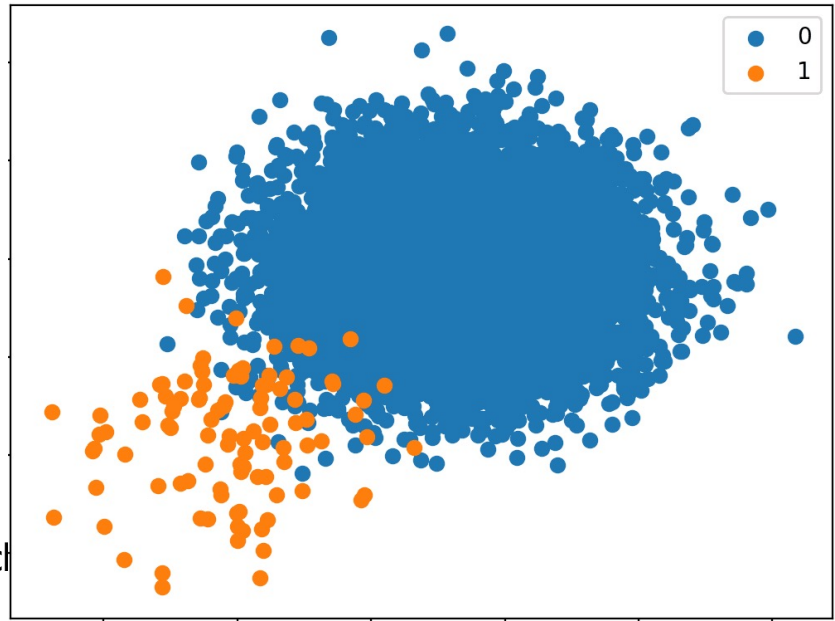
Imbalanced classification

Sang Yup Lee

Imbalanced classification

- What is it?
  - 특정 클래스에 대한 관측치의 수가 상대적으로 더 많은 (적은) 경우 (특히 학습 데이터에 대해서)
    - 일반적으로 학습데이터가 불균형이면 평가데이터나 실제 데이터도 불균형일 확률이 높다 (or vice versa)
  - 예) Iris_data_imbalanced.csv
  - 용어
    - Majority class
    - Minority class



Imbalanced classification

- 주요 문제
  - 예측력이 떨어진다. 특히 minority class 에 대해서
    - 즉, minority class 에 대한 precision과 recall 값이 좋지 못하다.
    - 보통 일반적으로 minority class에 더 많은 관심, 예) 질병 여부
    - Minority class의 특성을 잘 파악하지 못함 (minority class 를 noise 로 간주) 즉, minority class 를 majority class 로부터 구분을 잘 못함.
- Examples
  - 사기 예측, 질병 예측, 신용불량자 예측, 기계 고장 예측, 스팸 이메일 예측 등
- When classes are imbalanced, accuracy is not a good metric for model evaluation.
  - Precision, recall should be reported as well.
  - AUC is also preferred.
- See "LR_iris_example_imbalanced.ipynb"

Imbalanced classification

- 주요 원인 2가지
  - Biased sampling
    - Not a representative sample
    - 예) 모집단의 경우 0 과 1 (예, 물건 구매 고객)의 비중이 유사한데, sampling을 0이 상대적으로 많게 하는 경우
  - Properties of the domain
    - 특정 질병 여부
    - 네이버 영화 댓글 긍/부정

Imbalanced classification

- How to solve?
  - If possible, collect more data (especially for minority class)
  - Resampling the dataset
    - **Over-sampling (for minority class)**
      - Copy the existing data points (하지만, 새로운 정보가 추가되지는 않는다)
      - 데이터가 별로 없는 경우 사용
    - Under-sampling (for majority class)
      - Delete the existing data points (정보 손실 발생)
      - 데이터가 많은 경우 사용
  - Generate synthetic data points
    - SMOTE (Synthetic Minority Over-sampling Technique)*
  - Try alternative algorithms
  - Try penalized models (Cost-Sensitive Training)
    - Minority class 를 틀리는 경우 추가 cost 부여 (가중치)

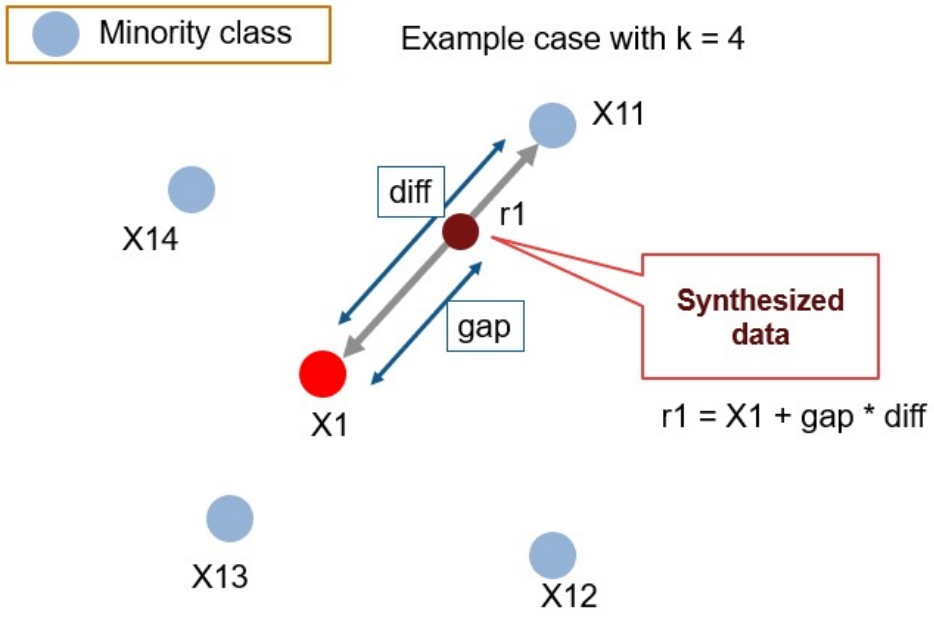
Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

SMOTE

- SMOTE: Synthetic Minority Oversampling Technique
  - An oversampling technique
    - Synthetic data points are generated for the minority class.
    - 데이터셋에 있는 기존 관측치들의 정보를 활용
  - Procedure
    - 아래 과정 반복 (다음 슬라이드 참고)
      - Randomly choose a data point of minority class
      - 해당 관측치와 유사도가 가장 큰 관측치 K 개 선택 (of minority class)
      - 관측치들 간 difference (거리) 계산
      - 0 ~ 1 사이의 값 (randomly chosen) 을 곱함
      - 이를 이용하여, 새로운 data point 생성

SMOTE

- Procedure (cont'd)



Source: https://github.com/minoue-xx/Oversampling-Imbalanced-Data

SMOTE

- Python code
  - See "Oversampling_methods_examples.ipynb"
  - You need to install "imbalanced-learn"
    - pip install imbalanced-learn
- 문제점
  - 한 (두개의 점을 연결하는) line 에서 여러 개의 data points ⇒ 추출 feature space를 잘 반영하지 못한다.
  - SMOTE tends to create a large number of noisy data points in feature space.
    - If there are observations in the minority class which are outlying and appears in the majority class, it causes a problem for SMOTE, by creating a line bridge with the majority class.

Border-line SMOTE

- Border-line SMOTE*
  - 주요 특징: 모든 minority data points를 사용한 것이 아니라 border line에 있는 즉, 상대적으로 분류하기 힘든 minority data points를 사용
  - 과정
    - Minority class에 속한 모든 관측치에 대하여 class 구분 없이 nearest neighbor 추출
    - 뽑아낸 nearest neighbor 중 절반 이상이 majority class인 minority 관측치를 DANGER라고 하는데, 이는 곧 borderline에 있는, 분류기가 어려워하는 example의 set을 의미
    - DANGER set 존재하는 minority data points에 대하여 nearest neighbor들을 다시 뽑는다.
    - 아래 식을 이용하여 synthetic points 생성

$$synthetic_j = p'_i + r_j \times dif_j, \qquad j = 1,2,\ldots,s$$

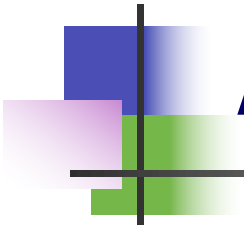
Han, H., Wang, W. Y., & Mao, B. H. (2005, August). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing* (pp. 878-887). Springer, Berlin, Heidelberg.

ADASYN

- ADASYN (Adaptive Synthetic Sampling)*
  - Minority class 관측치에 대해서 주변에 majority class 관측치가 많을수록 더 많이 oversampling 하는 방법
    - Minority class 관측치에 대해서 주변에 majority class 관측치가 많을수록 분류가 더 어려운 관측치라고 간주

ADASYN is based on the idea of adaptively generating minority data samples according to their distributions: more synthetic data is generated for minority class samples that are harder to learn compared to those minority samples that are easier to learn.

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008, June). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* (pp. 1322-1328). IEEE.

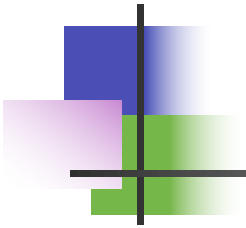


ADASYN

- Python code
  - See "Oversampling_methods_examples.ipynb"

Penalized models (Cost-Sensitive Training)

- Cost-Sensitive Logistic Regression for Imbalanced Classification
  - 종속변수값 (즉, 클래스)에 따라서 cost 에 다른 가중치를 주는 방법
  - 원래 비용함수 형태 (하나의 관측치에 대해서)
    - $-\{y_i \ln p(y_i = 1) + (1 - y_i) \ln p(y_i = 0)\}$
  - 가중치를 준 형태
    - $-\{w_1 y_i \ln p(y_i = 1) + w_0 (1 - y_i) \ln p(y_i = 0)\}$
  - Refer to "Cost_sensitive_Logistic.ipynb"



Q & A