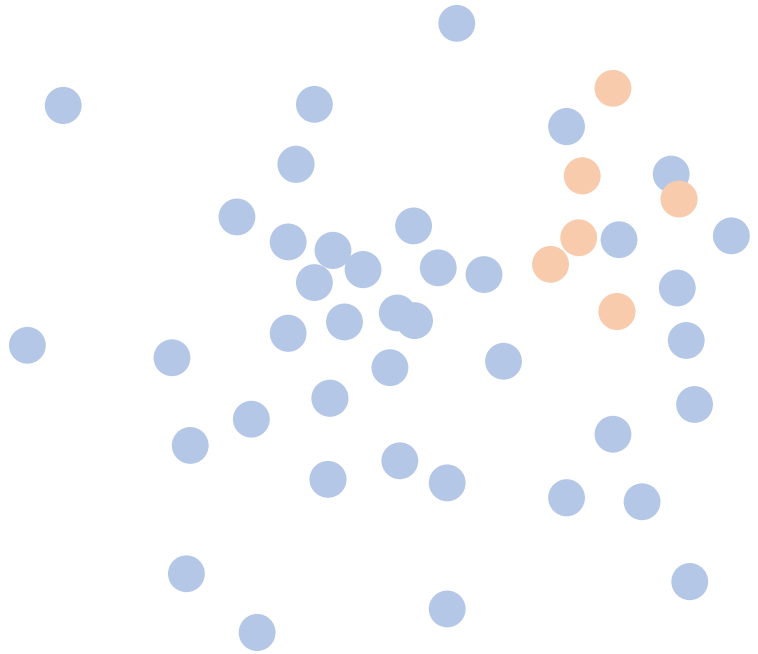
Chapter 08. 효과적이면서도 쉽게 쓸 수 있는 기법들

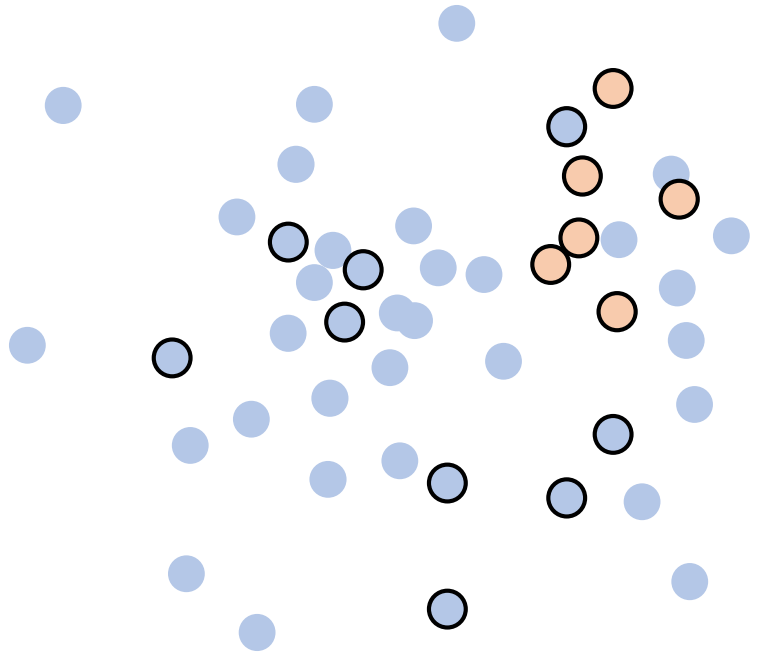
STEP2. SMOTE 알고리즘

불균형 데이터 Imbalanced Data



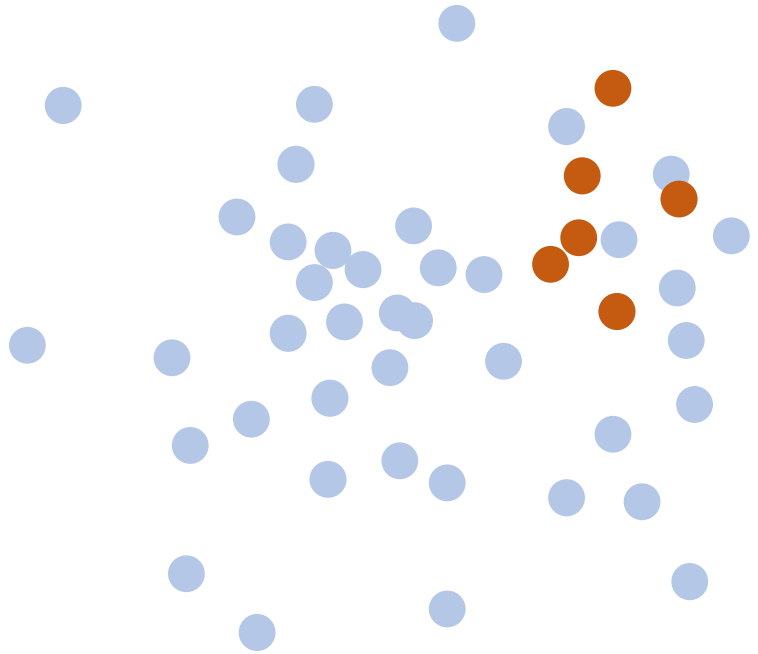
클래스별로 학습 데이터셋의 크기가 급격히 차이가 나는 데이터를 Imbalanced Data라고 부른다.

임의 언더 샘플링 Random Under Sampling



다수 클래스(Majority Class)에서 임의로 샘플링하여 크기를 맞추는 방법을 Random Under Sampling이라 한다.
이 경우, 임의로 선택된 샘플이 대표성이 떨어질 경우 학습이 잘못된 방향으로 될 수 있다.

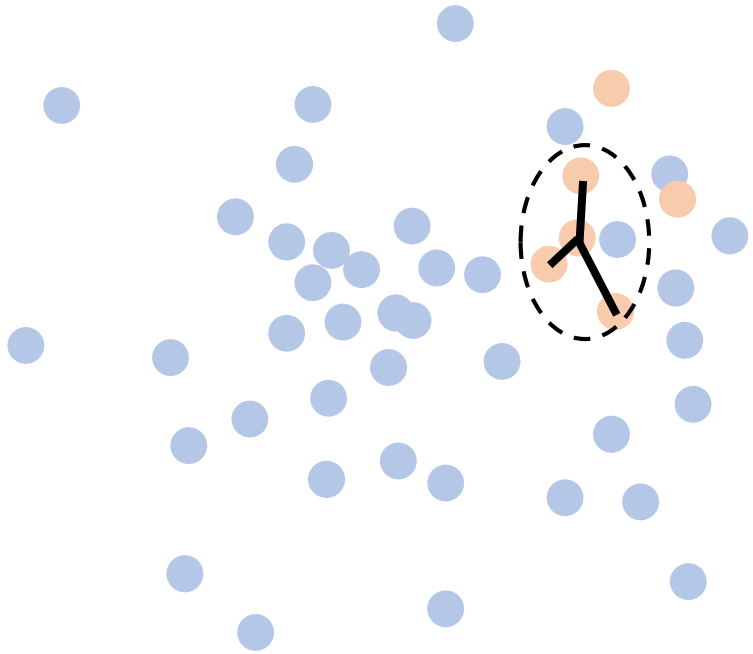
임의 오버 샘플링 Random Over Sampling

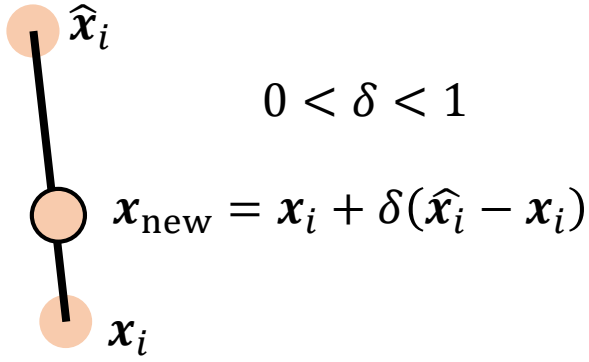


소수 클래스(Minority Class)의 데이터를 반복하여 양을 학습 데이터의 양을 맞추는 방법.

이는 학습 시 소수 클래스의 가중치를 증가시키는 것과 유사하다.

SMOTE Synthetic Minority Oversampling Technique

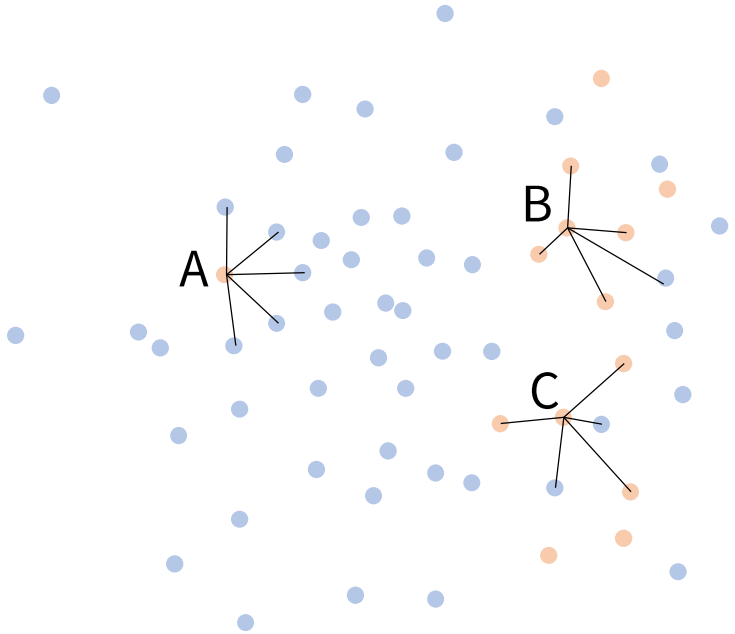




k-Nearest Neighbor 중, 랜덤으로 하나의 샘플을 선택하여 Linear Combination을 추가한다.

임의 오버 샘플링에 비해 다양한 데이터를 추가할 수 있는 장점이 있다.

Borderline-SMOTE



A : 잡음 (Noise) 샘플. $N_M = k$

B : 안전 (Safe) 샘플. $N_M < \frac{k}{2}$

C : 위험 (Danger) 샘플. $\frac{k}{2} \leq N_M < k$

안전한 지역에 있거나, 잡음으로 간주되는 샘플은 오버 샘플링 하지 않고,

위험 지역인 경계(Borderline)에 있는 샘플만 오버 샘플링하여 SMOTE를 효과적으로 개선했다.