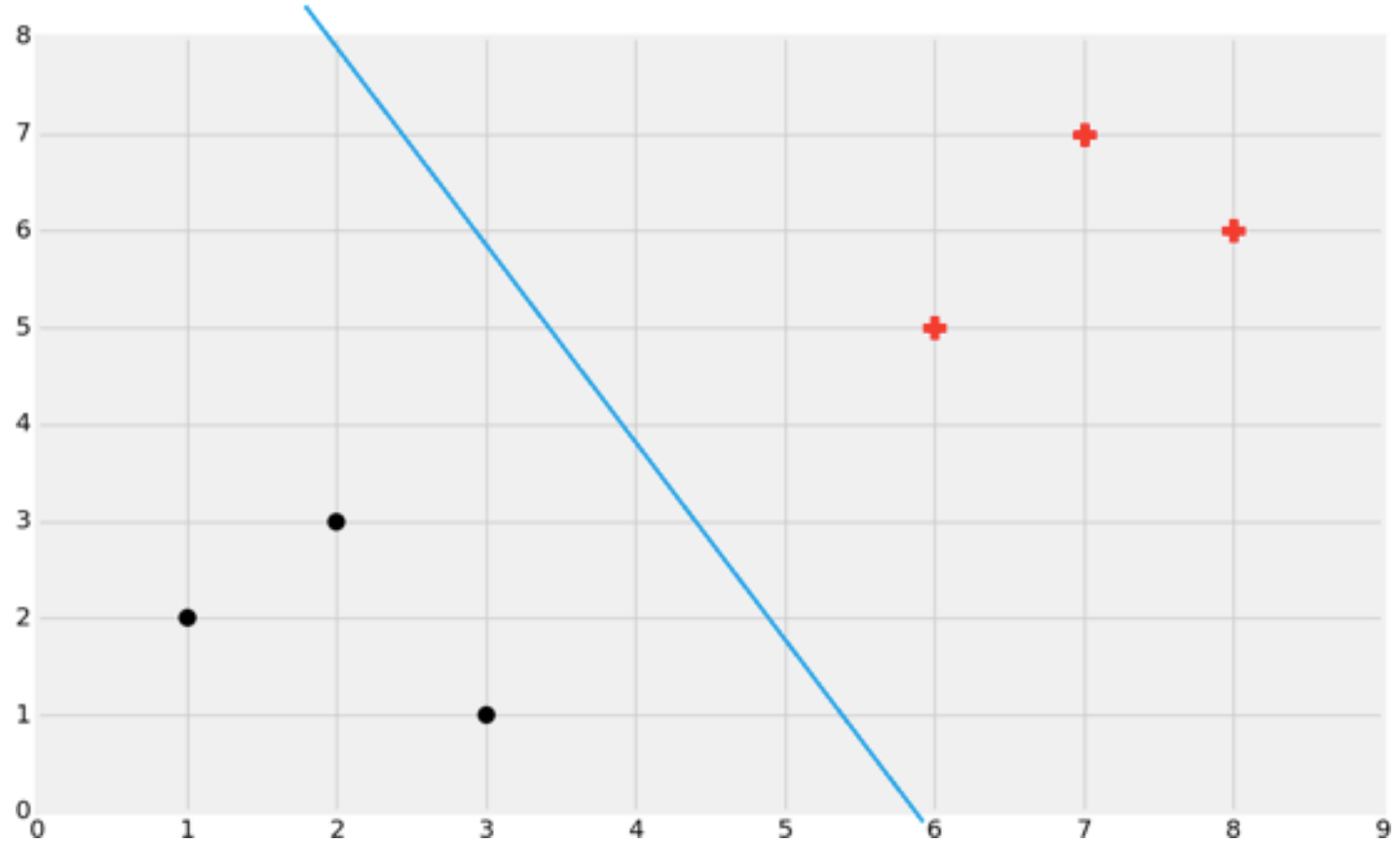
SVM Intuition
Mathematical Background
SVM on ScikitLearn

Why SVM?



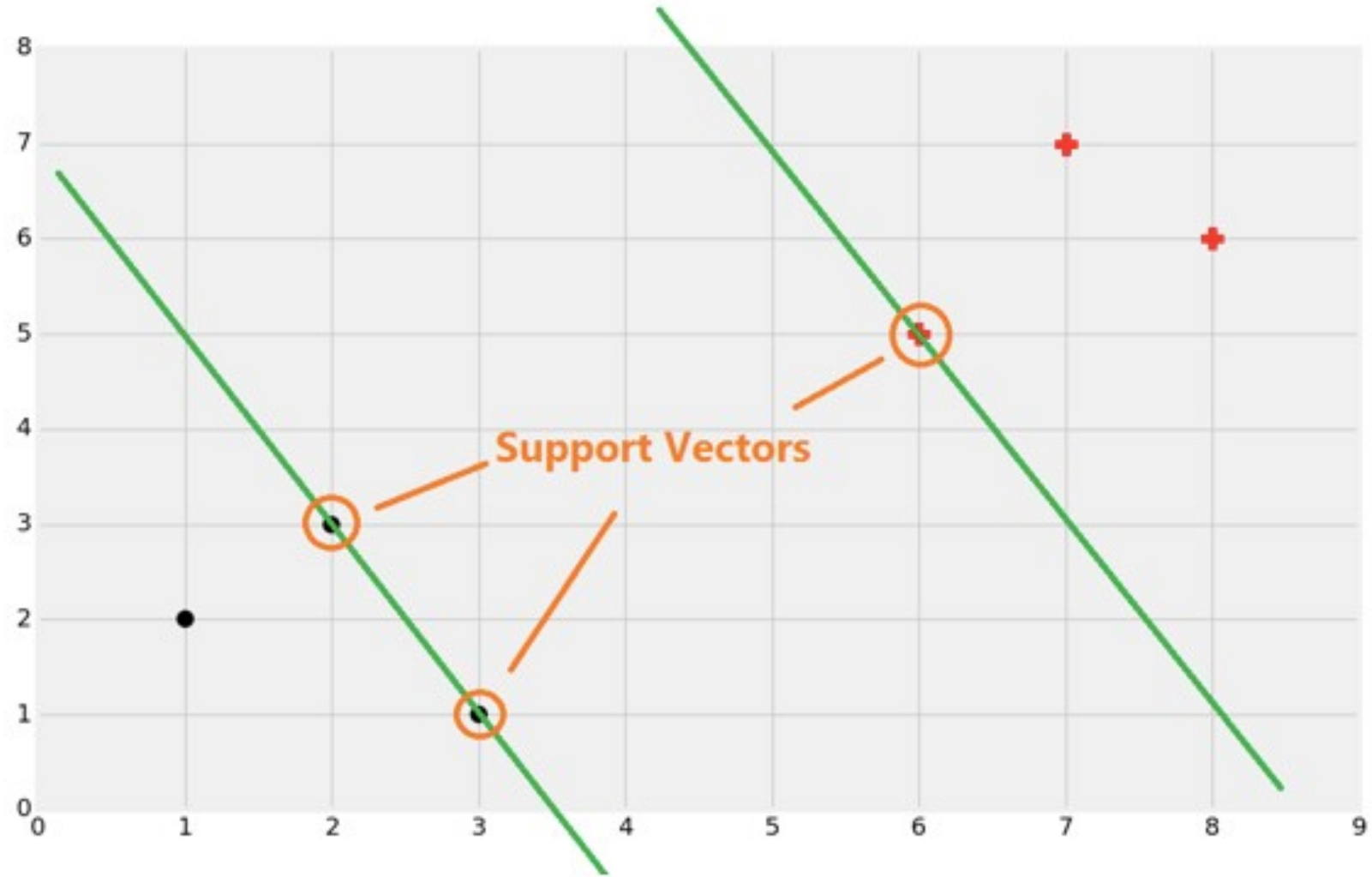
출처 : http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

목표 : Best fitting line 찾기

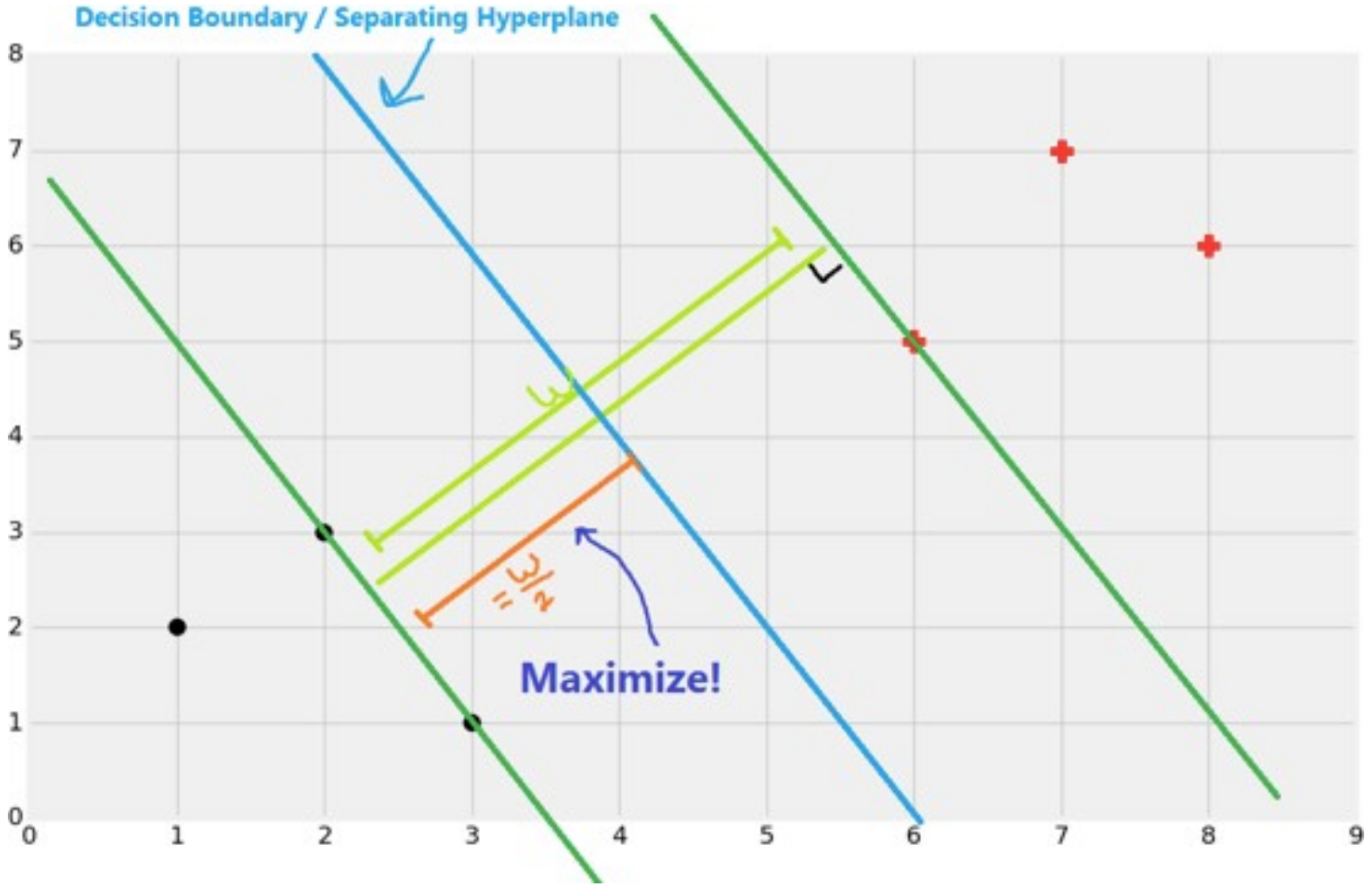


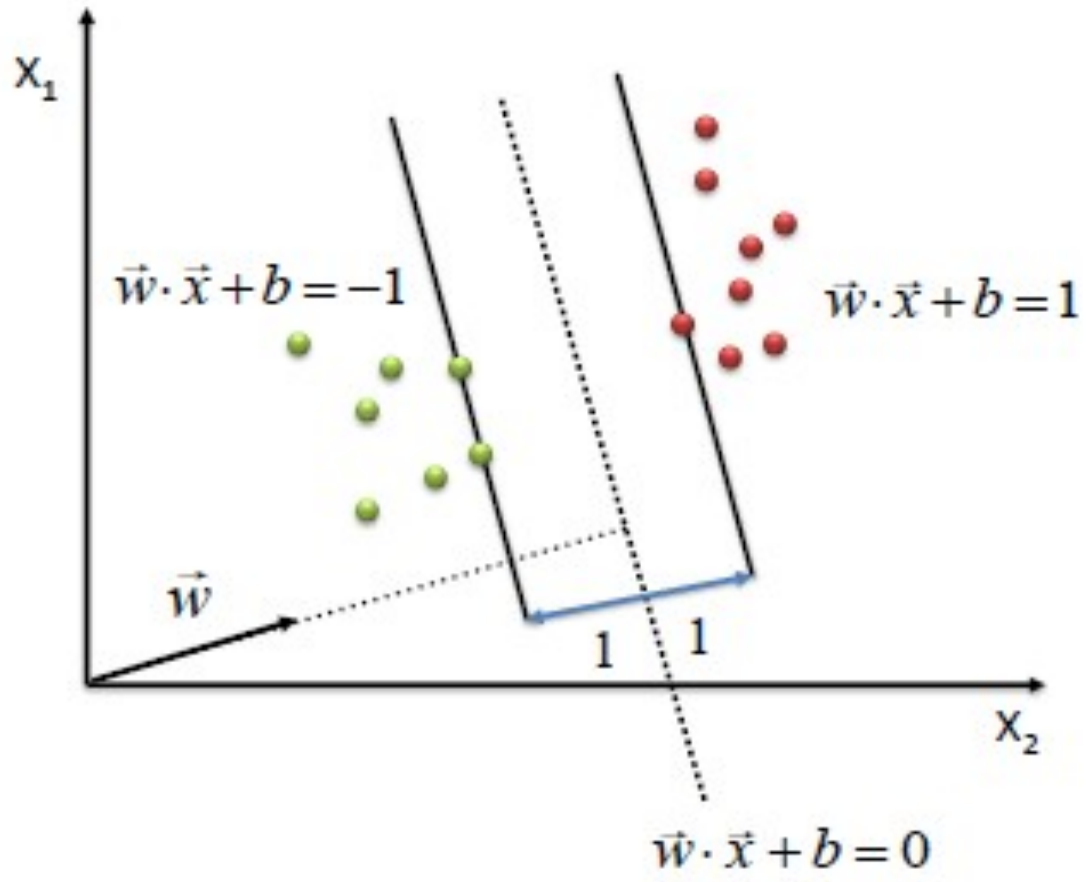
SVM : Intuition

How?



How?



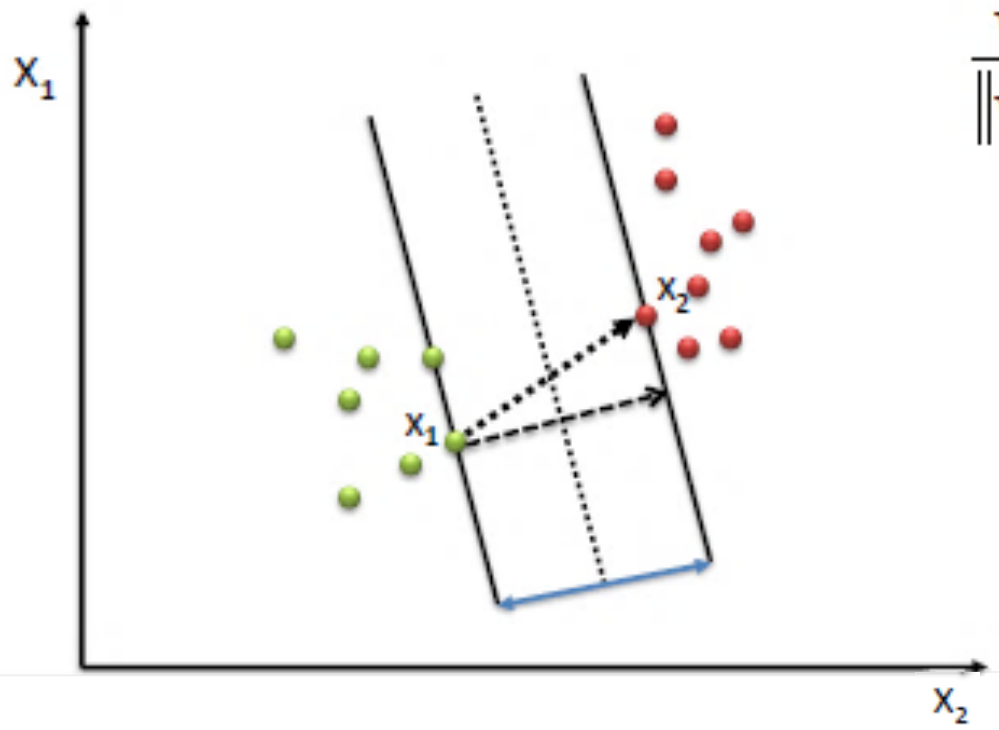


$$\max \frac{2}{\|w\|}$$

s.t.

$(w\cdot x+b)\geq 1, \forall x$ of class 1

$(w\cdot x+b)\leq -1, \forall x$ of class 2



$$\frac{w}{\|w\|}\cdot(x_2-x_1)=\text{width}=\frac{2}{\|w\|}$$

$$w\cdot x_2+b=1$$

$$w\cdot x_1+b=-1$$

$$w\cdot x_2+b-w\cdot x_1-b=1-(-1)$$

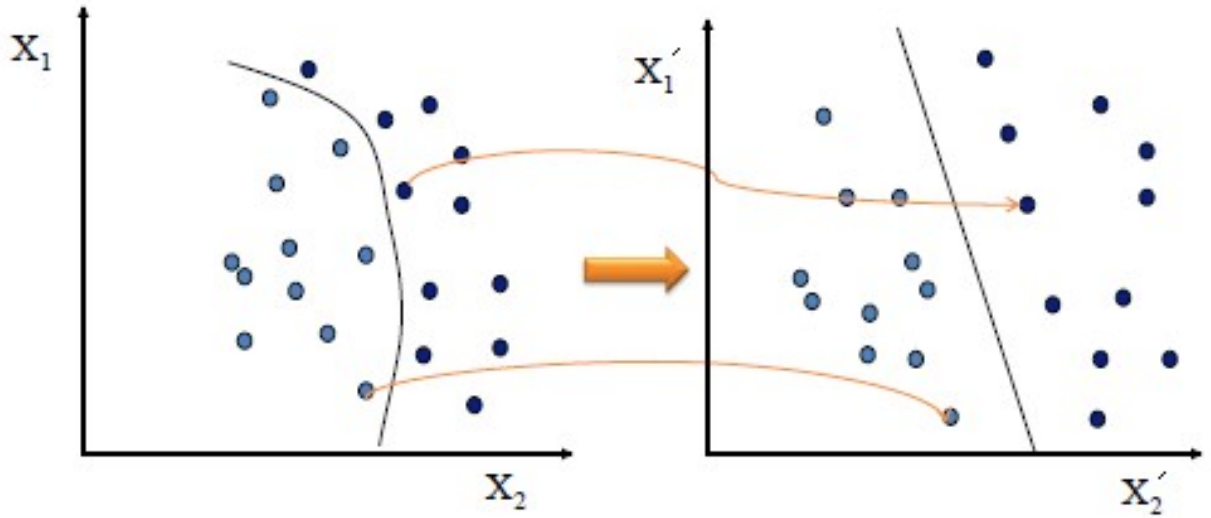
$$w\cdot x_2-w\cdot x_1=2$$

$$\frac{w}{\|w\|}(x_2-x_1)=\frac{2}{\|w\|}$$

$$\max_\alpha \quad W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle.$$

$$\text{s.t.} \quad \alpha_i \geq 0, \quad i = 1, \ldots, m$$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0,$$



Linear SVM

$$x_i \cdot x_j$$

Non-linear SVM

$$\phi(x_i) \cdot \phi(x_j)$$

Kernel function

$$k(x_i \cdot x_j)$$

Polynomial

$$k(\mathrm{x}_i, \mathrm{x}_j) = (\mathrm{x}_i . \mathrm{x}_j)^d$$

Gaussian Radial Basis function

$$k(\mathrm{x}_i, \mathrm{x}_j) = \exp\left(-\frac{\|\mathrm{x}_i - \mathrm{x}_j\|^2}{2\sigma^2}\right)$$

Sckit Learn

```
13
14 clf = svm.SVC()
15
16 clf.fit(X_train, y_train)
17 confidence = clf.score(X_test, y_test)
18 print(confidence)
19
20 example_measures = np.array([[4,2,1,1,1,2,3,2,1]])
21 example_measures = example_measures.reshape(len(example_measures), -1)
22 prediction = clf.predict(example_measures)
23 print(prediction)
24
```

Growth Hackers

Thank you