

2018  
DAEGU UNIVERSITY  
BIGDATA  
CAMP

BIG DATA  
MACHINE LEARNING  
OPTIMIZATION  
NEURAL NETWORK  
WITH PYTHON

29. JUNE - 4. JULY

|| 브라베룬 조 ||

대구대학교  
수리빅데이터학부 수학전공

최고훈  
김서영  
곽수빈  
이소정  
정지현  
김정엽



대구대학교  
DAEGU UNIVERSITY



# Contents

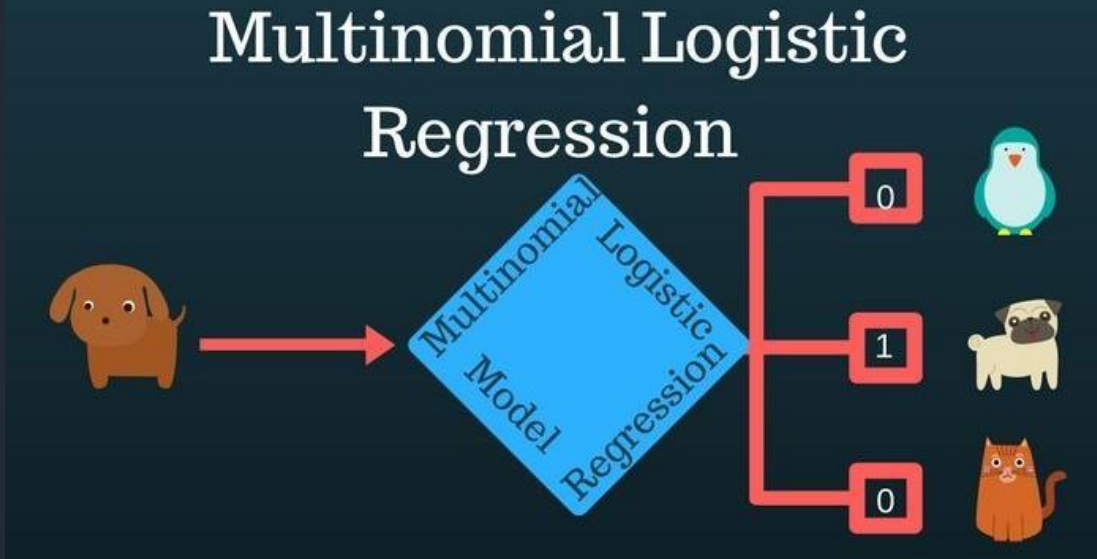
## 01. 모형 소개(Classification)

1. Logistic Regression
2. K-Nearest-Neighbor
3. Random Forest
4. SVM (Support Vector Machine)
5. Gradient boosting model

## 02. Parameters 최적화 방법

1. Grid Search
2. 10-Cross-Validation

# 1. Logistic Regression



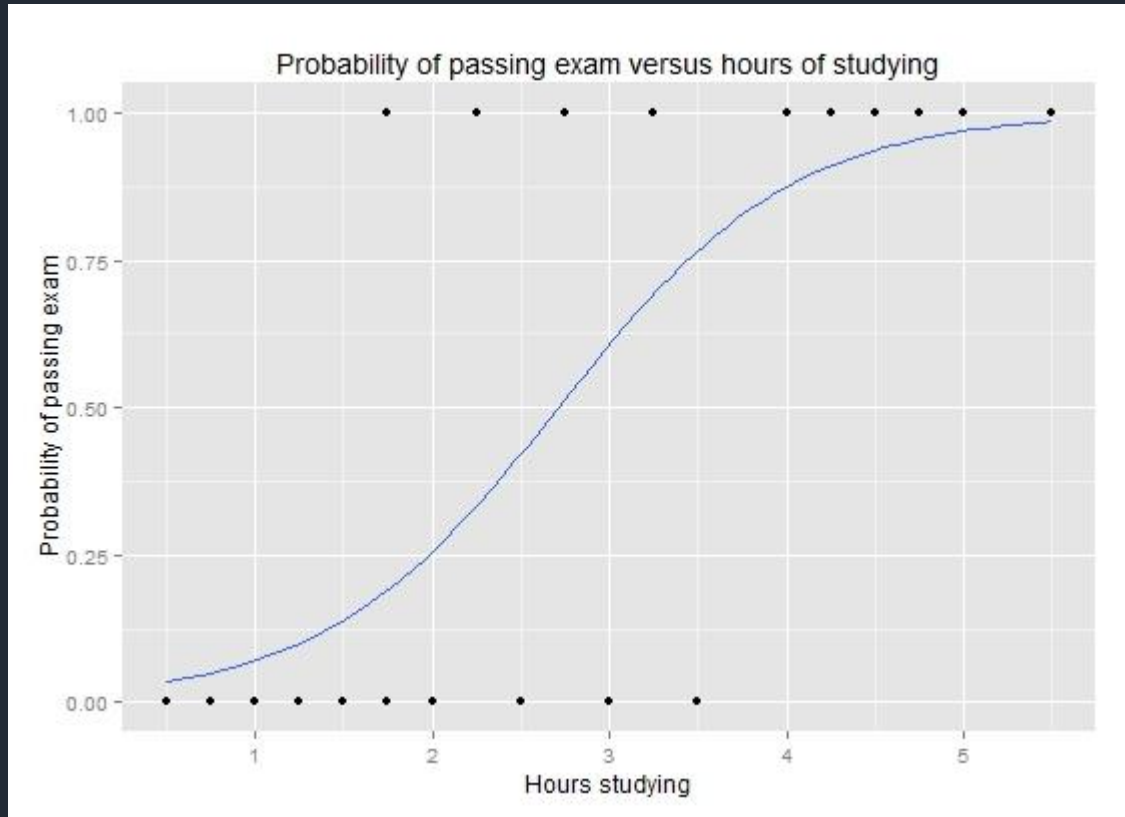
Logistic Regression (로지스틱 회귀)

독립 변수의 선형 결합을 이용하여 사건발생 가능성을 예측하는데 사용되는 통계 기법

목적은 일반적인 회귀 분석의 목표와 동일하게 종속 변수와 독립 변수 간의 관계를 구체적인 함수로 나타내어 향후 예측 모델에 사용하는 것이다.

독립 변수가  $[-\infty, \infty]$ 의 어느 숫자이든 상관없이 종속 변수 또는 결과 값이 항상 범위  $[0, 1]$  사이에 있도록 한다. 이는 오즈비(odds ratio)를 로짓(logit) 변환을 수행함으로써 얻어진다.

# 1. Logistic Regression



## 장점

- 계산 비용이 적고, 구현하기 쉬우며  
결과 해석을 위한 지식 표현이 쉽다.

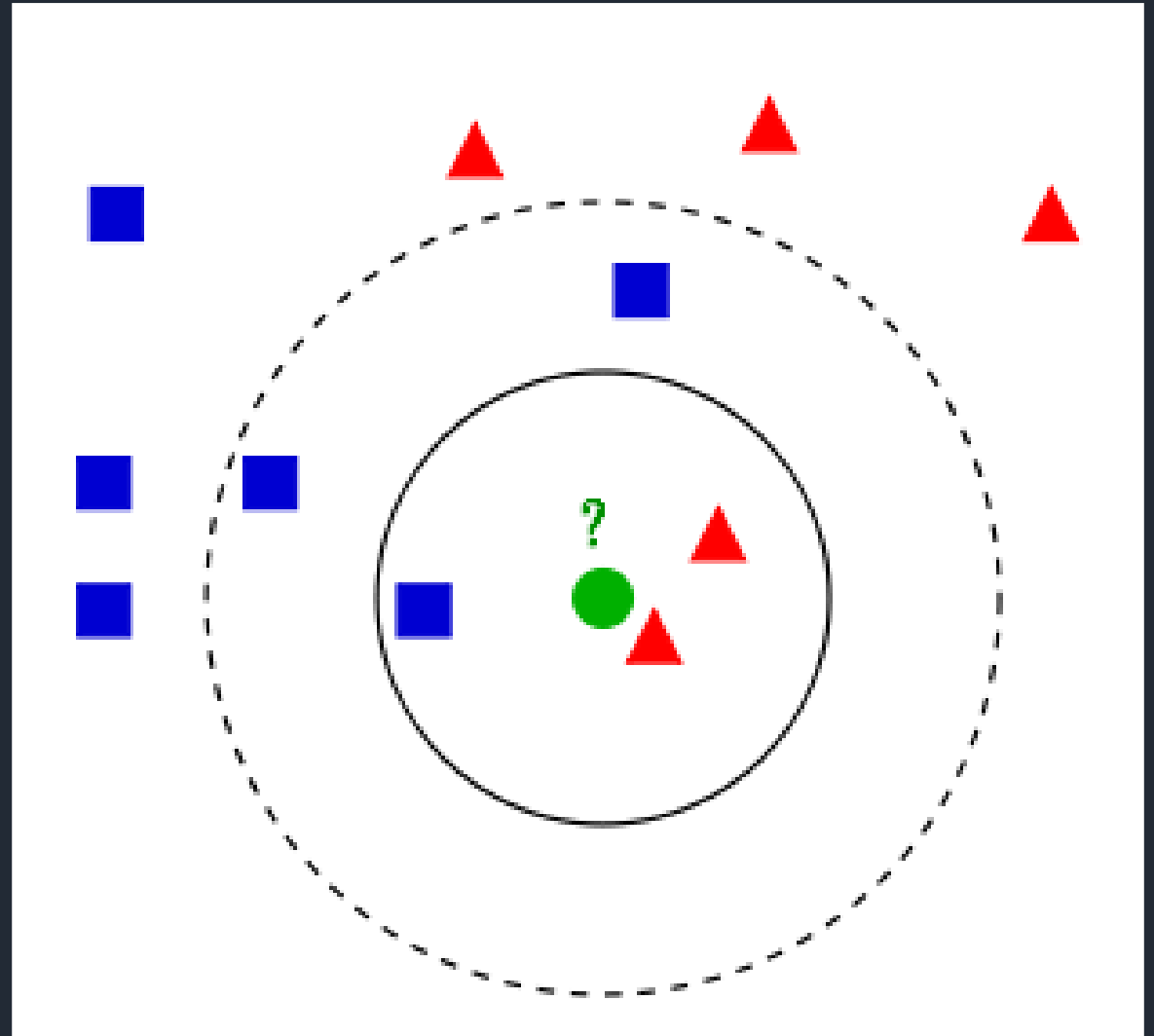
## 단점

- Underfitting(언더피팅) 경향이 있어  
정확도가 낮게 나올 수도 있다.

\* 활용 : 수치형 값, nominal 값

## 2. K-Nearest-Neighbor

테스트 데이터로부터 가장 가까운  $k$ 개의 훈련  
데이터들의 값을 토대로 테스트 데이터의 값을  
예측하는 방법



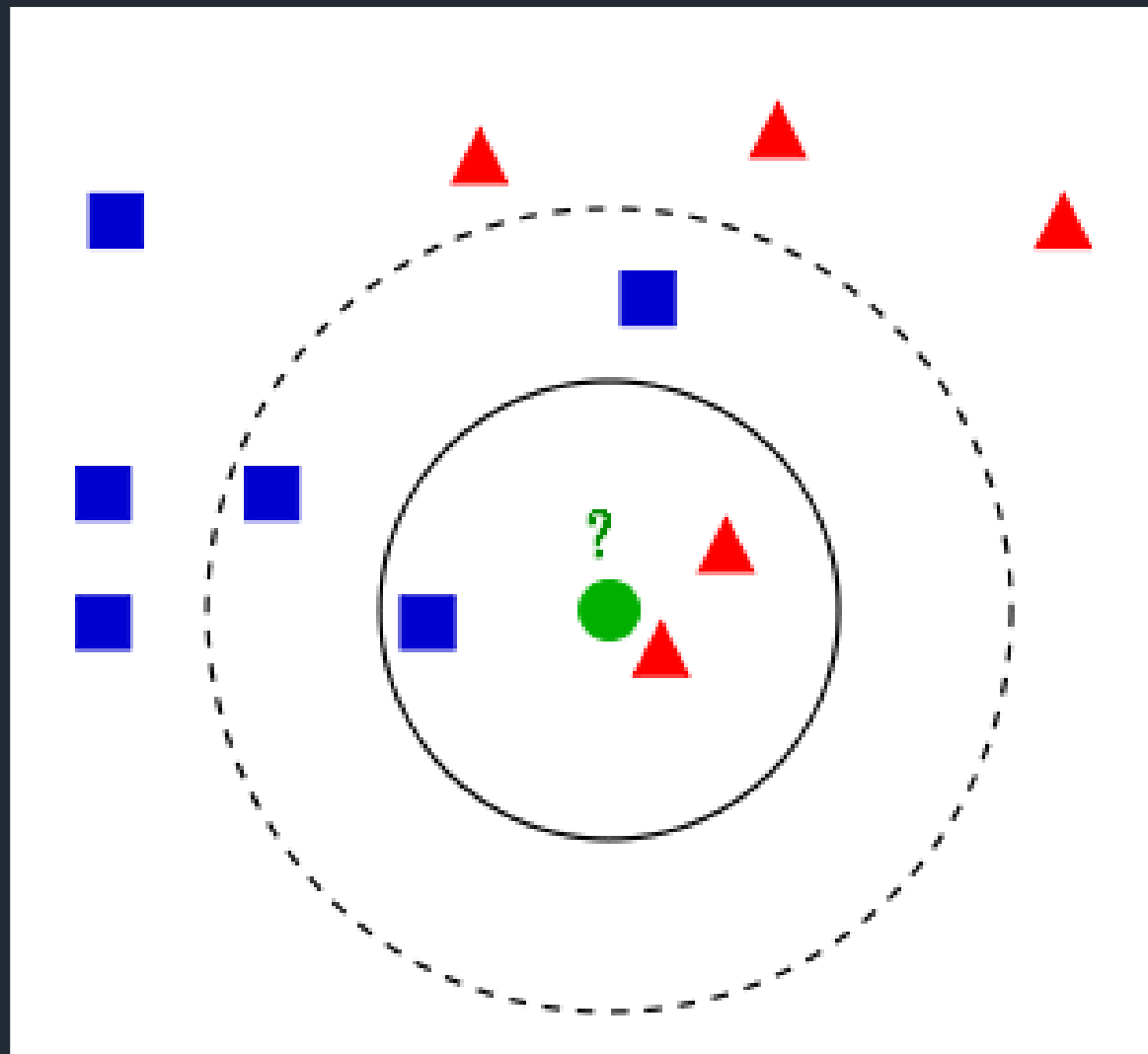
## 2. K-Nearest-Neighbor


### 장점

- 비모수적 방법이기 때문에  
어떤 분포든 상관없음
- 쉽고 이해하기 직관적
- 샘플 수가 많을 때 좋은 분류법이다

### 단점

- 최적의  $k$ 를 선택하기 어렵다
- 데이터가 많을 때 분석속도가 느릴 수 있다
- 특정 분포를 가정하지 않기 때문에  
샘플 수가 많이 있어야 정확도가 좋다





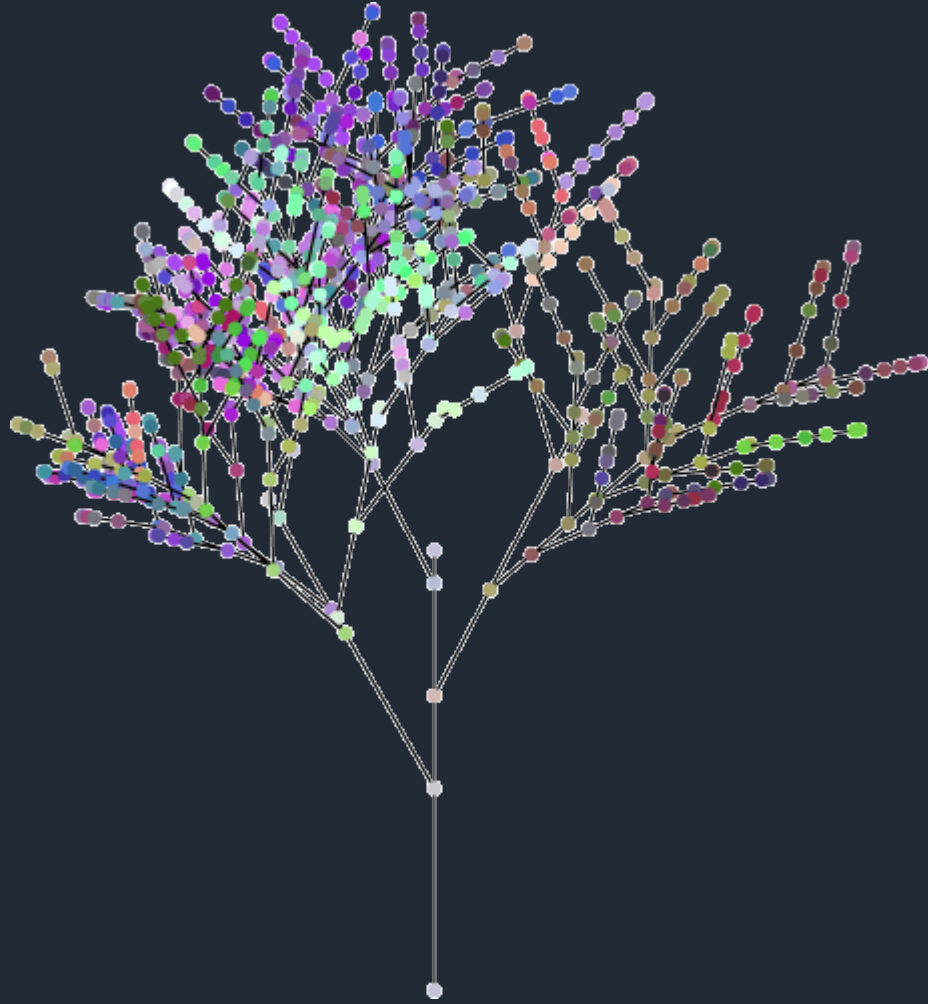
### 03. Random Forest

의사결정 트리의 단점을 개선하기 위한 알고리즘  
( 의사결정 트리는 과적합  
불안정성의 단점이 있다.)

다수의 의사결정 나무를 결합하여  
하나의 모형을 생성하는 방법

관측치, 변수에 임의성 적용하여 다수 트리 작성

### 3. Random Forest



#### 장점

- 예측력 우수 (다양성 극대화)
- 안정성 높음 (다수의 트리의 예측 결과 종합)

#### 단점

- 기존 트리의 장점인 설명력을 잃음  
(다수의 트리를 이용한 의사결정)
- 데이터 set이 큰 거중에서도 변수가 많을 때  
Random Forest는 권장되지 않는다.

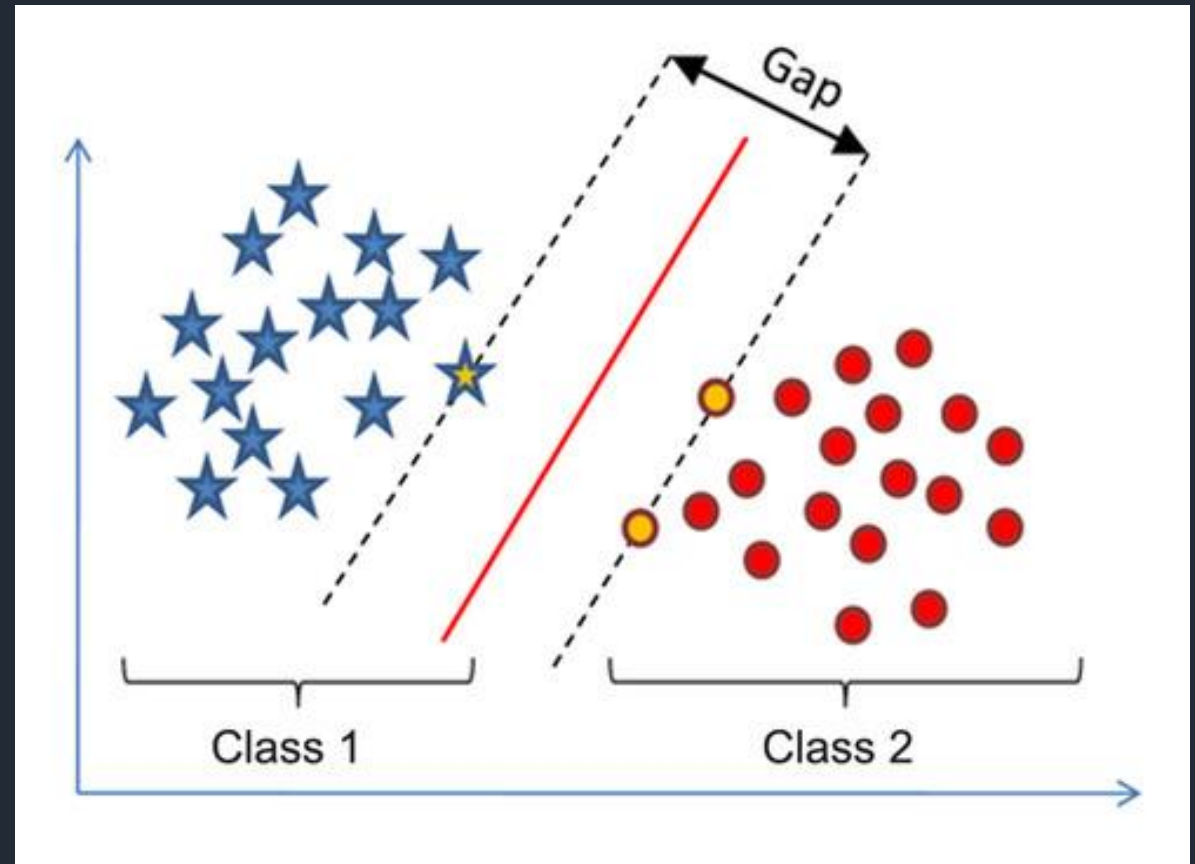


## 4. SVM (Support Vector Machine)

2차원에서 머무르는 것이 아닌, 3차원 등 차원을 높여서 데이터를 나누는 과정을 뜻한다.

고차원에서 데이터를 분류할 수 있는 일종의 칸막이를 만드는 것이다.

\* SVC (Support Vector Classifier) 보조 벡터 분류기



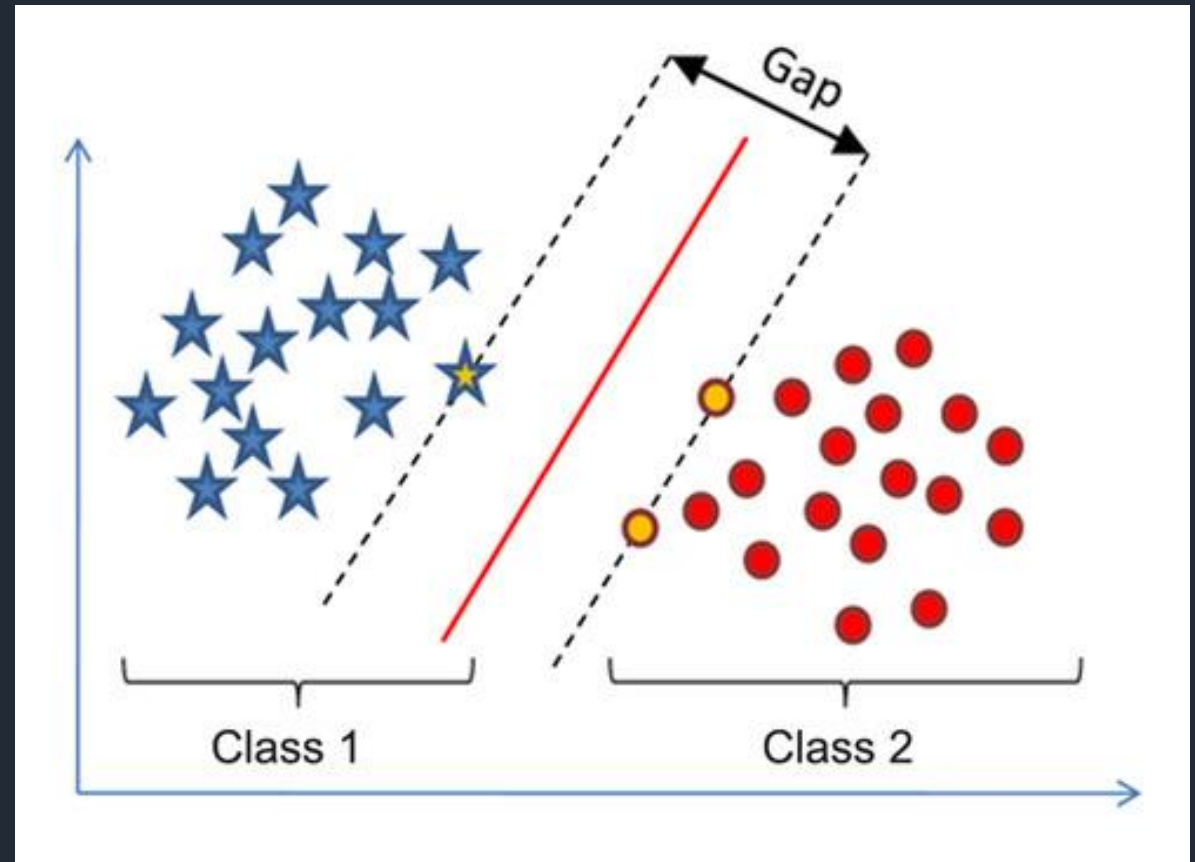
## 4. SVM (Support Vector Machine)

### 장점

- 분류문제나 예측문제 동시에 쓸 수 있다.
- 예측의 정확도가 높다.
- 사용하기 쉽다.

### 단점

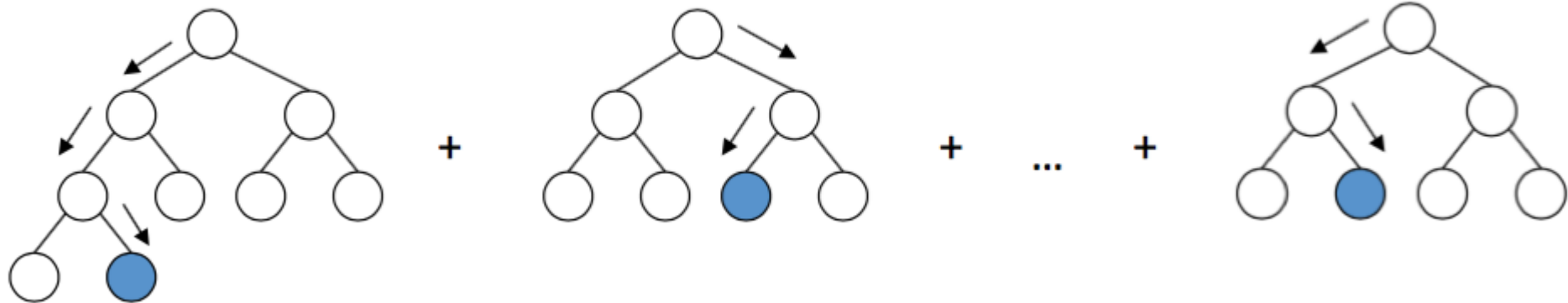
- 모형 구축 시간이 오래 걸린다.
- 결과에 대한 설명력이 떨어진다.



## 5. Gradient boosting model

회귀 및 분류 문제에 대한 기계 학습 기술로 약한 예측 모델,  
일반적으로 의사 결정 tree의 Ensemble 형태로 예측 모델을 생성합니다.

이 모델은 다른 boosting 방법과 같이 단계별로 모델을 만들고 임의로 차등화 할 수 있는 손실 함수를 최적화  
하여 모델을 일반화합니다.



## 5. Gradient boosting model

### 장점

- the decision tree보다 높은 예측력
- 무작위성이 없다  
(대신 강력한 사전 가지치기 사용,  
보통 하나에서 다섯 정도의 깊이 얇은 tree사용)
- 메모리가 작고 빠르다.

### 단점

- 모델의 해석력 떨어진다.
- 파라미터에 따라서 과적합 가능성 높아진다.
- 매개변수를 잘 조정해야 하는 것과 훈련 시간이 길다 .

Parameters 최적화 방법

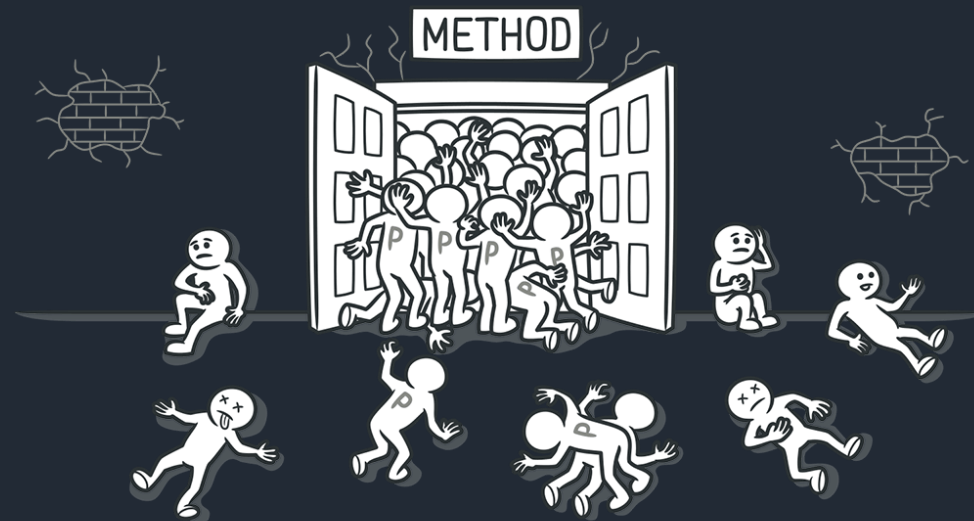
## Manual search

점을 몇가지 골라 찍은 후 결과 값을 찾는 것

## Random search

범위를 정하여 결과 값을 찾는 것

## Bayesian optimization



# 1. Grid Search

큰 틀에서 보면, Manual search와 큰 차이가 없으며, 개념적으로도 비슷하다.  
단, Grid search의 경우는 선험적인 지식을 활용하여 문제를 분석하고,  
hyperparameter의 범위를 정한다.

그리고 그 범위안에서 일정한 간격으로 점을 정하고  
그 점들에 대해 1 개씩 차례대로 실험을 해보면서 최적의 값을 찾는 방법이다.

그렇기 때문에 Grid search는 'Parameter sweep'이라고도 불린다.

Manual search나 Grid search를 할 때는 결과를 판정하기 위한 validation set이 필요하다.

## Grid Search 으로 Parameters 최적화

| N-estimate | Max_depth | Max_leaf_node | Max_feature |
|------------|-----------|---------------|-------------|
| 300        | 50        | 100           | log2        |
| 500        | 100       | 500           |             |
|            | 200       | 1000          |             |
|            | 300       | 2000          |             |
|            |           | 3000          |             |



Random Forest 시각화 하기

\* MAX\_LEAF\_NODES 변화

