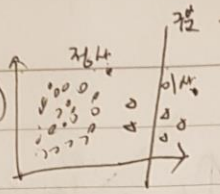
불균형 data 처리.

No. 20220307 고종규

불균형 데이터 ex)



경고 → 문제점.

→ 샘플링으로 해결

↙ 해결법

샘플링 기법 — 언더 샘플링 : 샘플개수를 축소.
- ! 계산시간↓ 클래스오버랩↓
- ! 정보손실↑

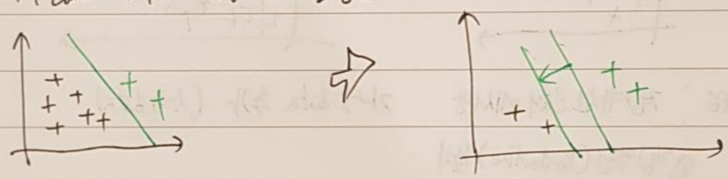
Random undersampling / Tomek link / condensed Nearest Neighbor Rule / One-sided selection

— 오버 샘플링 : 샘플개수 증가
- ! 정보손실↓ 높은분류정확도.
- ! 과적합가능성↑
- ! 계산 시간↑
- ! 노이즈 민감

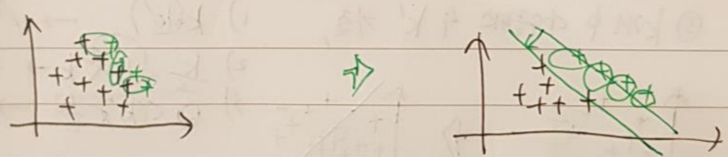
Resampling / SMOTE / Borderline — SMOTE / ADASYN

모델링 기법

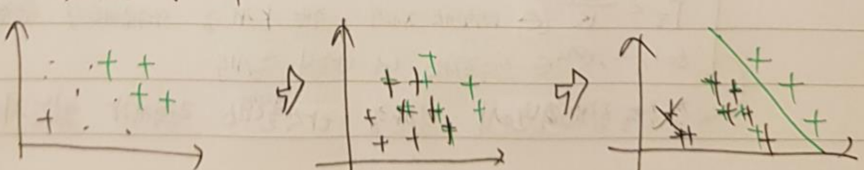
Random Sampling : 다수 data 에서 랜덤으로 샘플링.



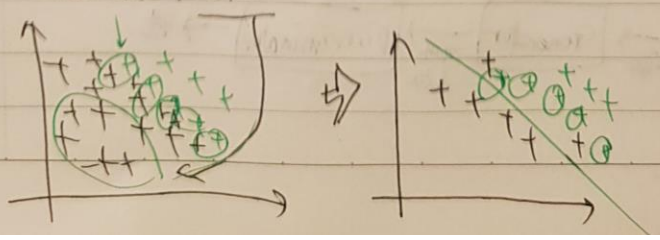
Tomek link : 다수 data 에서 Tomek link 형성된것 (경계선에 있는 data) 제거.



Condensed nearest neighbor ≒ CNN : ① 다수data 에서 1개만 남김 ② 나머지 다수 data 각각의 위치에서 1NN
③ 1NN 결과가 원래 다수data로 직원것들 다 제거



One Side Selection : Tomek link + CNN

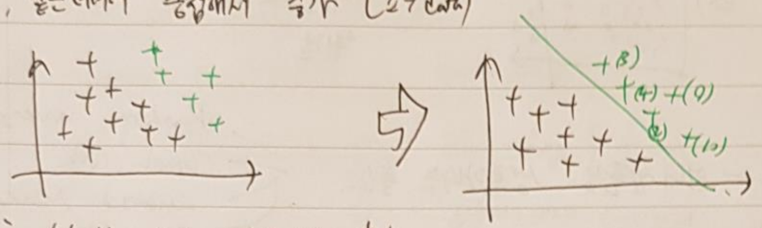


④) 모델링 기법으로 불균형 문제 해결

- 비용기반학습 : 소수 data에 가중치 부여.
- 단일 클래스 분류기법 : 보통 다수 data 기준, 그 거리 분류경계선 (ex 원 생성

Resampling

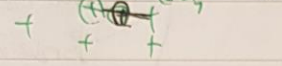
; 같은 데이터 중첩해서 증가 (소수 data)



SMOTE

; 소수 data 에서 가상 data 추가

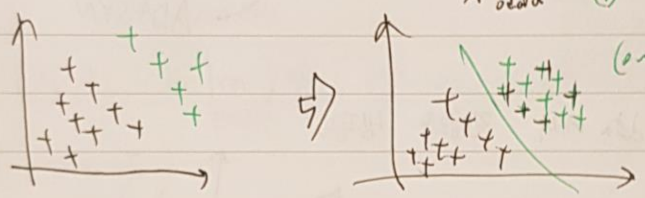
ex k=5,



→ ①② 방향으로 가상data ③ 추가.

공식) 생성된 가상data $= X + u \cdot (X(nn) - X)$

- X : ①
- u : (0~1) 값..
- $X(nn)$: ②, X : ①. → ①②



Borderline-SMOTE

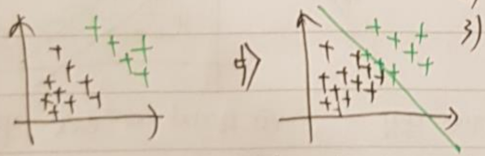
; 경계선 근처 에서만 가상 data 추가 (소수 data)

① 경계선 (Borderline) 찾기

② 소수클래스 中 x_i 지정. 그중 k개 주변탐색

③ k개 中 다수클래스 수 k' 기준,

1) $k = k'$ → Noise. 탈락.
2) $\frac{k}{2} < k' < k$ → Danger OK.
3) $0 < k' < \frac{k}{2}$ → Safe 탈락.



ADASYN

; 위치에 따라 적응적 가상data 추가 (소수data)

$r_i = \frac{\Delta_i}{K}$ ← 소수클래스 x_i의 주변 K개 중 다수클래스의 관측치 개수

$i = 1, \dots, m_s$ ← 소수클래스 내 관측치 총개수

→ 각 소수 클래스 주변에 얼마나 많은 다수클래스 관측치가 있는지에 대한 지표.

④) GAN 으로 오버샘플링.