Open DMQA Seminar

Handling imbalanced datasets

2021.11.19

Data Mining & Quality Analytics Lab.
발표자 : 황하은

발표자 소개

- **황하은 (Haeun Hwang)**
  - 고려대학교 산업경영공학부 재학 중
  - Data Mining & Quality Analytics Lab (김성범 교수님)
  - 석사과정 (2020.03 ~)
- 관심 연구 분야
  - Deep learning for tabular data
  - Design optimization in manufacturing
- E-mail: julyh777@korea.ac.kr

INDEX

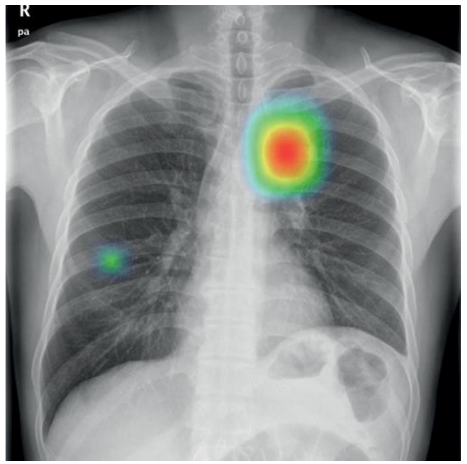
불균형 데이터란?

❖ 데이터의 비율이 균일하지 않고 한쪽으로 치우친 데이터

금융 사기 데이터

제조 불량 데이터

의료 진단 데이터



불균형 데이터란?

❖ **데이터의 비율이 균일하지 않고 한쪽으로 치우친 데이터**

❖ **다수** 클래스: 데이터 내에서 상대적으로 다수를 차지하는 class

❖ **소수** 클래스: 데이터 내에서 상대적으로 소수를 차지하는 class

범주형 데이터

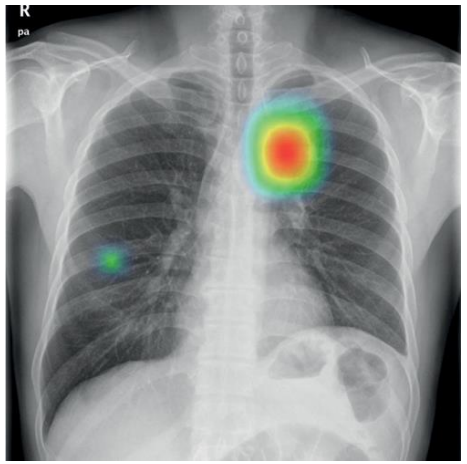
금융 사기 데이터

정상 사기

제조 불량 데이터

정상 불량

의료 진단 데이터



음성 양성

불균형 데이터란?

❖ 데이터 내 각각의 class들이 차지하는 데이터의 비율이 균일하지 않고 한쪽으로 치우친 데이터

❖ Left/ Right Skewed distribution

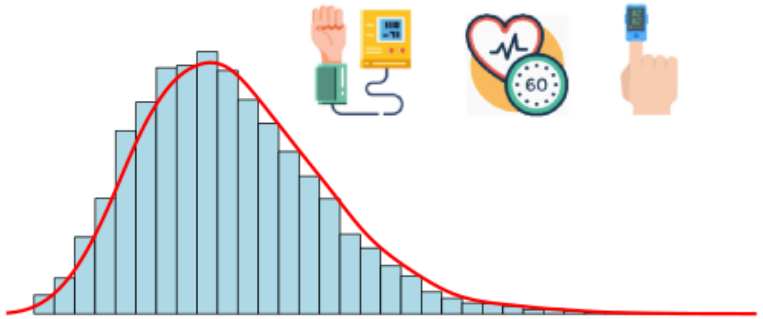
❖ Long tail distribution

연속형 데이터

연령 분포 데이터

1세 ~ 100세

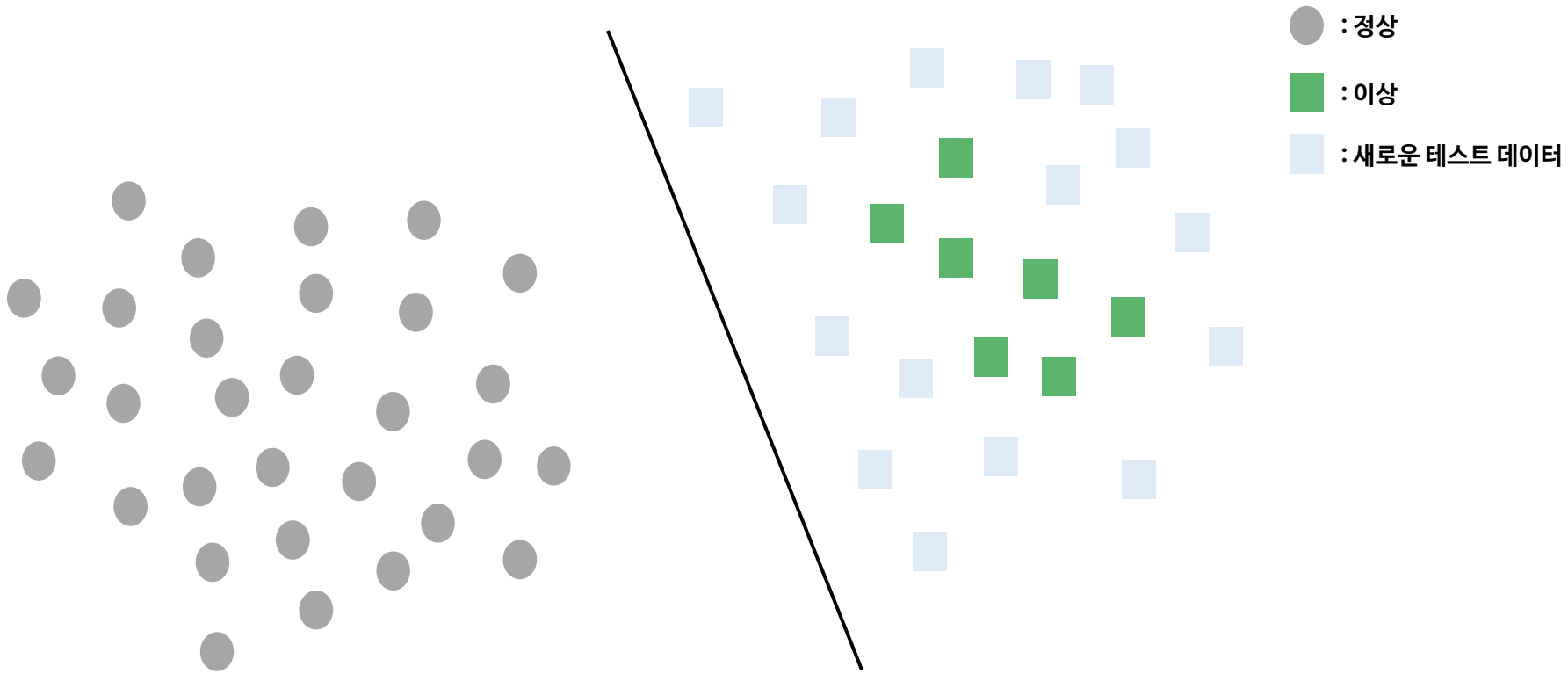
환자 활력 징후 데이터



혈압
맥박수
산소포화도

불균형 데이터의 문제점

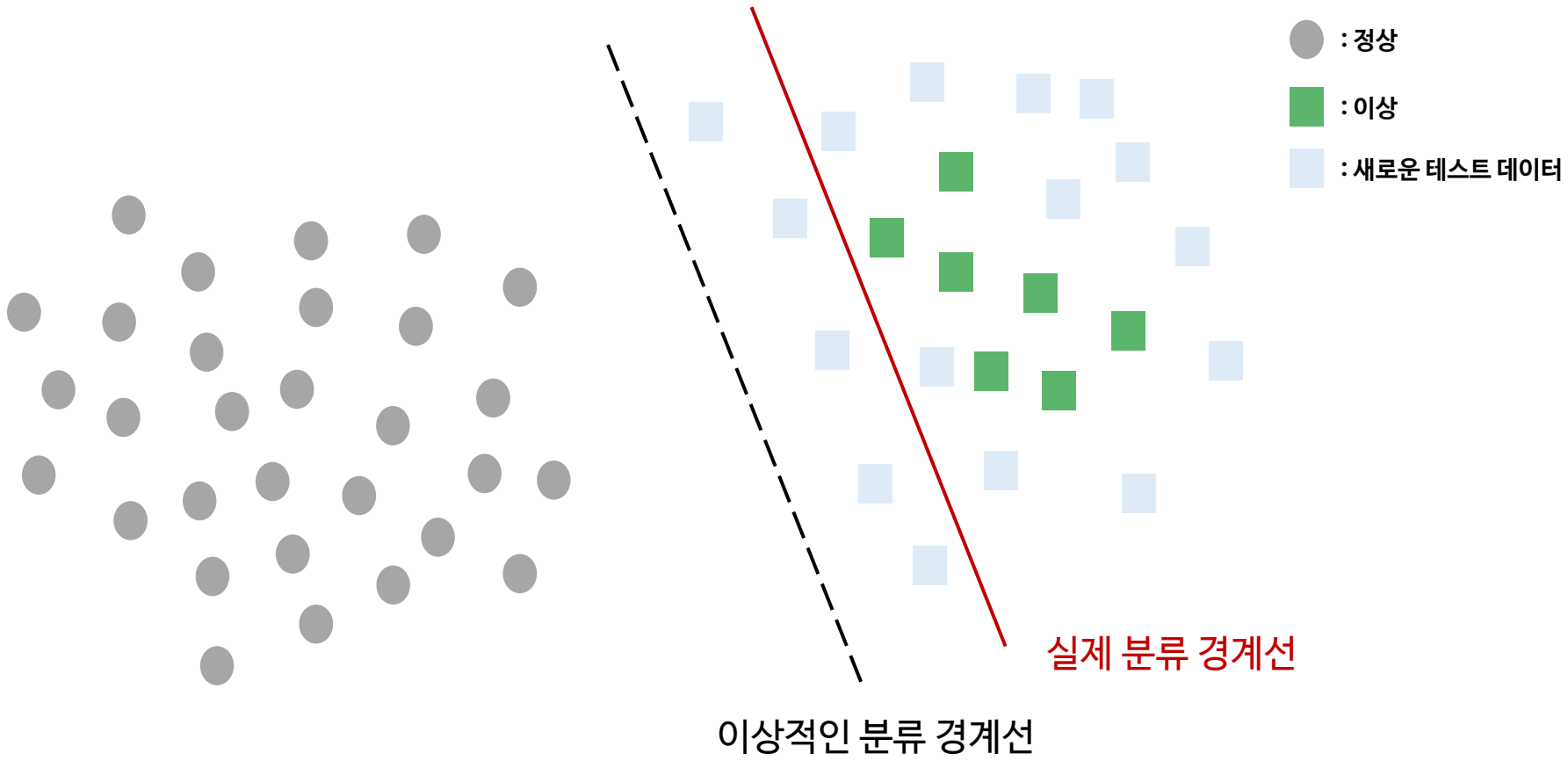
❖ 이상치에 편향된 분류 경계선이 학습됨



이상적인 분류 경계선 → 좋은 일반화 성능

불균형 데이터의 문제점

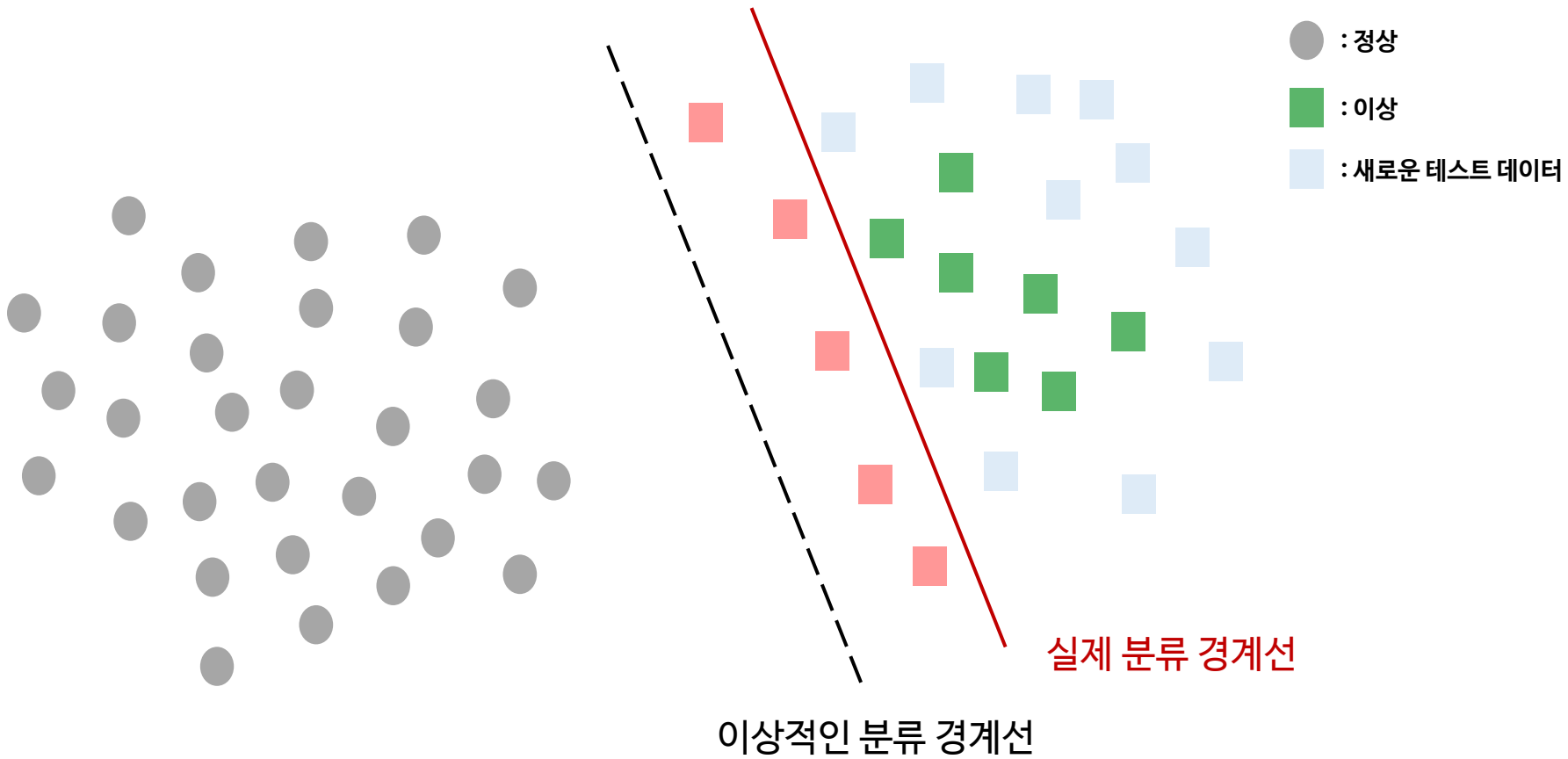
❖ 이상치에 편향된 분류 경계선이 학습됨



불균형 데이터의 문제점

❖ 이상치에 편향된 분류 경계선이 학습됨

- 테스트 단계에서의 오분류율이 높음



불균형 데이터의 문제점

- ❖ **이상치에 편향된 분류 경계선이 학습됨**
- ❖ **모델 성능에 대한 왜곡**
  - 높은 정확도를 보이지만, 이상 클래스(소수 클래스)에 대해서는 잘 분류하지 못하는 모델이 학습됨
  - 부족한 클래스 데이터를 예측하는 데 실패하더라도 그 수가 작아서 결국 전체 loss는 낮게 나오기 때문임

Confusion matrix

		실제 정답	
		정상	이상
분류 결과	정상	80	10
	이상	0	10

0.5

$$\text{Accuracy} = \frac{80+10}{80+10+0+10} = 0.9$$

분류 태스크 연구

분류 테스크 연구동향

❖ Data level 과 Algorithm level로 구분할 수 있음

Data - level 방법

- **Random resampling**
  - Random Over-Sampling (ROS)
  - Random Under-Sampling (RUS)
- **Synthetic sample**
  - SMOTE
  - GAN 활용

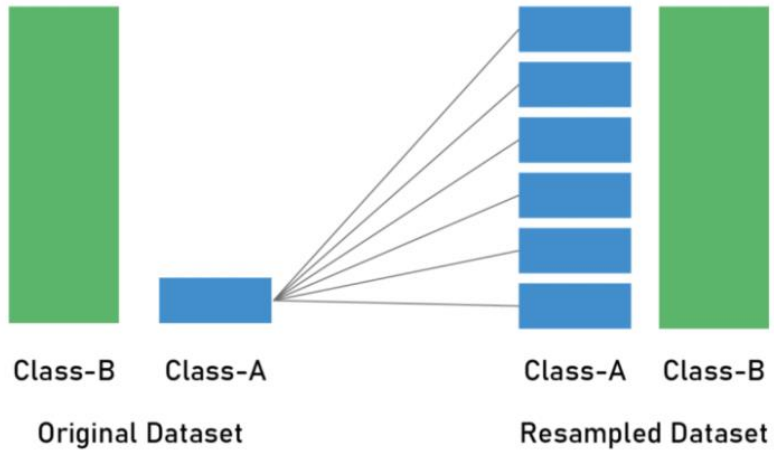
Algorithm - level 방법

- **Cost-sensitive learning**
  - Inverse frequency weight
  - Square root weight
  - Focal loss
- **Two Stage Training**

Data-level 방법: Random resampling

❖ Random resampling

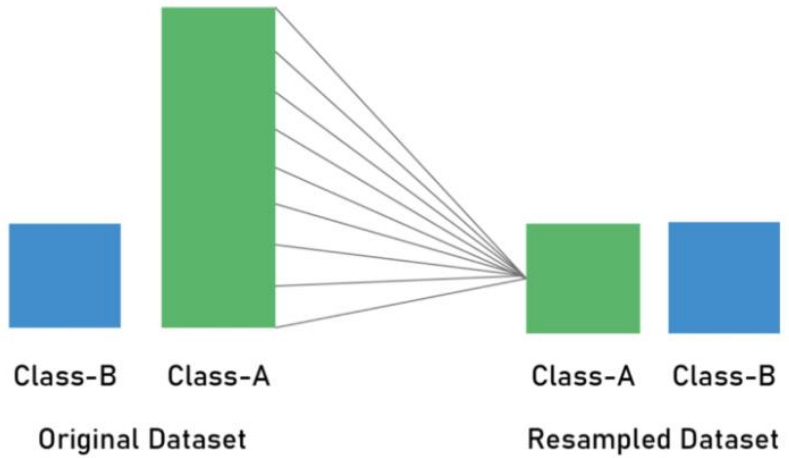
Random Over Sampling (ROS)



소수 클래스의 데이터를 임의로 복제하는 방법

소수 클래스 분류시 Overfitting이 발생할 가능성이 있음

Random Under Sampling (RUS)



다수 클래스의 데이터를 임의로 샘플링하는 방법

학습데이터의 유실이 발생
임의로 뽑은 샘플로 인해 편향된 분류결과가 나올 수 있음

Data-level 방법: Synthetic sample 생성

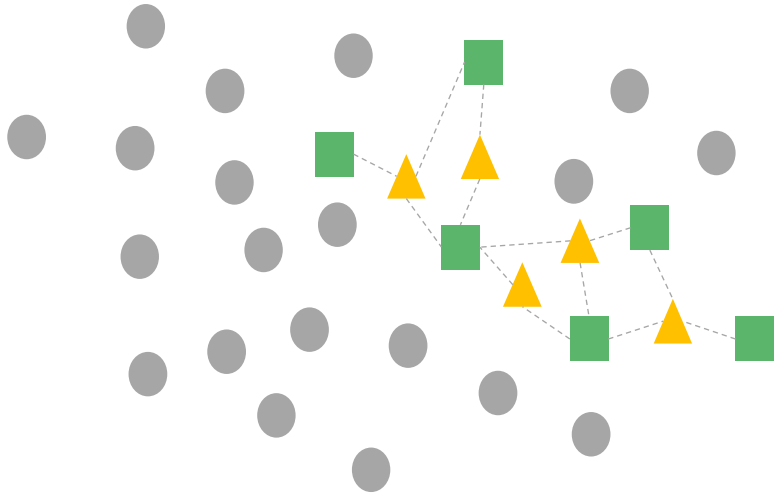
❖ SMOTE(Synthetic Minority Over-sampling Technique)

- 소수 클래스에서 가상의 데이터를 생성하는 방법
- 소수 클래스의 인스턴스를 단순 복제함으로 생기는 overfitting 문제를 보완

Data-level 방법: Synthetic sample 생성

❖ SMOTE(Synthetic Minority Over-sampling Technique)

- 소수 클래스에서 가상의 데이터를 생성하는 방법
- 소수 클래스의 인스턴스를 단순 복제함으로 생기는 overfitting 문제를 보완



: 다수 클래스

: 소수 클래스

: Synthetic 데이터

① x_i 기준으로 KNN방식으로 주변 K개 관측치 중 랜덤으로 하나의 관측치 (x_{zi})선택

② 수식을 통해 새로운 Synthetic 데이터 생성

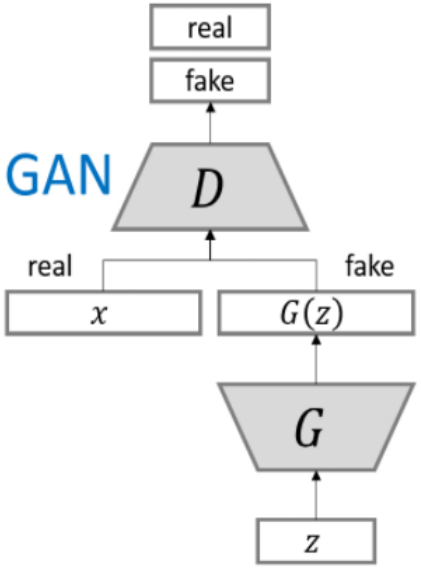
③ 소수 클래스내 모든 데이터에 대하여 반복

$$x_{synthetic} = x_i + \lambda(x_{zi} - x_i), \qquad \lambda \in [0, 1]$$

Data-level 방법: Synthetic sample 생성

❖ **GAN 활용**

- GAN은 가상의 데이터를 생성하는 Generator와 가상 데이터와 실제 데이터를 구분하는 Discriminator로 구성됨
- **Generator**: 최대한 실제 데이터같이 데이터를 만들어 Discriminator가 구별하기 힘들도록 학습
- **Discriminator**: 실제 데이터와 Generator가 만든 가상 데이터를 구별하도록 학습

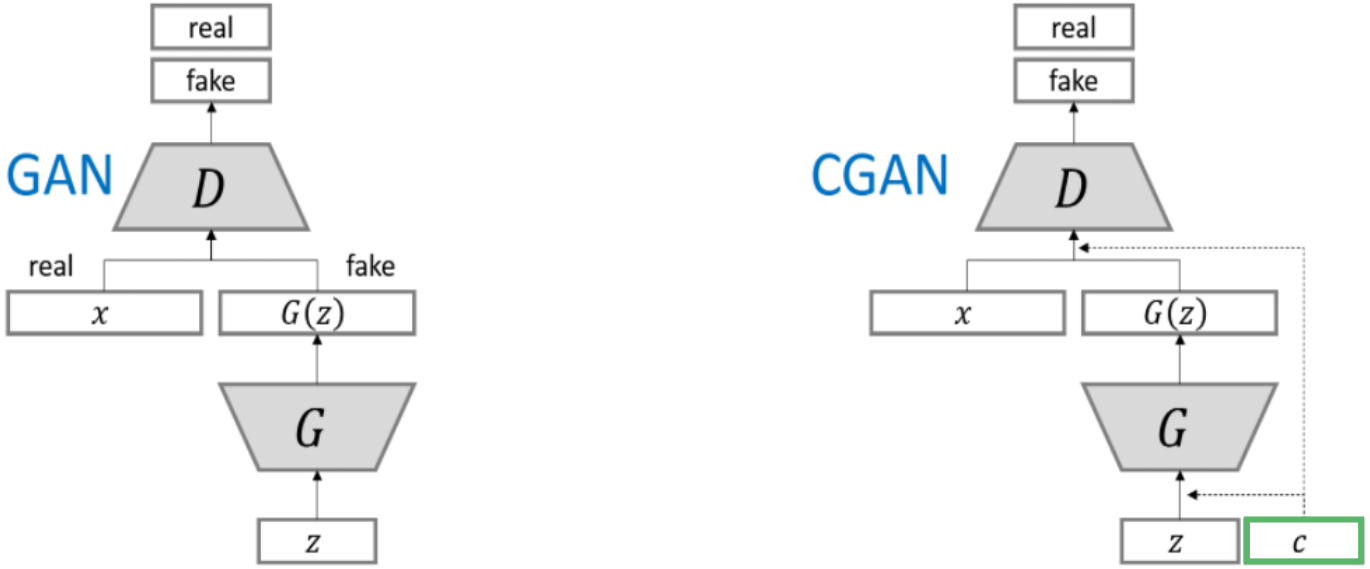


$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{data}(x)}[logD(x)] + \mathbb{E}_{x\sim P_z(z)}[\log(1-D(G(z)))]$$

Data-level 방법: Synthetic sample 생성

❖ **GAN 활용: Conditional GAN (CGAN)**

- Generator와 Discriminator를 활용하여 학습하는 방식은 같지만, **조건(c)**을 부여할 수 있음
- CGAN을 이용하면 우리가 원하는 class에 해당하는 가상 데이터를 생성 가능



$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{data}(x)}[logD(x|c)] + \mathbb{E}_{x\sim P_z(z)}[\log(1-D(G(z|c)))]$$

Algorithm-level 방법: Cost sensitive learning

❖ **Cost sensitive reweighting**

- 각 클래스별 개수를 반영하여 목적함수에 가중치를 부여
- 각 클래스별 데이터 개수의 반비례를 가중치로 설정
- Inverse frequency weight: $w = \frac{1}{f_i}$
- Square root weight: $w = \frac{1}{\sqrt{f_i}}$

f_i: 각 클래스별 데이터 개수

$$Loss_{CE} = -t_i \cdot \log(p_i)$$

$$Loss_{wCE} = -w_i \cdot t_i \cdot \log(p_i)$$

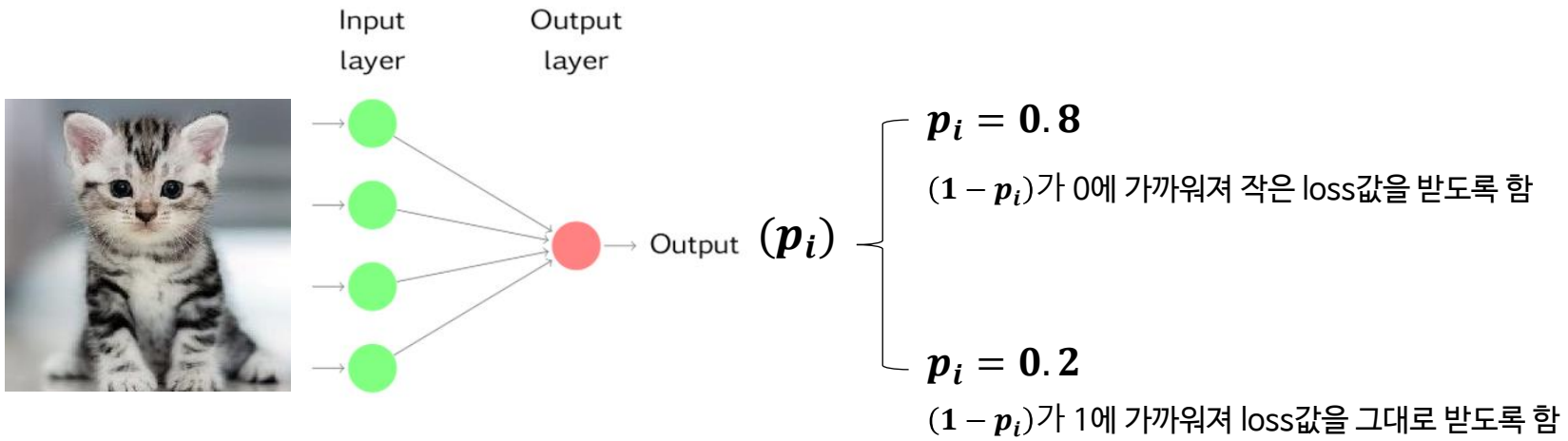
$$t_i = \begin{cases} 1\ (i = true\ label) \\ 0\ (i \neq true\ label) \end{cases}$$

Algorithm-level 방법: Cost sensitive learning

❖ Focal loss

- 맞추기 쉬운 샘플의 가중치를 줄이고 어려운 샘플에 대한 학습에 초점을 맞춤
- 가중치를 샘플 단위로 부여함
- γ 는 맞추기 쉬운 샘플에 대한 loss의 비중을 낮추는 역할

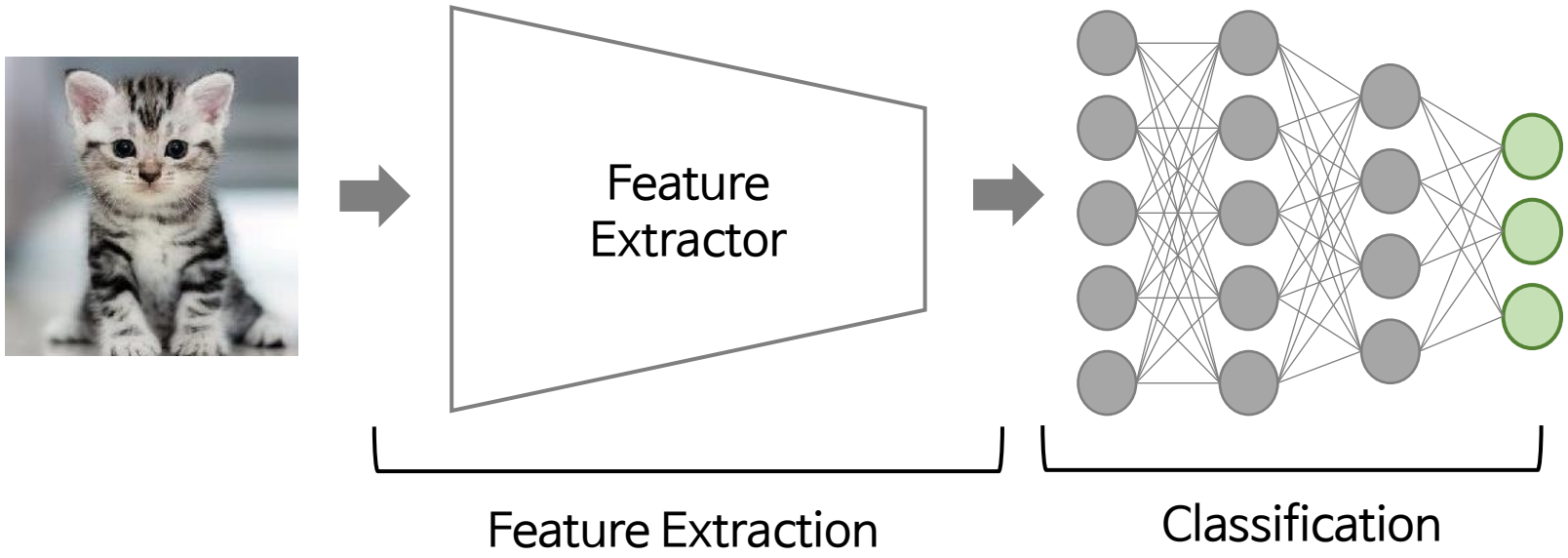
$$Loss_{focal} = -\sum_{i=1}^{C}(1-p_i)^{\gamma} \cdot t_i \cdot \log(p_i)$$



Algorithm-level 방법: Two Stage Training

❖ **Two Stage Training**

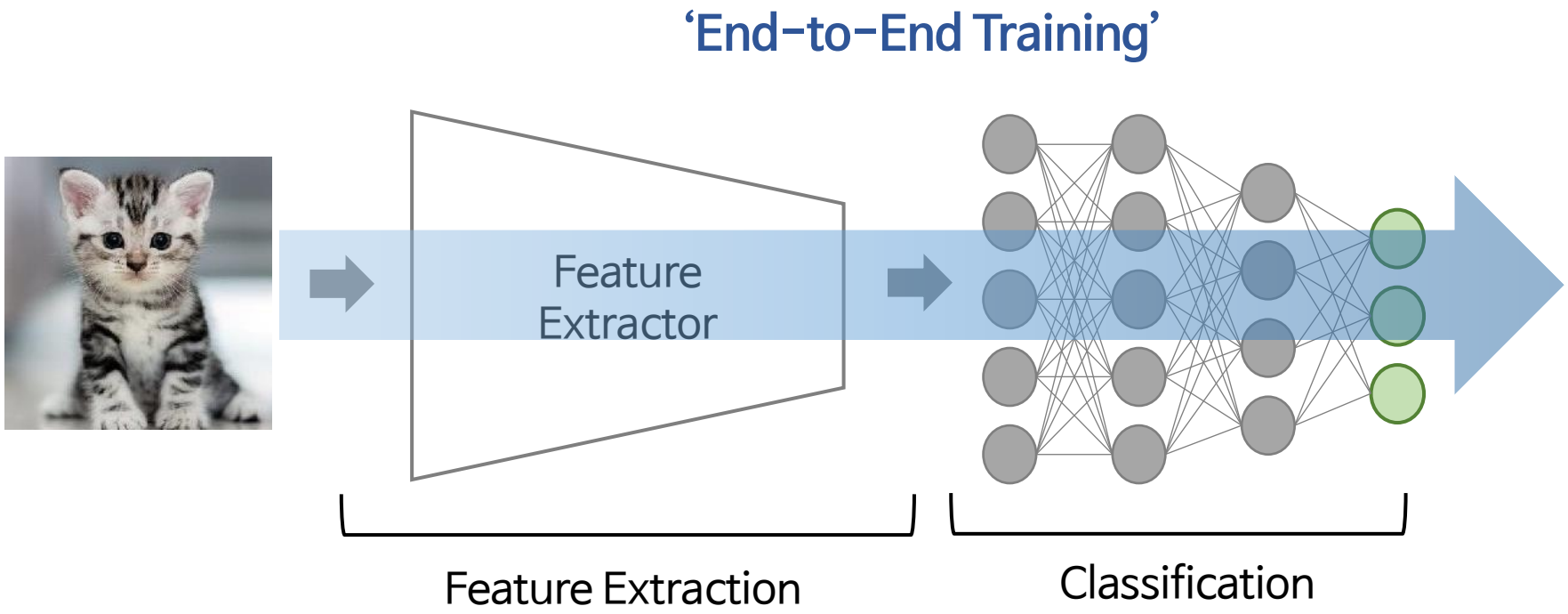
- Decoupling Representation and Classifier for Long-Tailed Recognition (ICLR, 2020)
- **학습 방법**에 주목함



Algorithm-level 방법: Two Stage Training

❖ Two Stage Training

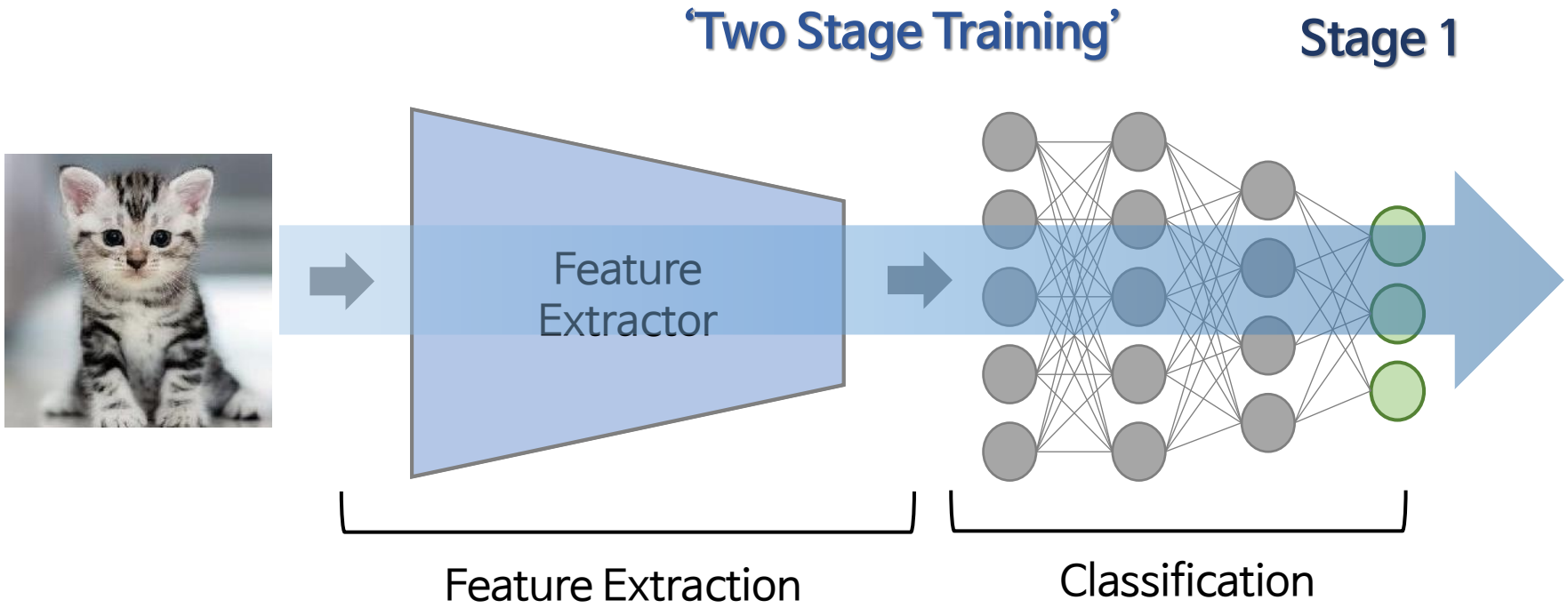
- Decoupling Representation and Classifier for Long-Tailed Recognition (ICLR, 2020)
- 기존 학습 방법 : Feature extractor 와 Classifier를 한번에 학습



Algorithm-level 방법: Two Stage Training

❖ Two Stage Training

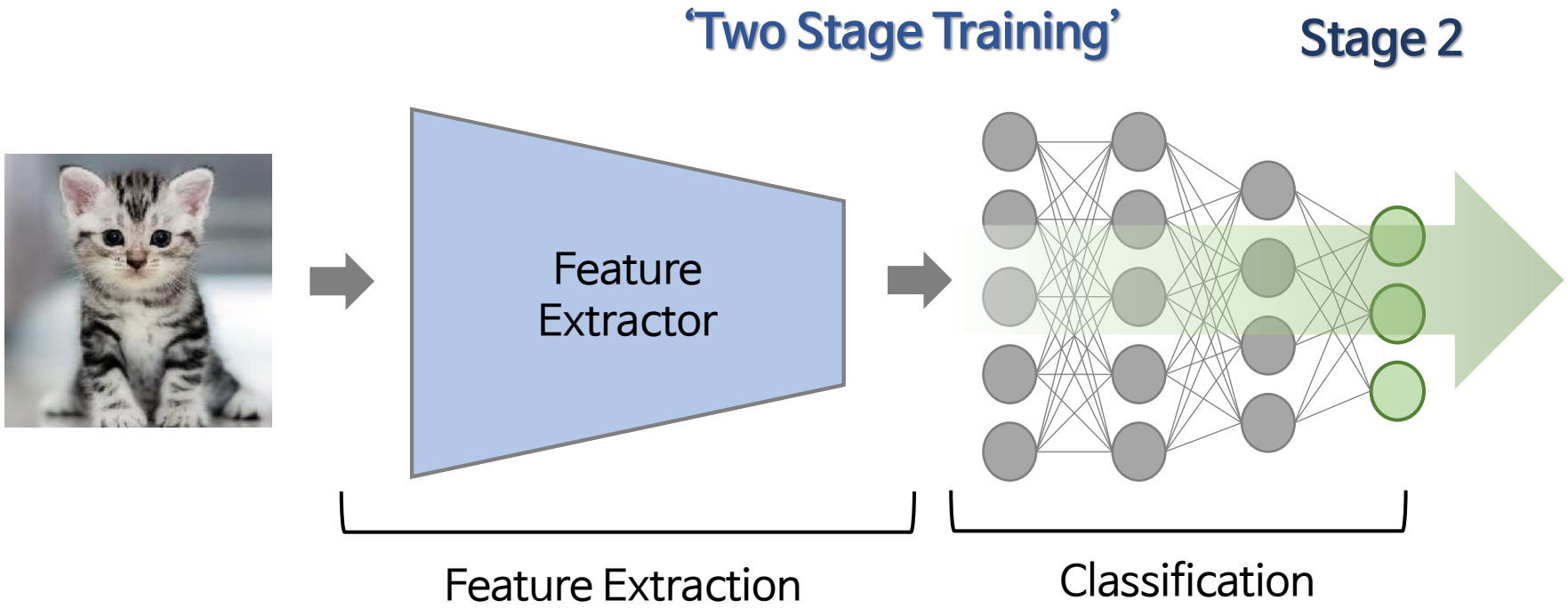
- Decoupling Representation and Classifier for Long-Tailed Recognition (ICLR, 2020)
- 제안 학습 방법: Classifier re-training
  - ➢ Stage 1: Feature extractor와 Classifier를 한번에 학습 시키는 End-to-End Training
  - ➢ Stage 2: Feature extractor부분은 고정하고, Classifier 부분만 재학습



Algorithm-level 방법: Two Stage Training

❖ Two Stage Training

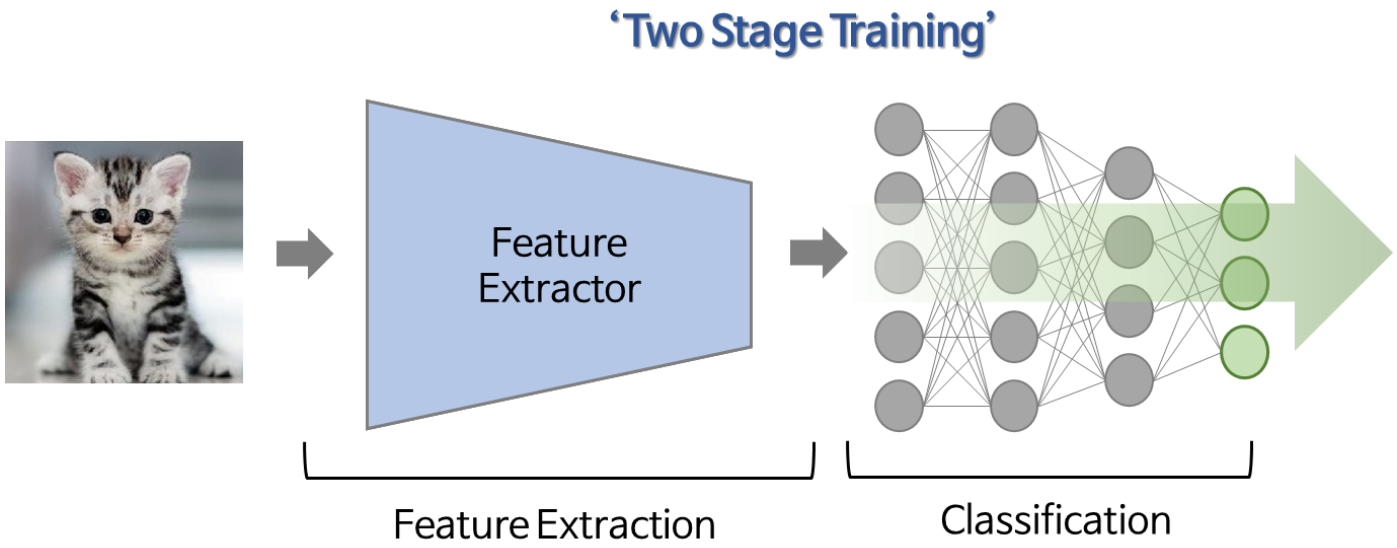
- Decoupling Representation and Classifier for Long-Tailed Recognition (ICLR, 2020)
- 제안 학습 방법: Classifier re-training
  - ➢ Stage 1: Feature extractor와 Classifier를 한번에 학습 시키는 End-to-End Training
  - ➢ Stage 2: Feature extractor부분은 고정하고, Classifier 부분만 재학습



Algorithm-level 방법: Two Stage Training

❖ Two Stage Training

- Decoupling Representation and Classifier for Long-Tailed Recognition (ICLR, 2020)
- 제안 학습 방법: Classifier re-training
- **기존 연구에서는 feature extractor을 학습 시키기에 충분하지 못한 양의 데이터에 초점**
- 분류경계선은 Classifier를 통해 형성 되기 때문에 **오히려 Classifier를 잘 학습시키는 것이 중요함**



예측 태스크 연구

불균형 연속형 데이터

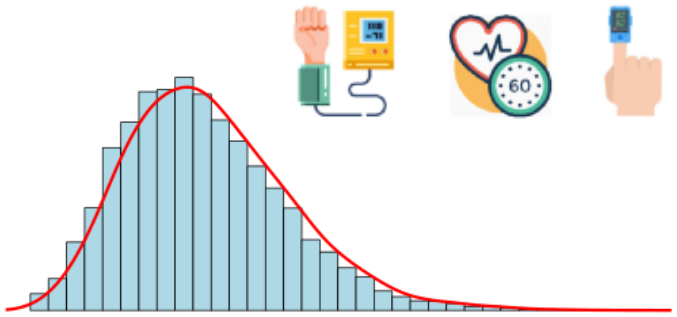
❖ 연속형 데이터에 대한 연구

- 불균형 데이터에서 학습하기 위한 기존 솔루션은 범주형 인덱스가 있는 대상에 중점
- 그러나 실제 데이터에서 **연속형**인 경우가 많음

연령 분포 데이터

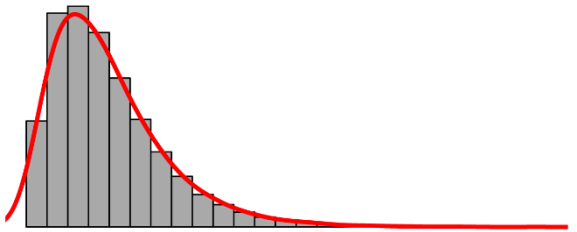
1세 ~ 100세

환자 활력 징후 데이터



혈압
맥박수
산소포화도

응급실 체류 시간 데이터



체류 시간

예측 태스크 관련 연구

❖ Delving into Deep Imbalanced Regression

- ICML Conference Long talk에서 2021년 발표된 논문
- 2021년 11월 13일 기준 9회 인용

Delving into Deep Imbalanced Regression

Yuzhe Yang [1] Kaiwen Zha [1] Ying-Cong Chen [1] Hao Wang [2] Dina Katabi [1]

Abstract

Real-world data often exhibit imbalanced distributions, where certain target values have significantly fewer observations. Existing techniques for dealing with imbalanced data focus on targets with categorical indices, i.e., different classes. However, many tasks involve continuous targets, where hard boundaries between classes do not exist. We define Deep Imbalanced Regression (DIR) as learning from such imbalanced data with continuous targets, dealing with potential missing data for certain target values, and generalizing to the entire target range. Motivated by the intrinsic difference between categorical and continuous label space, we propose distribution smoothing for both labels and features, which explicitly acknowledges the effects of nearby targets, and calibrates both label and learned feature distributions. We curate and benchmark large-scale DIR datasets from common real-world tasks in computer vi-

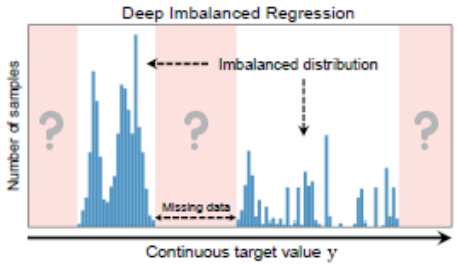


Figure 1. Deep Imbalanced Regression (DIR) aims to learn from imbalanced data with continuous targets, tackle potential missing data for certain regions, and generalize to the entire target range.

Existing solutions for learning from imbalanced data, however, focus on targets with categorical indices, i.e., the targets are different classes. However, many real-world tasks involve continuous and even infinite target values. For example, in vision applications, one needs to infer the age of

예측 태스크 관련 연구

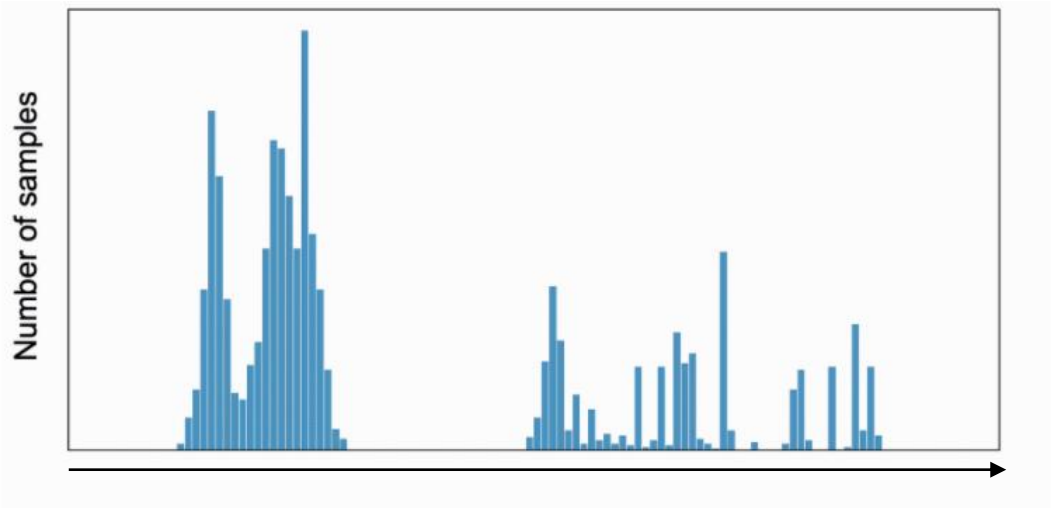
❖ 불균형 연속형 데이터 특징

- **기존 불균형 범주형 데이터와 다른 특징을 보임**

① 클래스 경계가 존재하지 않음: resampling, reweighting 방법을 적용하기 어려움

② 타겟값끼리 연속성 및 유사성: 주변값의 분포에 따라 다른 수준의 불균형을 겪음

③ 특정 대상값에 대한 데이터가 아예 없을 수 있음



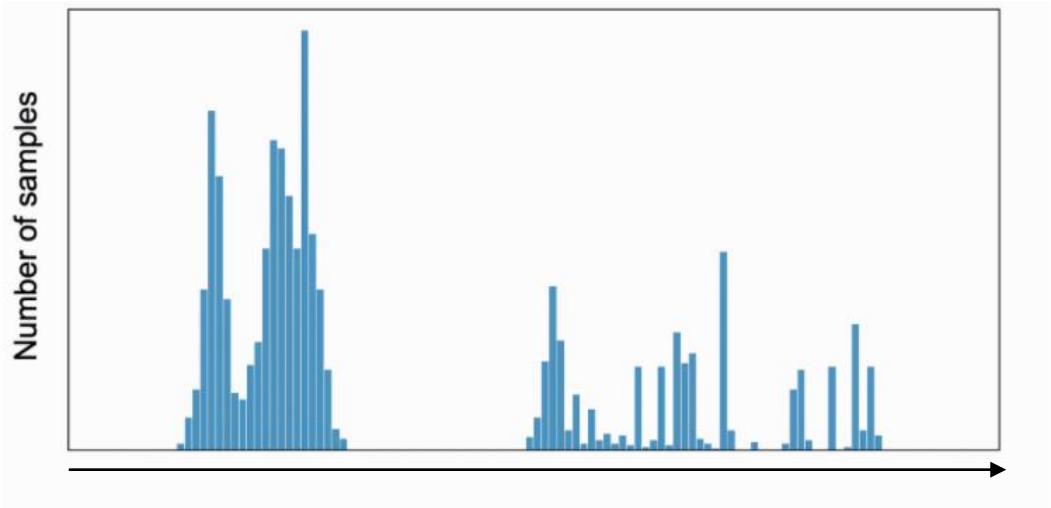
예측 태스크 관련 연구

❖ 불균형 연속형 데이터 특징

① **클래스 경계가 존재하지 않음: resampling, reweighting 방법을 적용하기 어려움**

② 타겟값끼리 연속성 및 유사성: 주변값의 분포에 따라 다른 수준의 불균형을 겪음

③ 특정 대상값에 대한 데이터가 아예 없을 수 있음



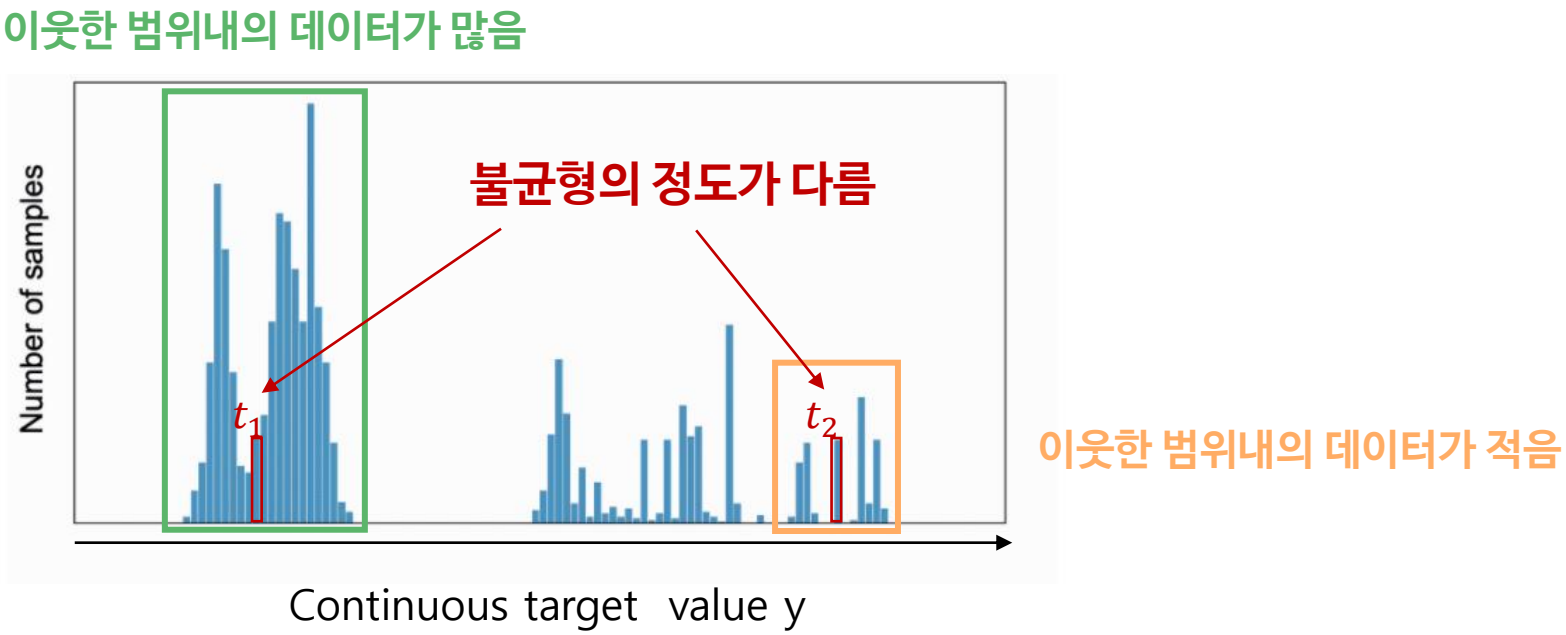
예측 태스크 관련 연구

❖ 불균형 연속형 데이터 특징

① 클래스 경계가 존재하지 않음: resampling, reweighting 방법을 적용하기 어려움

② 타겟값끼리 연속성 및 유사성: 주변값의 분포에 따라 다른 수준의 불균형을 겪음

③ 특정 타겟값에 대한 데이터가 아예 없을 수 있음



예측 태스크 관련 연구

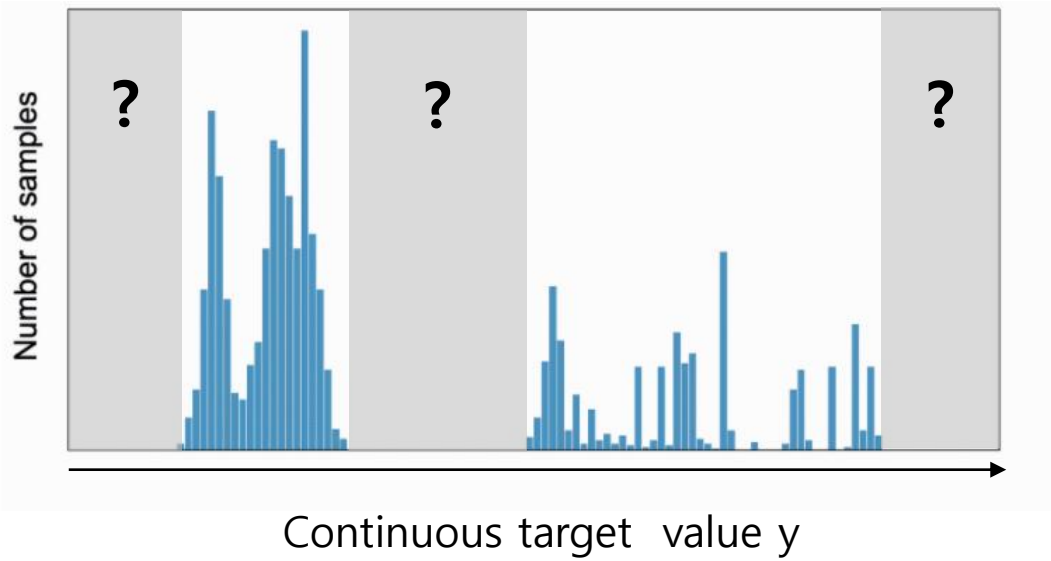
❖ 불균형 연속형 데이터 특징

① 클래스 경계가 존재하지 않음: resampling, reweighting 방법을 적용하기 어려움

② 타겟값끼리 연속성 및 유사성: 주변값의 분포에 따라 다른 수준의 불균형을 겪음

③ 특정 타겟값에 대한 데이터가 아예 없을 수 있음

➢ 주변 데이터를 통해 interpolation 또는 extrapolation 가능

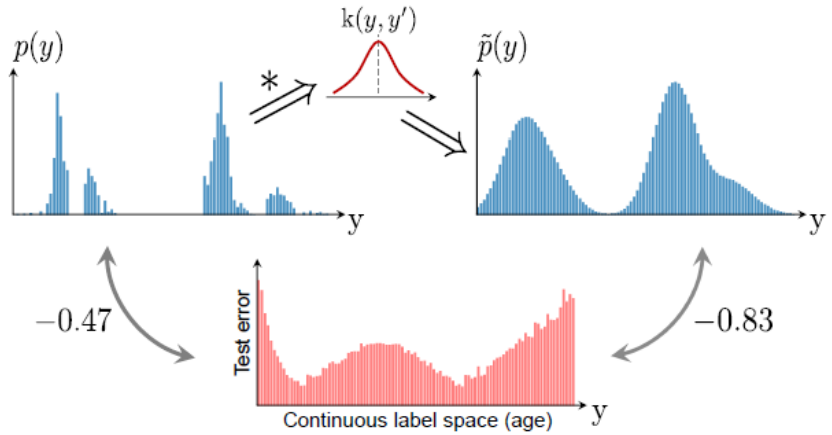


예측 태스크 관련 연구

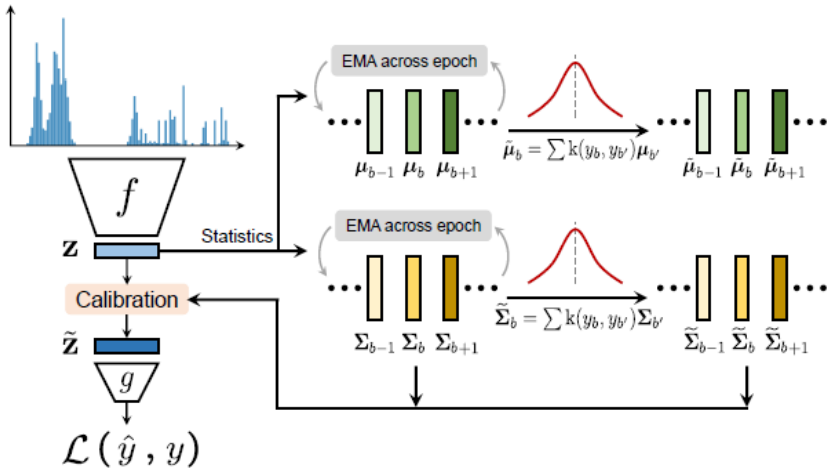
❖ 제안 방법론

- 인접 데이터간 **유사성**을 활용
- **커널 함수**를 활용하여 불균형 문제 해소
- Label Distribution Smoothing (LDS) : 레이블 공간 관점
- Feature Distribution Smoothing (FDS) : 특징 공간 관점

LDS



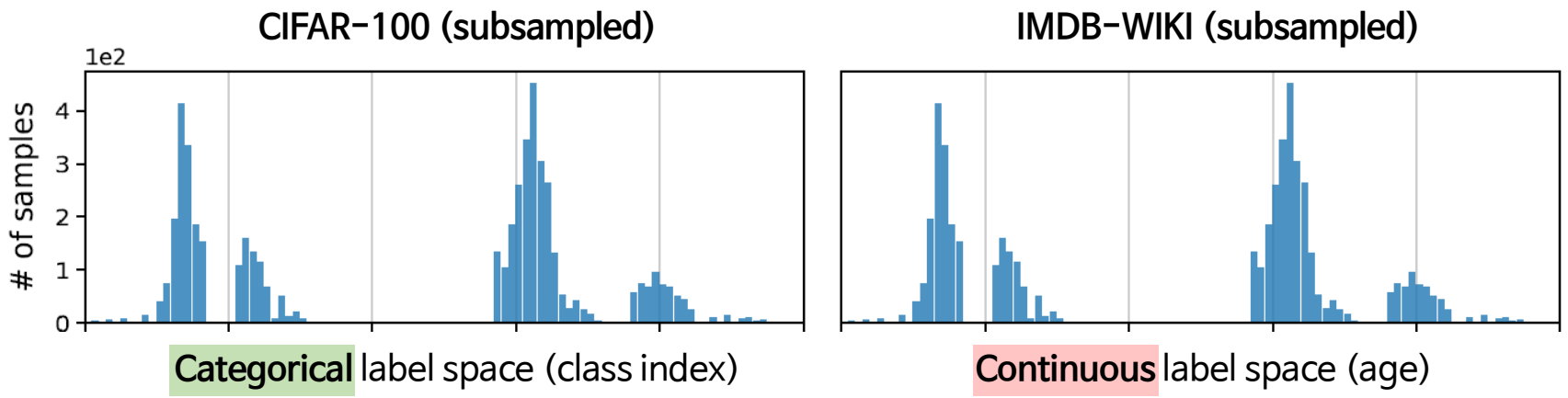
FDS



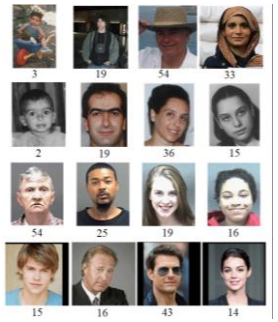
예측 태스크 관련 연구

❖ Label Distribution Smoothing (LDS)

- Motivation: 연속형 데이터의 학습 결과는 범주형 데이터의 학습 결과와 다소 다른 양상을 보임



레이블: 100가지 클래스



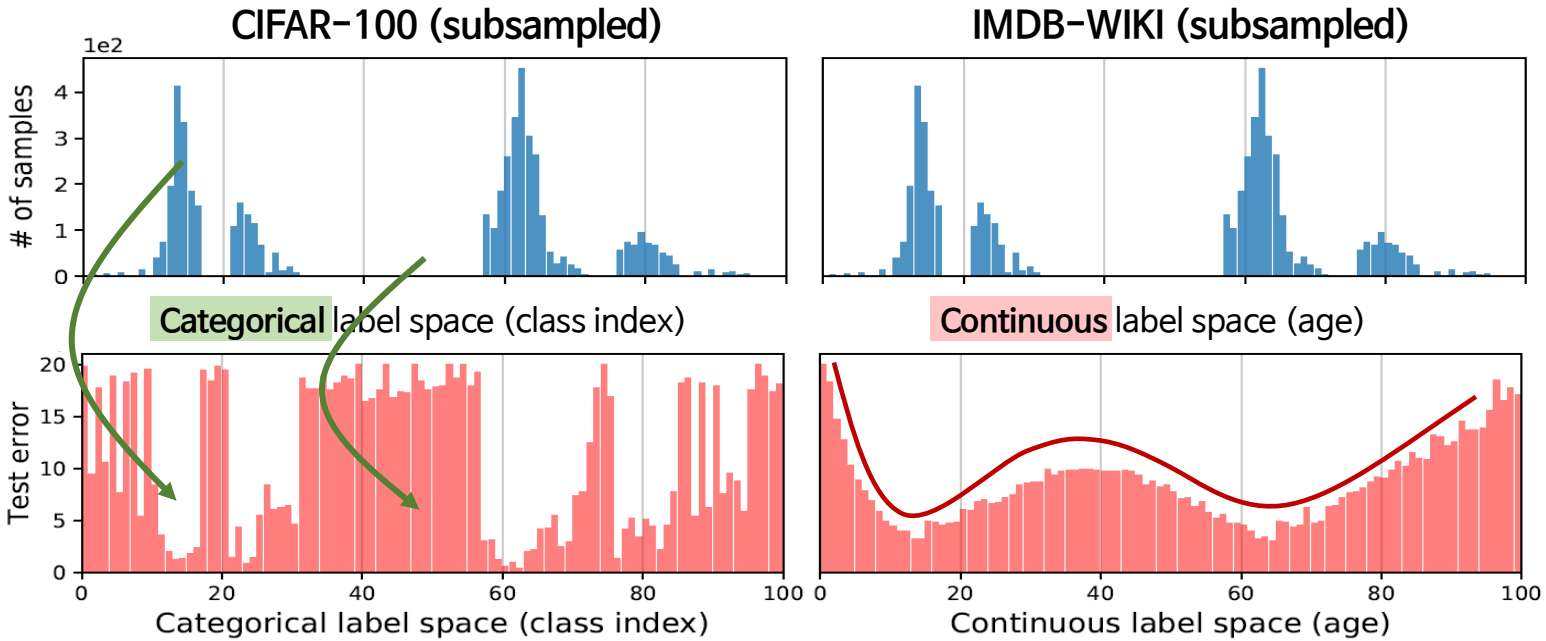
레이블: 0~99세까지 연령

예측 태스크 관련 연구

❖ Label Distribution Smoothing (LDS)

- Motivation: 연속형 데이터의 학습 결과는 범주형 데이터의 학습 결과와 다소 다른 양상을 보임
- 불균형이 심한 범주형 데이터는 오분류율 분포에 불균형의 정도가 그대로 반영됨
- **반면에, 연속형 데이터의 오분류율 분포는 불균형 정도를 정확하게 반영하지 않음**

상관계수
−0.76

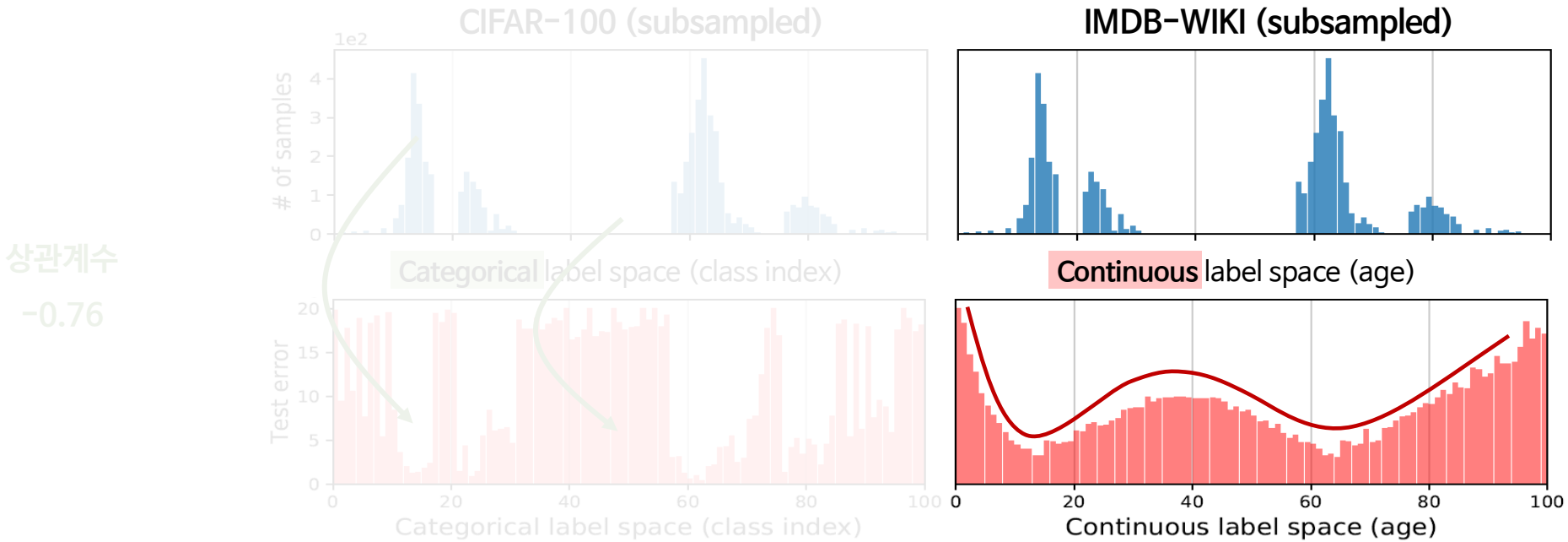


상관계수
−0.47

예측 태스크 관련 연구

❖ Label Distribution Smoothing (LDS)

- Motivation: 연속형 데이터의 학습 결과는 범주형 데이터의 학습 결과와 다소 다른 양상을 보임
- 불균형이 심한 범주형 데이터는 오분류율 분포에 불균형의 정도가 그대로 반영됨
- **반면에, 연속형 데이터의 오분류율 분포는 불균형 정도를 정확하게 반영하지 않음**



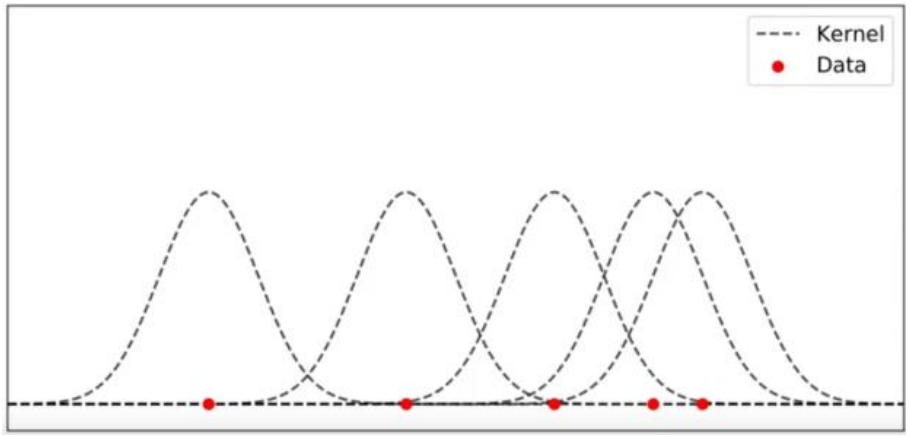
Empirical label distribution ≠ Real label density distribution

주변 레이블간 연관성을 가짐

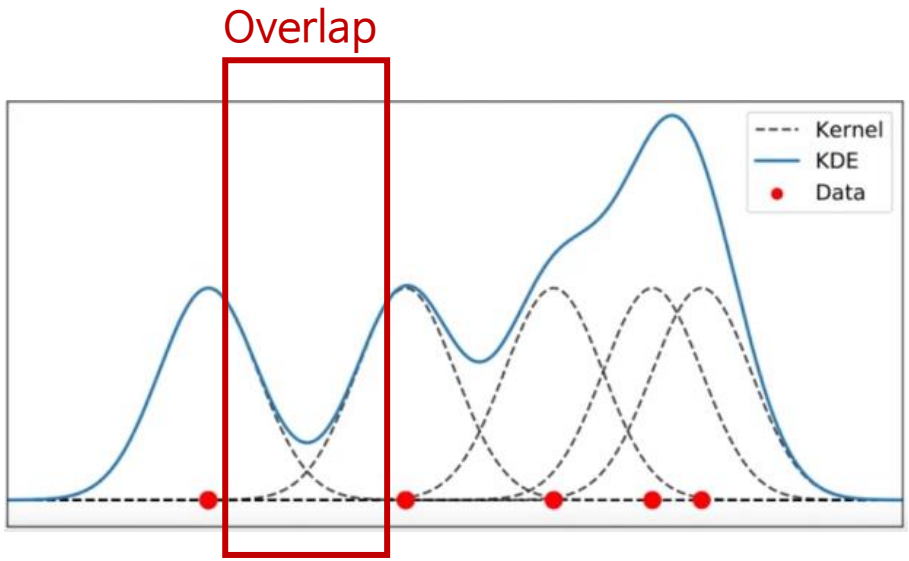
예측 태스크 관련 연구

❖ Label Distribution Smoothing (LDS)

- '커널 밀도 추정' 개념 활용함
- 커널 함수(K): 원점을 중심으로 대칭이며 적분값이 1인 non-negative 함수 (e.g. Gaussian 함수)
- 커널 밀도 추정: 커널 함수를 통해 부드러운 확률밀도함수를 추정



x와 다른 모든 데이터 포인트들의 커널 함수를 생성



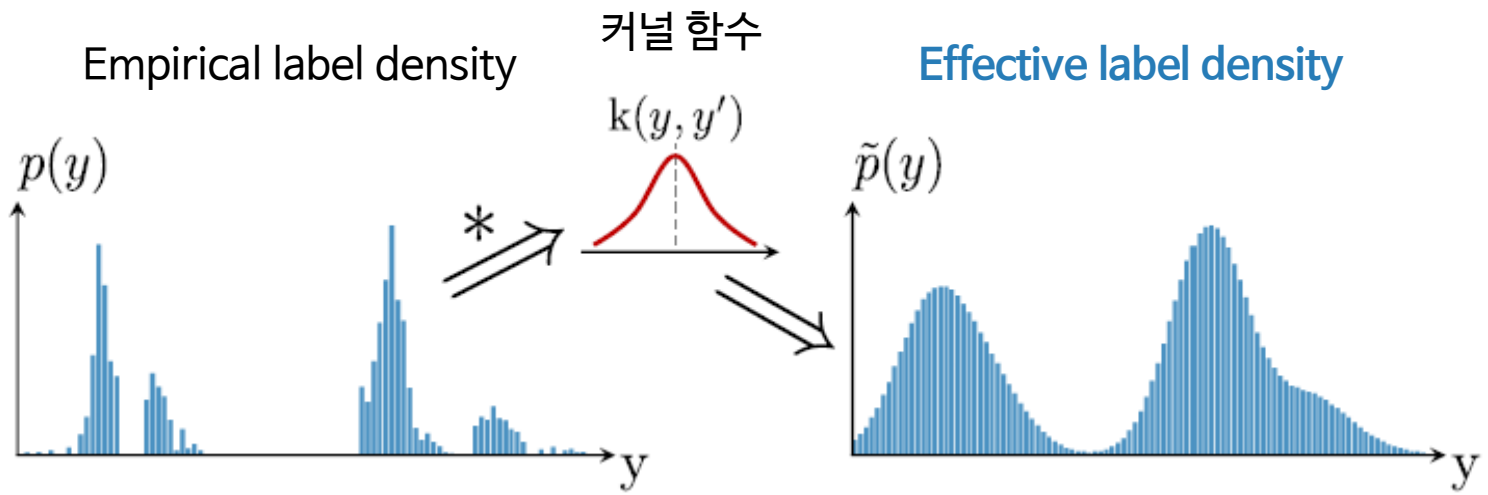
모든 커널 함수를 더한 후 전체 데이터 개수로 나눔

$$f_x(x) = \frac{1}{N}\sum_{i=1}^{N} K(x - x_i)$$

예측 태스크 관련 연구

❖ Label Distribution Smoothing (LDS)

- 커널 밀도 추정: 커널 함수를 통해 부드러운 확률밀도함수를 추정
- LDS을 통해 주변 데이터의 연속성을 반영한 'Kernel-smoothed' 버전을 추출

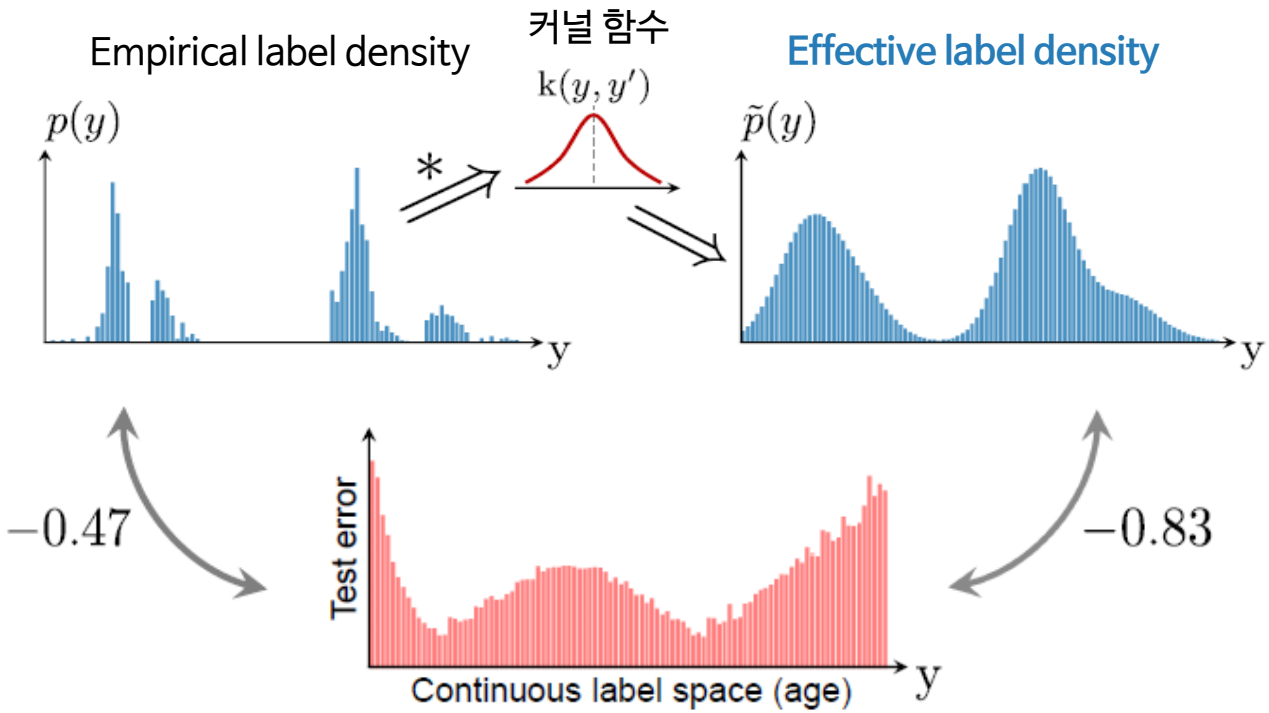


$$\tilde{p}(y') \triangleq \int_y K(y, y') p(y)\, dy$$

예측 태스크 관련 연구

❖ Label Distribution Smoothing (LDS)

- 커널 밀도 추정: 커널 함수를 통해 부드러운 확률밀도함수를 추정
- LDS을 통해 주변 데이터의 **연속성**을 반영한 'Kernel-smoothed' 버전을 추출
- **Reweighting에 활용: 손실함수의 가중치** $w_i = \frac{1}{\tilde{p}(y_i)}$

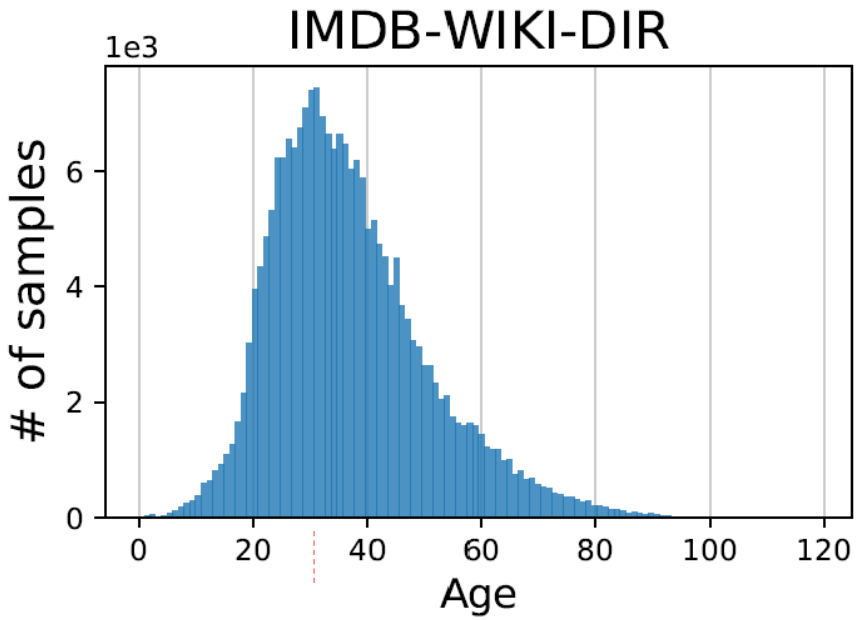


예측 테스크에 영향을 미치는 실제 불균형 정도를 반영함

예측 태스크 관련 연구

❖ Feature Distribution Smoothing (FDS)

- Motivation: 타겟 공간에서의 연속성은 잘 학습된 모델의 특징공간에도 반영됨



bin = 타겟 공간을 b개로 나누는 동일한 간격 (e.g. 연령: 1살)

$$y_{b+1} - y_b = 1$$

예측 태스크 관련 연구

❖ Feature Distribution Smoothing (FDS)

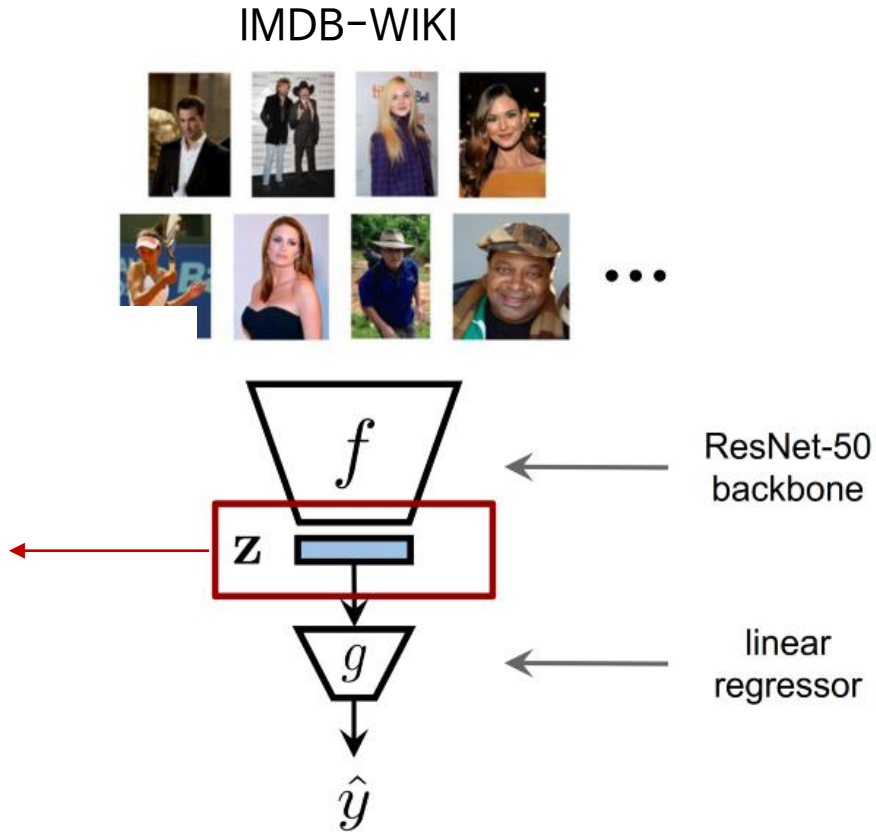
- Motivation: 타겟 공간에서의 연속성은 잘 학습된 모델의 특징공간에도 반영됨
- **학습된 feature space (z)를 통해 모든 b의 평균, 분산을 계산할 수 있음**

Feature statistics

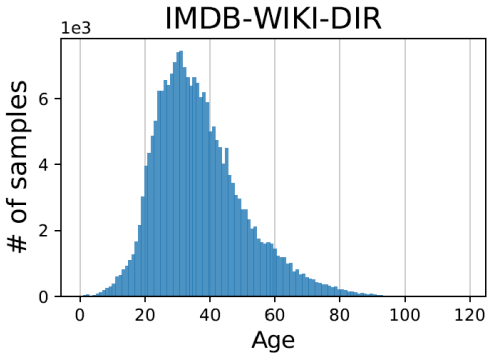
z_i: feature element
N_b: total number of samples in b−th bin

$$\mu_b = \frac{1}{N_b}\sum_{i=1}^{N_b} z_i$$

$$\Sigma_b = \frac{1}{N_b - 1}\sum_{i=1}^{N_b} (z_i - \mu_b)\,(z_i - \mu_b)^\top$$



예측 태스크 관련 연구

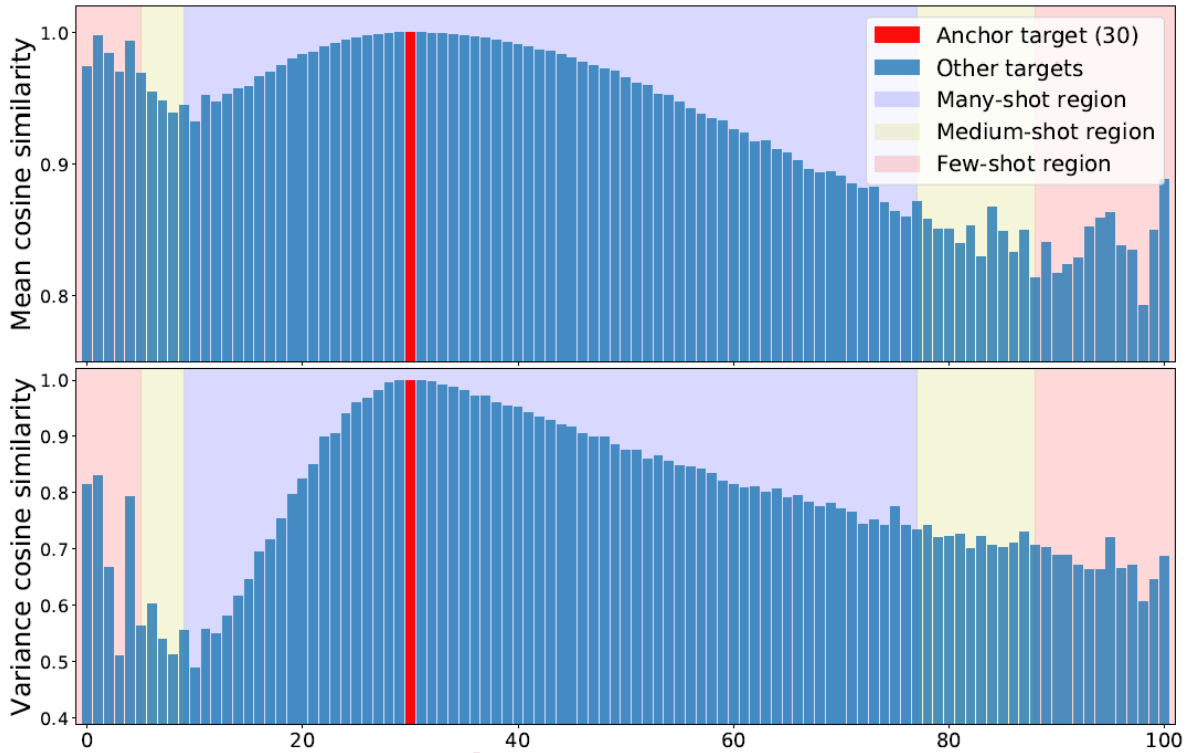


❖ Feature Distribution Smoothing (FDS)

- Motivation: 타겟 공간에서의 연속성은 잘 학습된 모델의 특징공간에도 반영됨
- b_0 **(Anchor target)과 다른** b**의 평균, 분산의 코사인 유사도 계산**

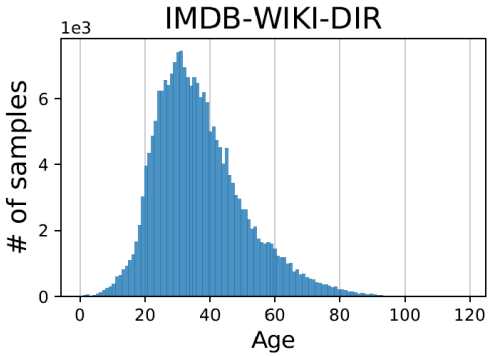
$\cos(\mu_b, \mu_{b_0})$

$\cos(\sigma_b, \sigma_{b_0})$



Anchor target b_0 = 30살

예측 태스크 관련 연구

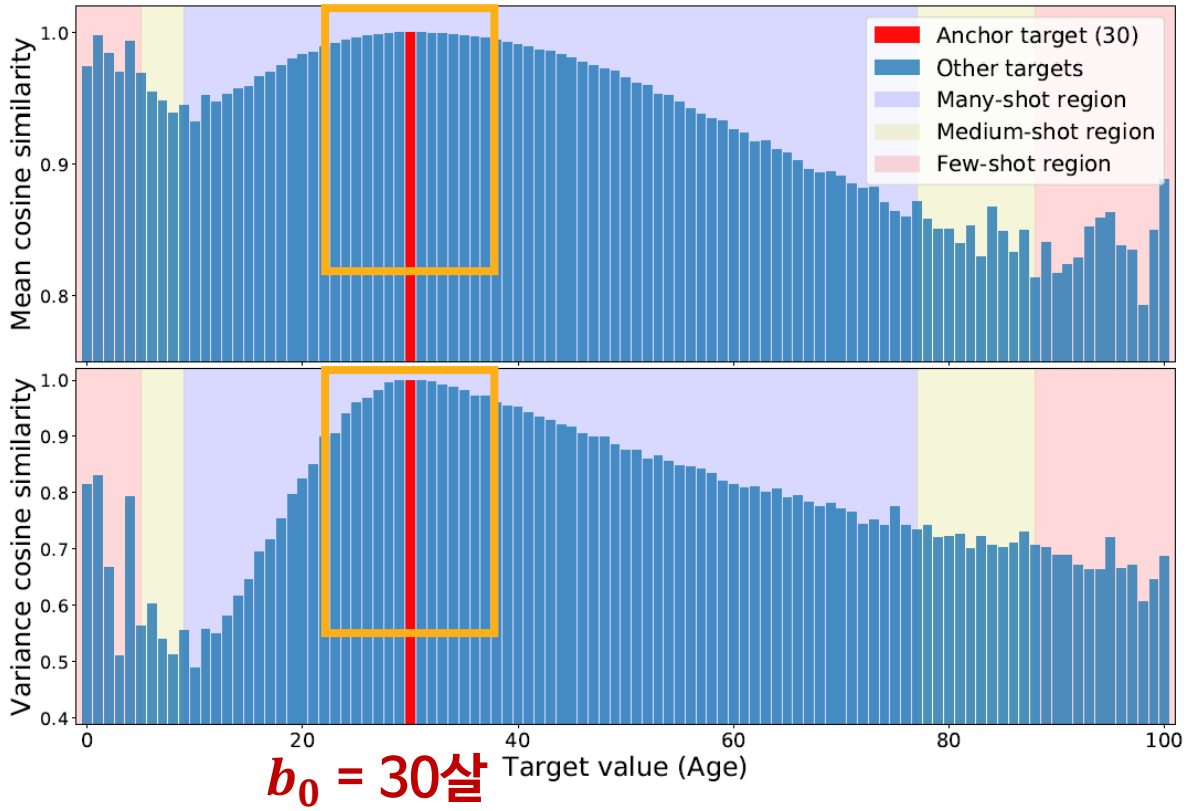


❖ Feature Distribution Smoothing (FDS)

- Motivation: 타겟 공간에서의 연속성은 잘 학습된 모델의 특징공간에도 반영됨
- b_0 **(Anchor target)과 다른** b**의 평균, 분산의 코사인 유사도 계산**

$\cos(\mu_b, \mu_{b_0})$

$\cos(\sigma_b, \sigma_{b_0})$

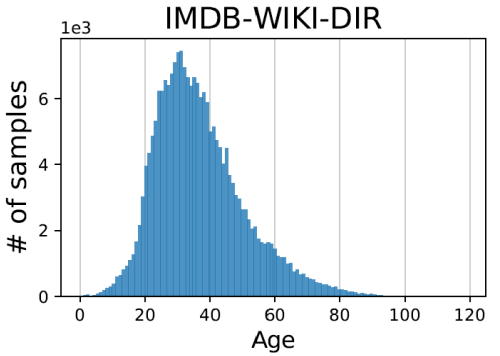


b_0 = 30살

30살 vs 20대 후반, 30대 초반

유사한 나이대의 평균, 분산은 높은 유사도를 보임

예측 태스크 관련 연구

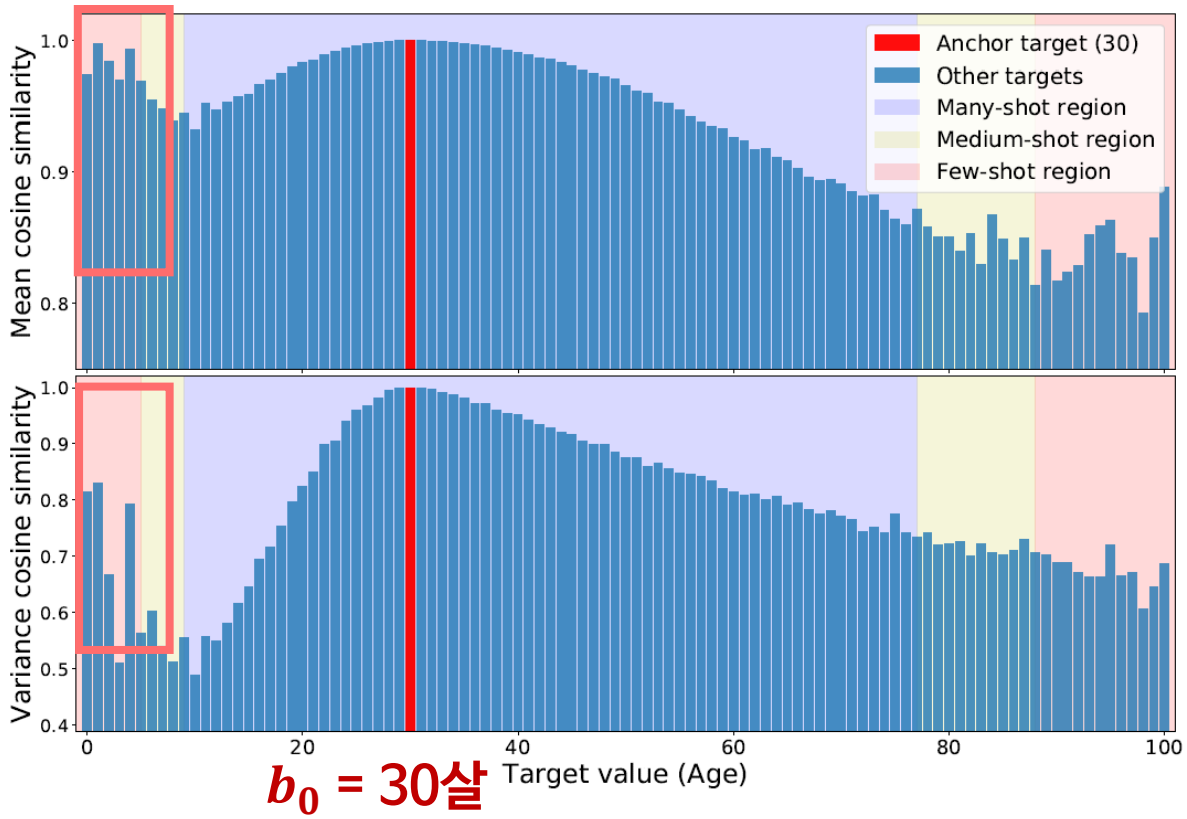


❖ Feature Distribution Smoothing (FDS)

- Motivation: 타겟 공간에서의 연속성은 잘 학습된 모델의 특징공간에도 반영됨
- b_0 **(anchor target)과 다른** b**의 평균, 분산의 코사인 유사도 계산**
- **데이터 불균형의 영향을 feature space에서도 확인 할 수 있음**

$\cos(\mu_b, \mu_{b_0})$

$\cos(\sigma_b, \sigma_{b_0})$



b_0 = 30살

30살 vs 20대 후반, 30대 초반

유시한 나이대의 평균, 분산은 높은 유사도를 보임

30살 vs 6세 이하

데이터의 수가 확연히 적은 0~6세와의 유사도가 높게 나옴

예측 태스크 관련 연구

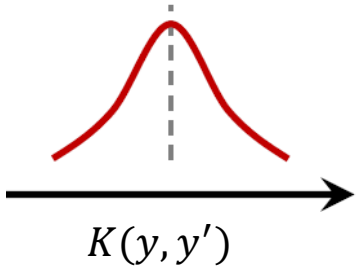
❖ Feature Distribution Smoothing (FDS)

- **LDS와 마찬가지로 커널 함수 이용해 스무딩**

$$\mu_b = \frac{1}{N_b}\sum_{i=1}^{N_b} z_i$$

$$\Sigma_b = \frac{1}{N_b - 1}\sum_{i=1}^{N_b}(z_i - \mu_b)\,(z_i - \mu_b)^\top$$

커널 함수



$$\tilde{\mu}_b = \sum_{b' \in \mathcal{B}} K(y_b, y_{b'})\,\mu_{b'}$$

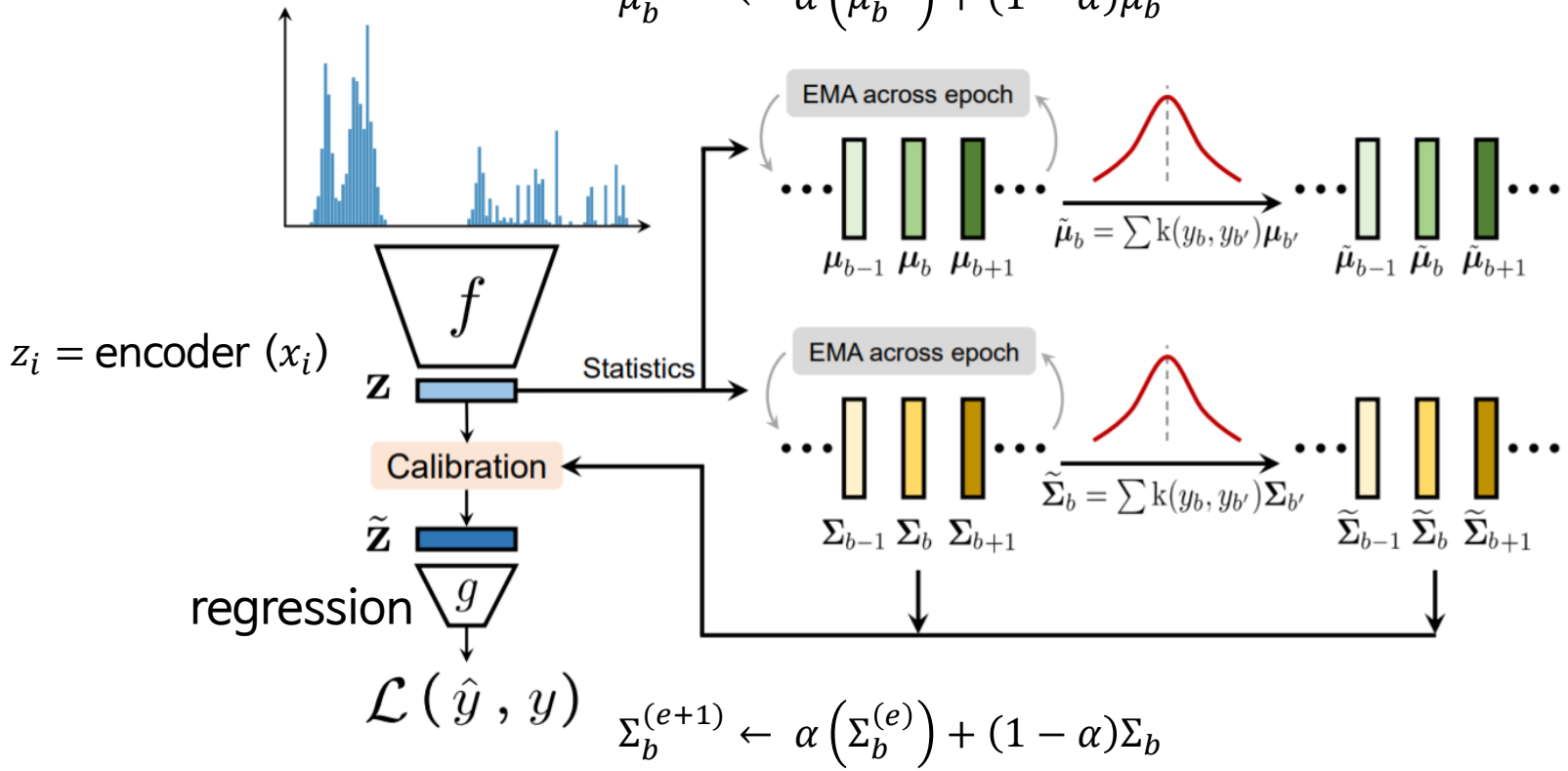
$$\tilde{\Sigma}_b = \sum_{b' \in \mathcal{B}} K(y_b, y_{b'})\,\Sigma_{b'}$$

예측 태스크 관련 연구

❖ Feature Distribution Smoothing (FDS)

- 특징공간 z 다음에 calibration 레이어를 추가하여 FDS를 진행
- 현재 에폭의 통계량을 추정하기 위해 모멘텀 업데이트 방식을 활용

$$\mu_b^{(e+1)} \leftarrow \alpha\left(\mu_b^{(e)}\right) + (1-\alpha)\mu_b$$



$$\Sigma_b^{(e+1)} \leftarrow \alpha\left(\Sigma_b^{(e)}\right) + (1-\alpha)\Sigma_b$$

결론

- 불균형 데이터를 처리하기 위해 분류, 예측 태스크로 구분하여 연구동향을 살펴봄
- 데이터 레벨 방법: Random resampling, Synthetic sample 등
- 알고리즘 레벨 방법: Cost sensitive learning, Two stage training, LDS, FDS
- 데이터 형태에 따른 특징을 고려하여 연구 진행
- 앞으로도 연속형 데이터, 시계열 데이터에 대한 연구가 활발히 진행되길 희망함

참고문헌

- Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357
- Lin, Tsung-Yi, et al. "Focal loss for dense object detection." Proceedings of the IEEE international conference on computer vision. 2017..
- Kang, Bingyi, et al. "Decoupling representation and classifier for long-tailed recognition." arXiv preprint arXiv:1910.09217 (2019).
- Yang, Yuzhe, et al. "Delving into Deep Imbalanced Regression." arXiv preprint arXiv:2102.09554 (2021).
- https://github.com/YyzHarry/imbalanced-regression
- https://towardsdatascience.com/strategies-and-tactics-for-regression-on-imbalanced-data-61eeb0921fca
- https://www.youtube.com/watch?v=Vhwz228VrIk&t=1277s

Thank you