

# k-인접 이웃 분석 (k-NN)



## 정의

새로운 데이터가 주어졌을 때 각 개체별 **거리가 가까운 k개 대상의 정보를 활용**하여 **새로운 데이터의 범주를 분류 또는 값을 예측**하는 알고리즘

## 목적

k-인접 이웃 분석에서 적정 k를 어떻게 찾는지 이해하고 k-인접 이웃 분석의 절차 및 장점과 단점을 파악하기 위함

1. k-인접 이웃 분석은 어떻게 **활용**될 수 있는가?

2. K-인접 이웃 분석의 **원리**를 알아보자.

## Goal!

3. **k-인접 이웃 분류**는 어떤 절차에 따라서 수행하는가?

4. 예제 R 코드를 통해 k-인접 이웃 분석을 **실습**해보자.

# Goal 1. k-인접 이웃 분석은 어떻게 활용될 수 있는가?



## 영화 추천 시스템 : 협업 필터링

박스오피스 추천영화 평가누리기


장르 | 국가 | 추천이유

높게 평가한 영화와 비슷한 영화 x


## k-인접 이웃 분석 응용사례




## 추천영화리스트


 WATCHA PLAY 홈 카테고리 평가하기


Q 검색 보고싶어요 김찬경 ▾


그는 당신에게 반하지 않았다 리미트리스 더 셰프 더 스토리: 세상에 숨겨진 사랑 뉴욕 아이 러브 유


재미있게 본 '베를린'과 비슷한 작품


 더 테러 라이브

 감시자들


 의형제


 불한당: 나쁜 놈들의 세상


 리크루트


 무간


재미있게 본 '가디언즈 오브 갤럭시'와 비슷한 작품


 퍼스트 어벤저

 캡틴 아메리카: 윈터 솔져

 빅 히어로

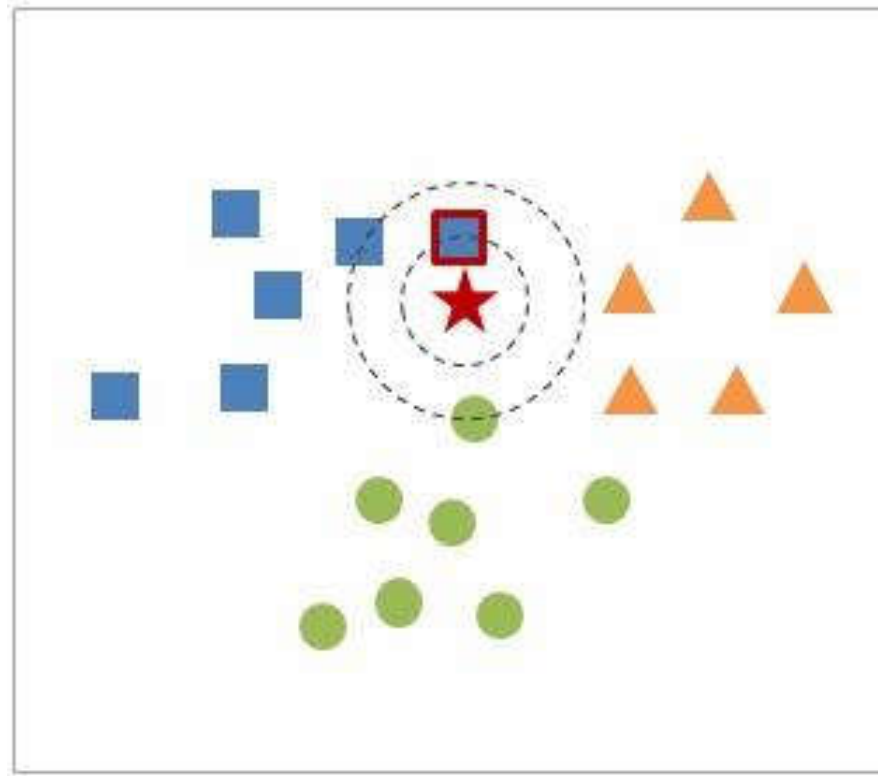
 아이언맨 2

 토르: 천둥의 신

 토르

## k-인접 이웃 분석 응용사례

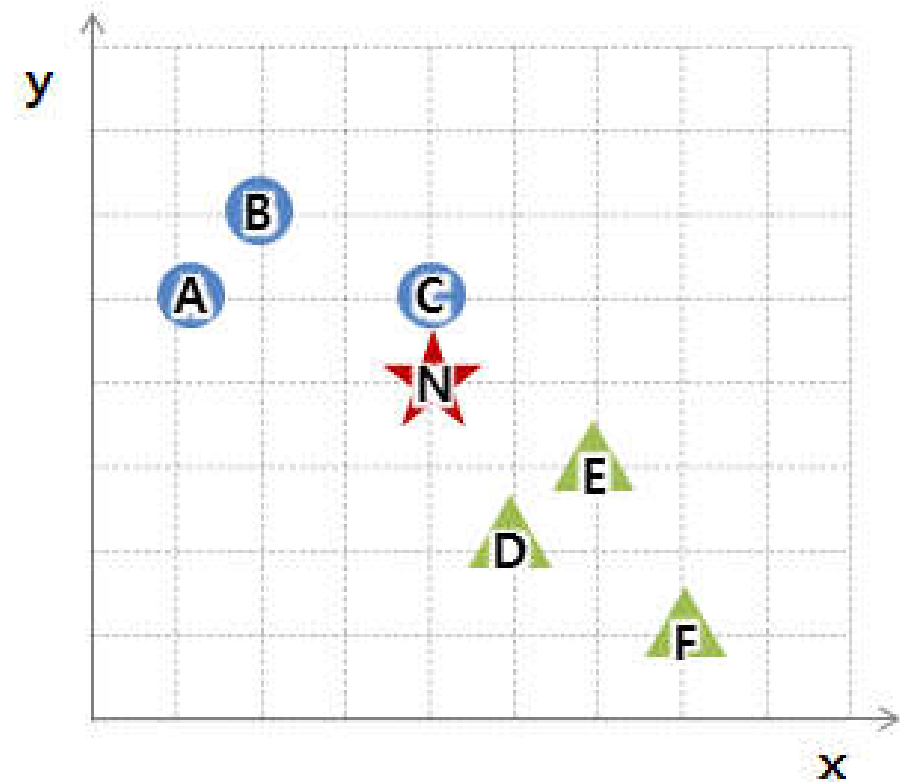
- 📝 k-NN(k-Nearest Neighbor)은 “새로 들어온 ★은 ■ 그룹의 데이터와 가장 가까우니 ★은 ■ 그룹이다.” 라고 분류하는 알고리즘이다.



✅ 예시 출처: <http://kkokkilkon.tistory.com/14>

🔍 6개의 기존 데이터 A-F와 1개의 신규 데이터 N이 있다고 하자.

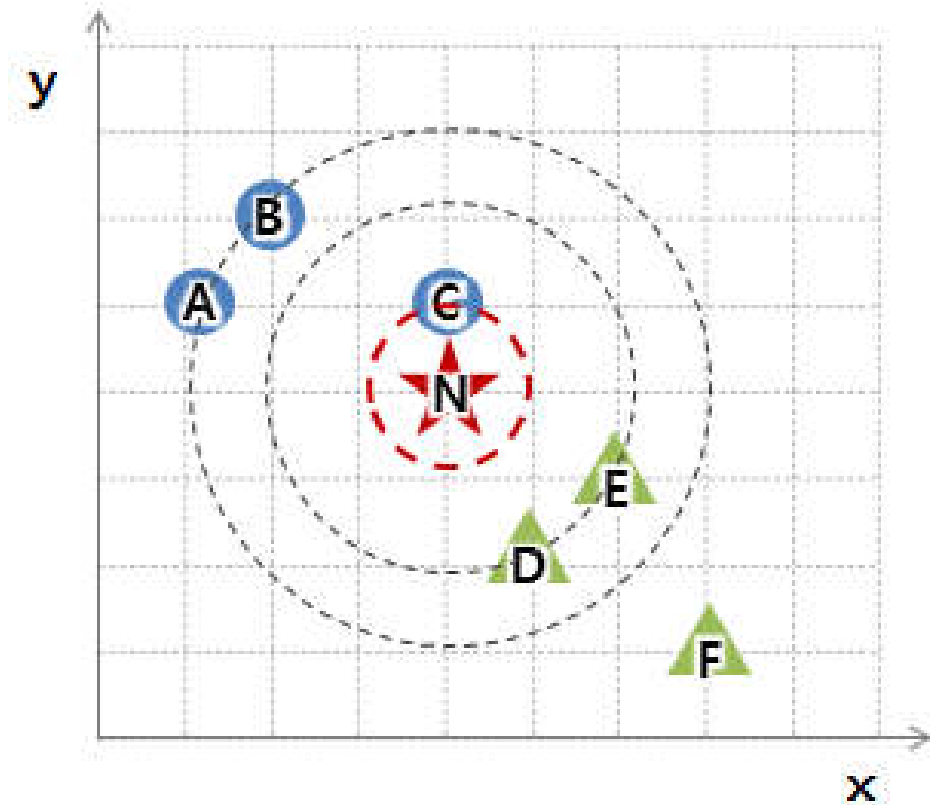
데이터	X좌표	y좌표	그룹
A	1	5	●
B	2	6	●
C	4	5	●
D	5	2	▲
E	6	3	▲
F	7	1	▲
N	4	4	?



✅ k-NN의 k의 역할은 '몇 번째로 가까운 데이터까지 살펴볼 것인가'를 정한 숫자이다.

🔍 6개의 기존 데이터 A-F와 1개의 신규 데이터 N이 있다고 하자.

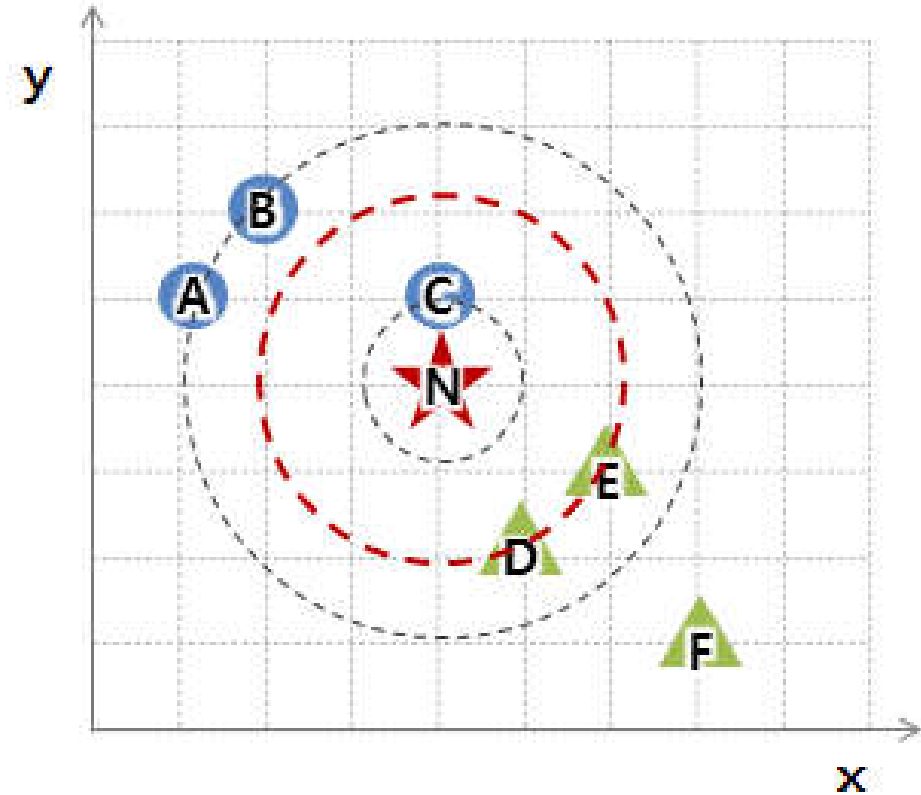
데이터	X좌표	y좌표	그룹
A	1	5	●
B	2	6	●
C	4	5	●
D	5	2	▲
E	6	3	▲
F	7	1	▲
N	4	4	?



- ✓ If  $k=1$ , 거리가 1번째로 가까운 C만을 보고 신규데이터 N을 분류한다. 따라서 N은 C와 같은 그룹인 ●로 분류된다.

🔍 6개의 기존 데이터 A-F와 1개의 신규 데이터 N이 있다고 하자.

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	?

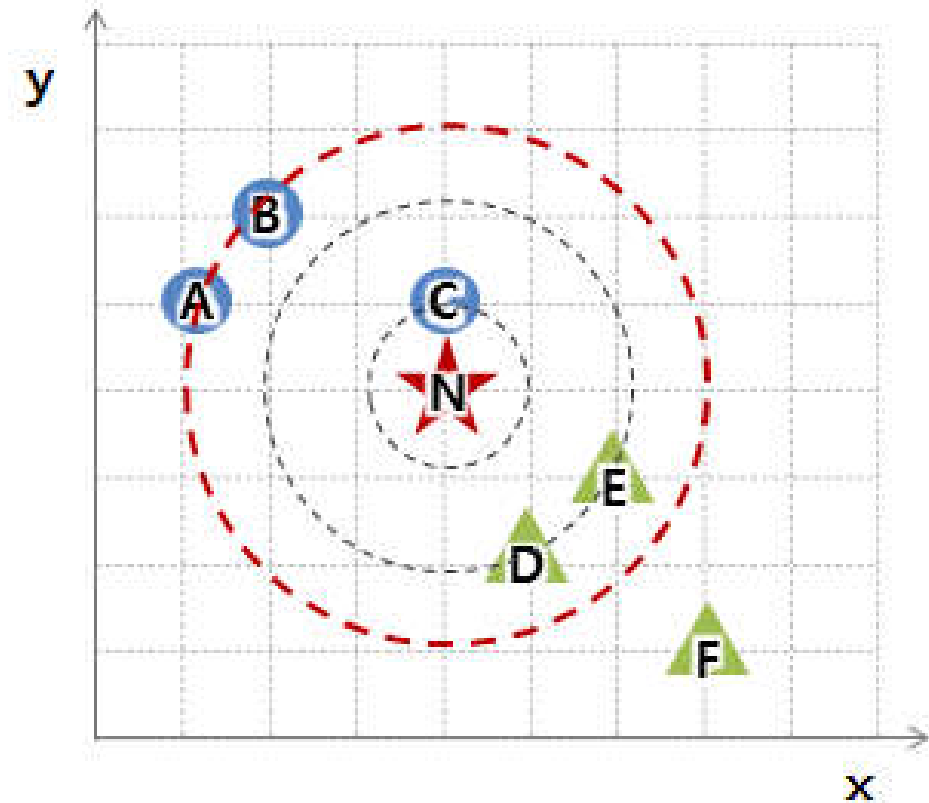


- ✅ If  $k=3$ , 거리가 3번째로 가까운 C, D, E까지 보고 신규데이터 N을 분류한다. 이 때 그룹이 같으면 다수결의 원칙을 따른다. 따라서 N은 ▲로 분류한다.



🔍 6개의 기존 데이터 A-F와 1개의 신규 데이터 N이 있다고하자.

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	?



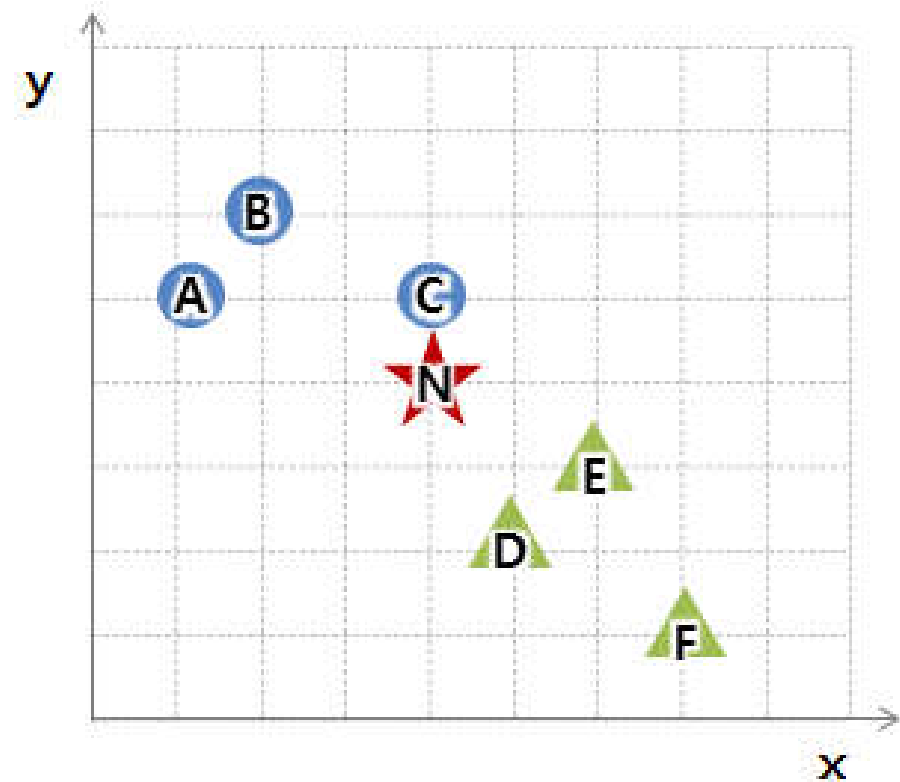
✅ If  $k=5$ , 거리가 5번째로 가까운 C, D, E, B, A까지 보고 신규데이터 N을 분류한다. 여기서는 3:2가 되어 N은 ●로 분류한다.

## Goal 2. k-인접 이웃 분석의 원리를 알아보자.



🔍 6개의 기존 데이터 A-F와 1개의 신규 데이터 N이 있다고 하자.

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	?



- ☑ 이처럼 같은 데이터임에도 k가 얼마냐에 따라 N이 ●로 분류되기도 하고 ▲로 분류되기도 한다. 그만큼 적절한 k를 정해주는 것이 중요하다.

사실 여기까지가 k-NN의 기본적인 아이디어의 전부이다.

분류 문제만 설명했지만, 회귀 문제 역시 비슷하게 풀 수 있다.  
인접한 k개의 평균으로  $y$ 의 값을 예측하면 된다.