DMQA Open Seminar

# Handling Imbalanced data

정진용

Data Mining & Quality Analytics Lab

2022.05.27(금)

고려대학교 KOREA UNIVERSITY

Data Mining Quality Analytics

# 발표자 소개

- **정진용 (Jinyong Jeong)**
  - ✔ 고려대학교 산업경영공학과
  - ✔ Data Mining & Quality Analytics Lab. (김성범 교수님)
  - ✔ 석박통합과정(2021.09~)
- **관심분야**
  - ✔ Machine learning / Deep learning Algorithms
  - ✔ Semi-Supervised Learning
- **이메일**
  - ✔ jy_jeong@korea.ac.kr

# 목차

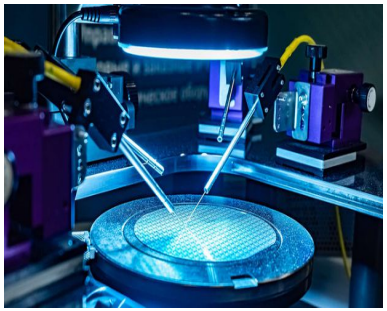
# Introduction

불균형 데이터란?

## ❖ 불균형 데이터란?

- 데이터 내에서 각 클래스가 갖고 있는 데이터의 양에 차이가 큰 경우
- 다수 클래스는 데이터 셋에서 상대적으로 큰 비중을 차지하는 클래스
- 소수 클래스는 데이터 셋에서 상대적으로 작은 비중을 차지하는 클래스

**제조 공정 데이터**



정상 불량

**보험료 청구 데이터**


정상 사기

**질병 진단 데이터**

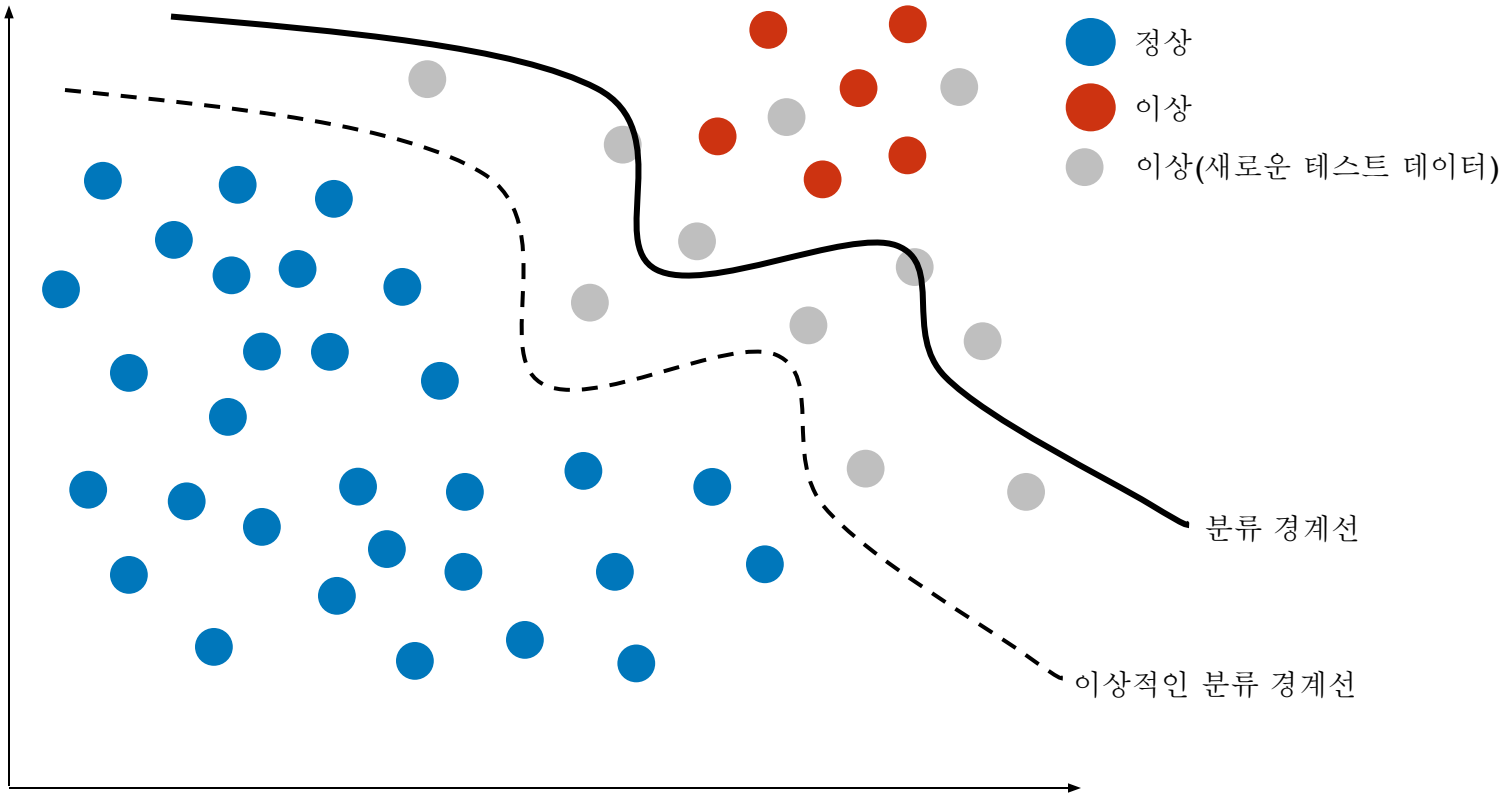
음성 양성

# Introduction

불균형 데이터의 문제점

## ❖ 불균형 데이터의 문제점

- 일반적으로 다수 정상 클래스보다 소수의 이상 클래스를 정확히 분류하는 것이 중요
- 다수의 정상 클래스에 의해 적절한 분류 경계선이 생성되지 못함

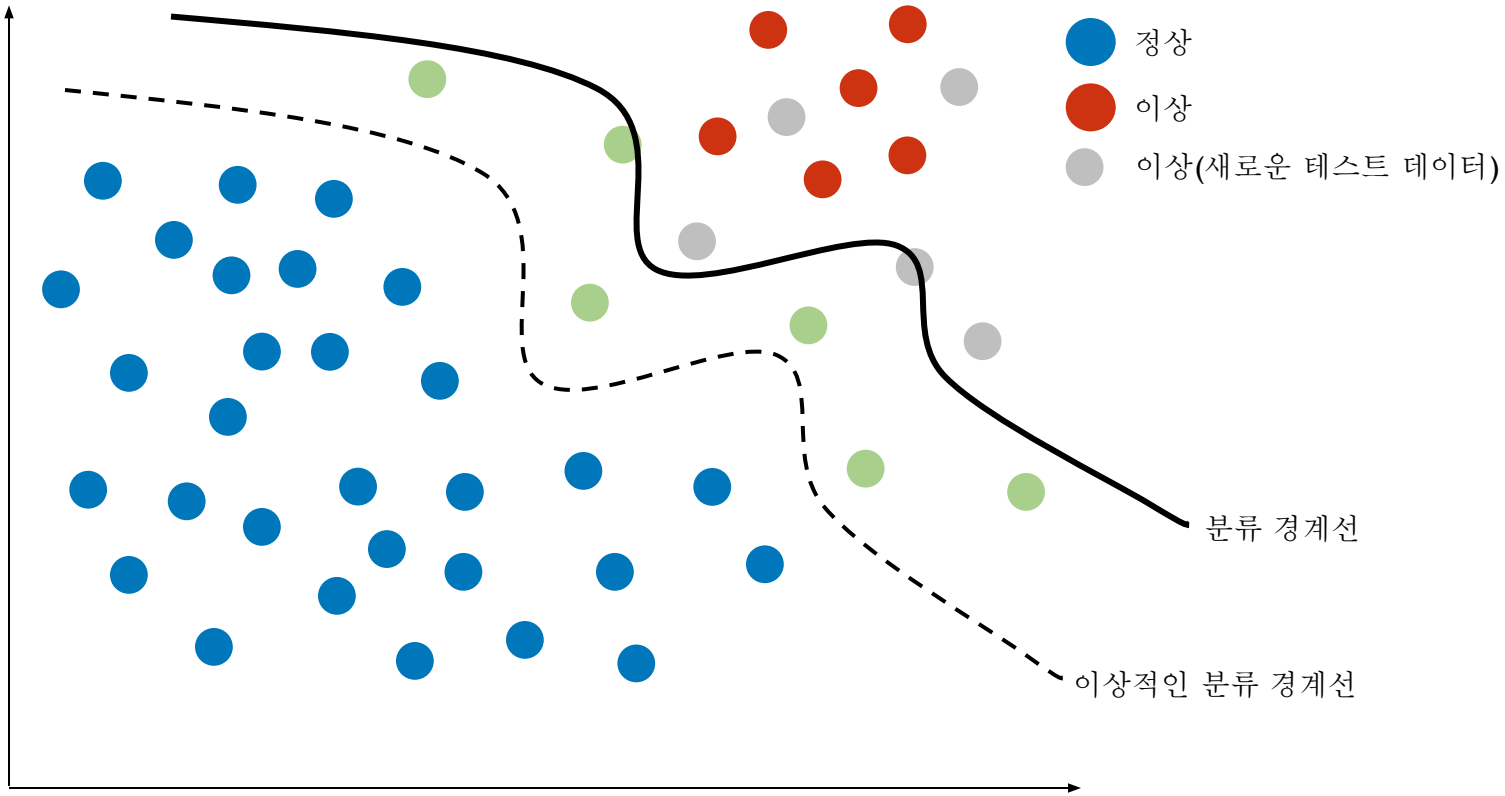


# Introduction

불균형 데이터의 문제점

## ❖ 불균형 데이터의 문제점

- 일반적으로 다수 정상 클래스보다 소수의 이상 클래스를 정확히 분류하는 것이 중요
- 다수의 정상 클래스에 의해 적절한 분류 경계선이 생성되지 못함



# Introduction

불균형 데이터의 문제점

❖ **불균형 문제에서 사용하는 모델 성능 지표**

- 소수 클래스를 잘 분류하지 못하는 모델이 학습되지만 높은 정확도를 보임
- F1 measure, Recall 등 소수 클래스 분류를 고려하는 지표를 사용해야 함

| **정오 행렬 (Confusion matrix)** | | 모델 예측 | |
|---|---|---|---|
| | | 정상 | 이상 |
| 실제 | 정상 | 85 | 5 |
| | 이상 | 5 | 5 |

$$\text{정확도} = \frac{85+5}{85+5+5+5} = 0.9$$

그러나 소수 클래스 절반 분류 못함

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + reall} = 2 \cdot \frac{0.5 \cdot 0.5}{0.5+0.5} = 0.5$$

# Introduction

## ❖ 불균형 데이터 해결 방안

- 크게 Data-level과 Algorithm-level로 구분 가능함

### ① Data-level 방법

- Under Sampling 기법
  - ✔ Random under sampling
  - ✔ CNN
  - ✔ Tomek-link
- Over Sampling 기법
  - ✔ Random over sampling
  - ✔ SMOTE
  - ✔ Borderline SMOTE
  - ✔ ADASYN
  - ✔ Major-to-minor Translation

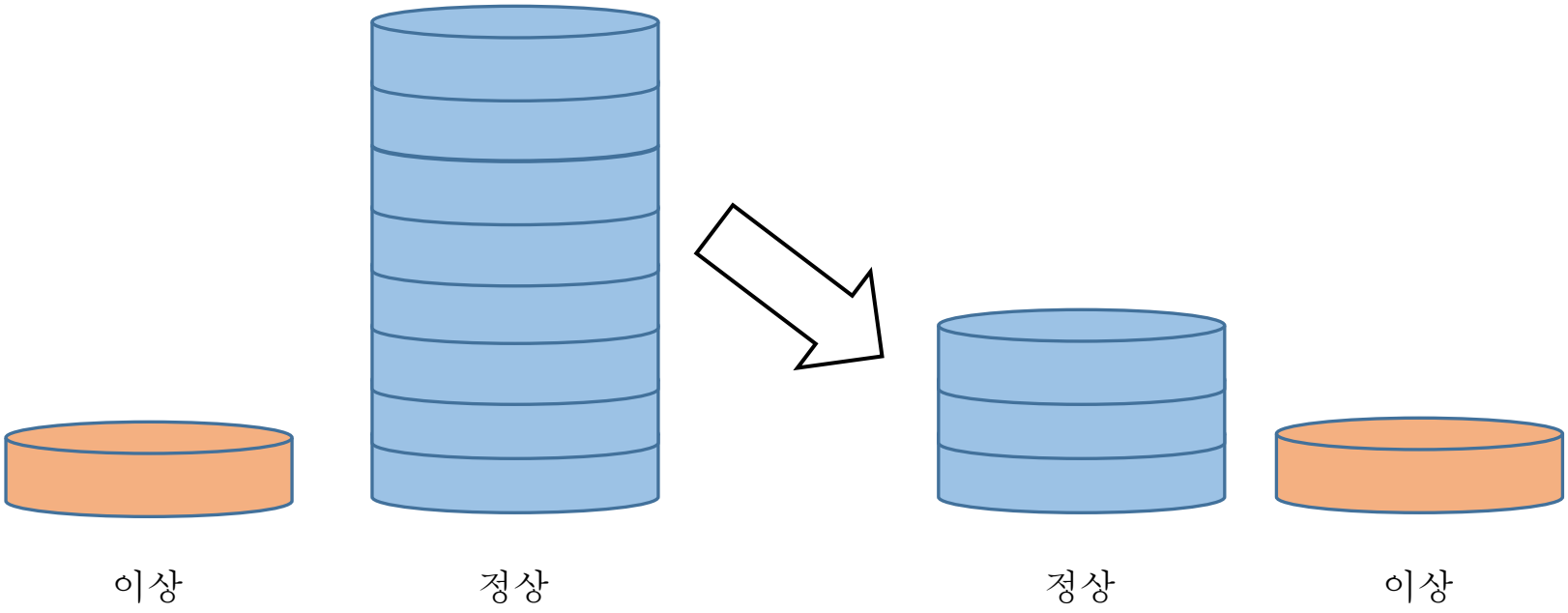
### ② Algorithm-level 방법

- Cost-sensitive learning
  - ✔ WCE Loss
  - ✔ Focal loss
  - ✔ Class Balanced Loss
  - ✔ LDAM Loss
- Two-stage training

# Introduction

❖ **Data-Level 방법**

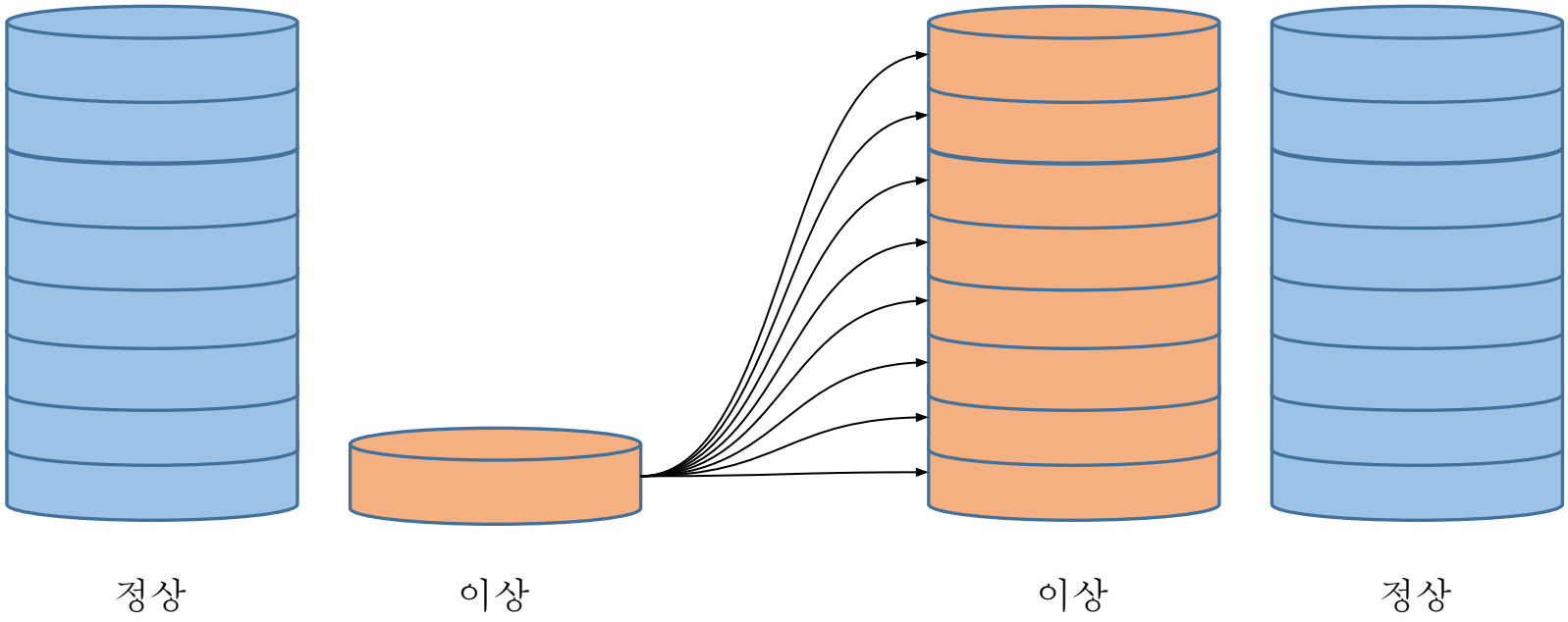
- Random under sampling : 다수 클래스의 비중을 줄이기 위해 다수 클래스 샘플 무작위 제거
- 중요한 정보를 가지고 있는 다수 클래스의 샘플이 제거 될 수 있음



# Introduction

❖ **Data-Level 방법**

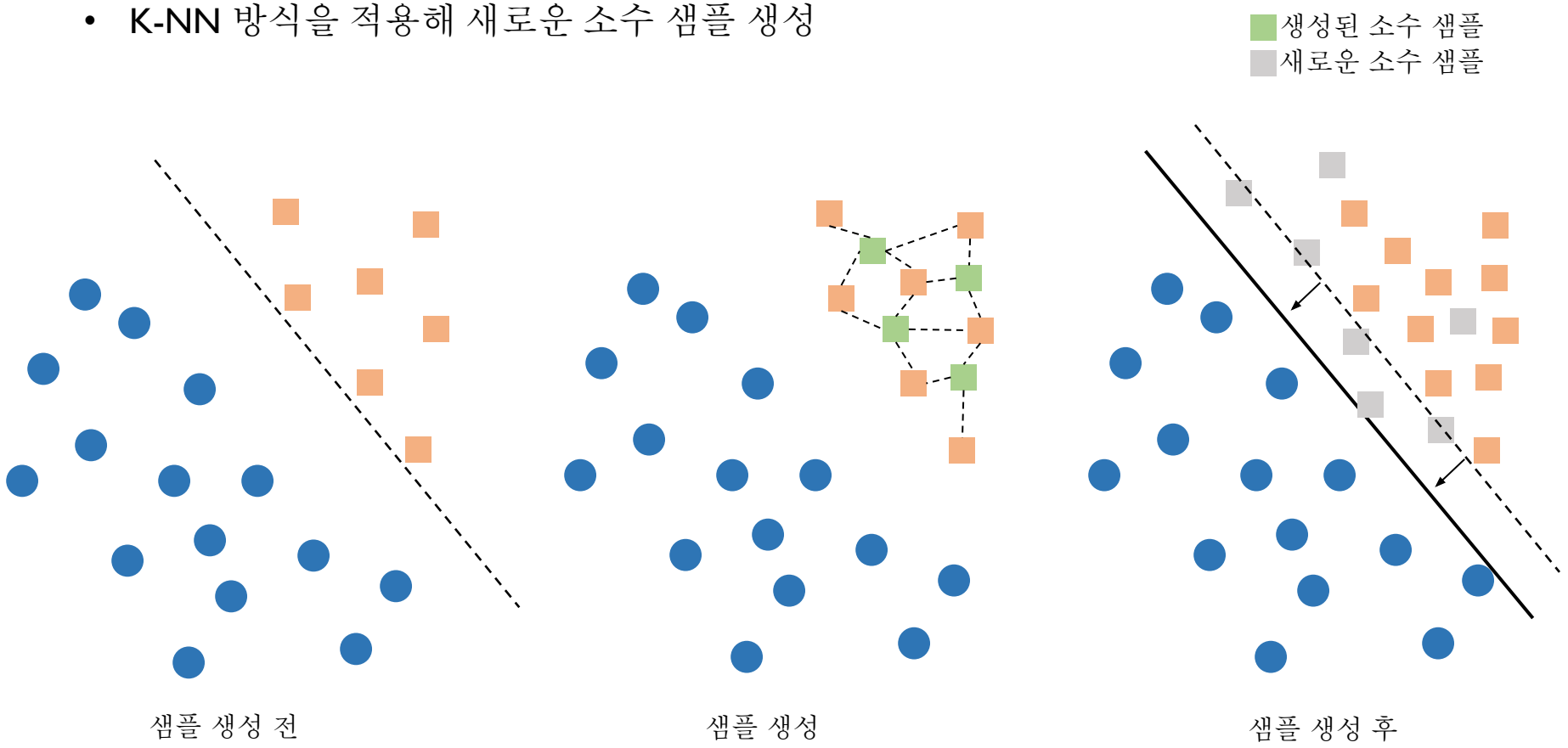
- Random over sampling : 다수 클래스 수와 같아지도록 소수 클래스 샘플을 무작위 복제
- 소수 클래스에 대해서 과적합이 발생할 수 있음



# Introduction

❖ **Data-Level 방법**

- SMOTE(Synthetic Minority Oversampling Technique)
- K-NN 방식을 적용해 새로운 소수 샘플 생성



# Cost-sensitive learning methods

❖ **Cost-sensitive learning이란?**

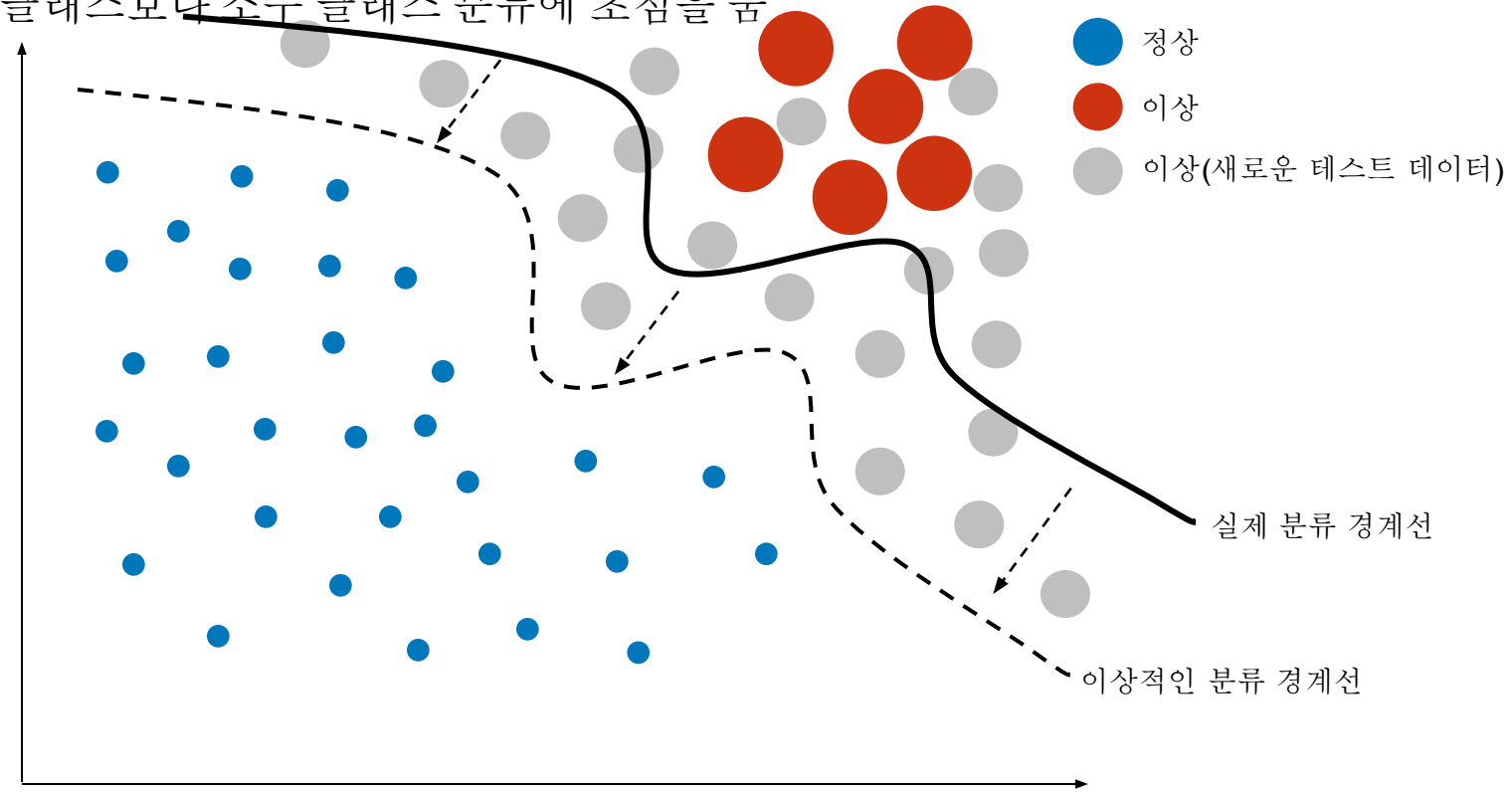
- 소수 클래스 관측치에 대한 cost값에 가중치를 더 많이 주어 균형 잡힌 학습으로 만드는 방법
- 다수 클래스보다 소수 클래스 분류에 초점을 둠


*https://datascience.foundation/sciencewhitepaper/understanding-imbalanced-datasets-and-techniques-for-handling-them*

# Cost-sensitive learning methods

❖ **Cost-sensitive learning이란?**

- 소수 클래스 관측치에 대한 cost값에 가중치를 더 많이 주어 균형 잡힌 학습으로 만드는 방법
- 다수 클래스보다 소수 클래스 분류에 초점을 둠

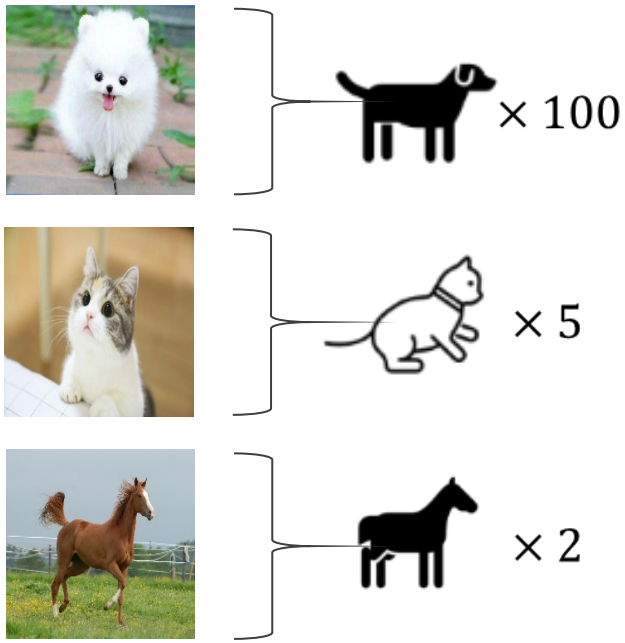


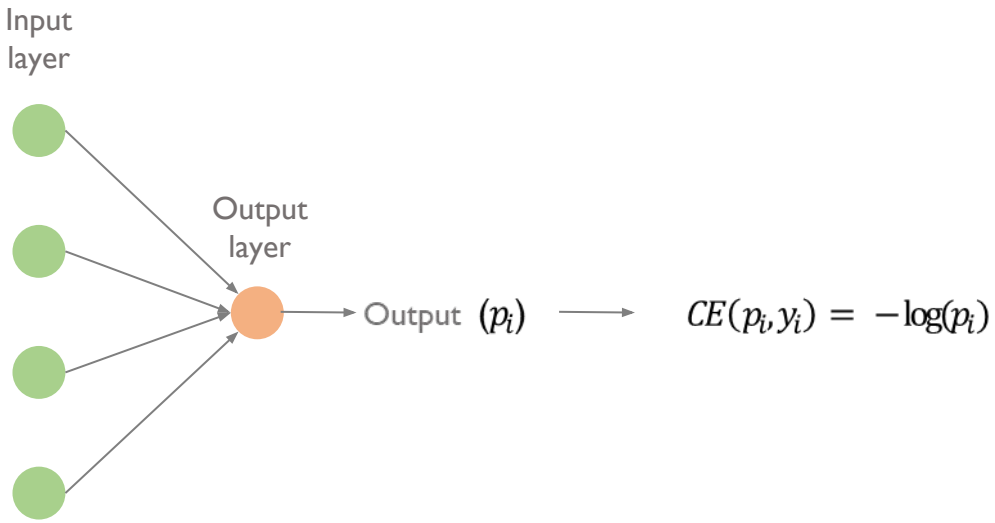
# Cost-sensitive learning methods

❖ **Cross Entropy Loss**

- Cross Entropy Loss : 잘 분류한 경우 보다 잘못 예측한 경우의 페널티에 초점을 두는 loss

$$CE(p_i, y_i) = -y_i \cdot log(p_i)$$
$$= -log(p_i)$$

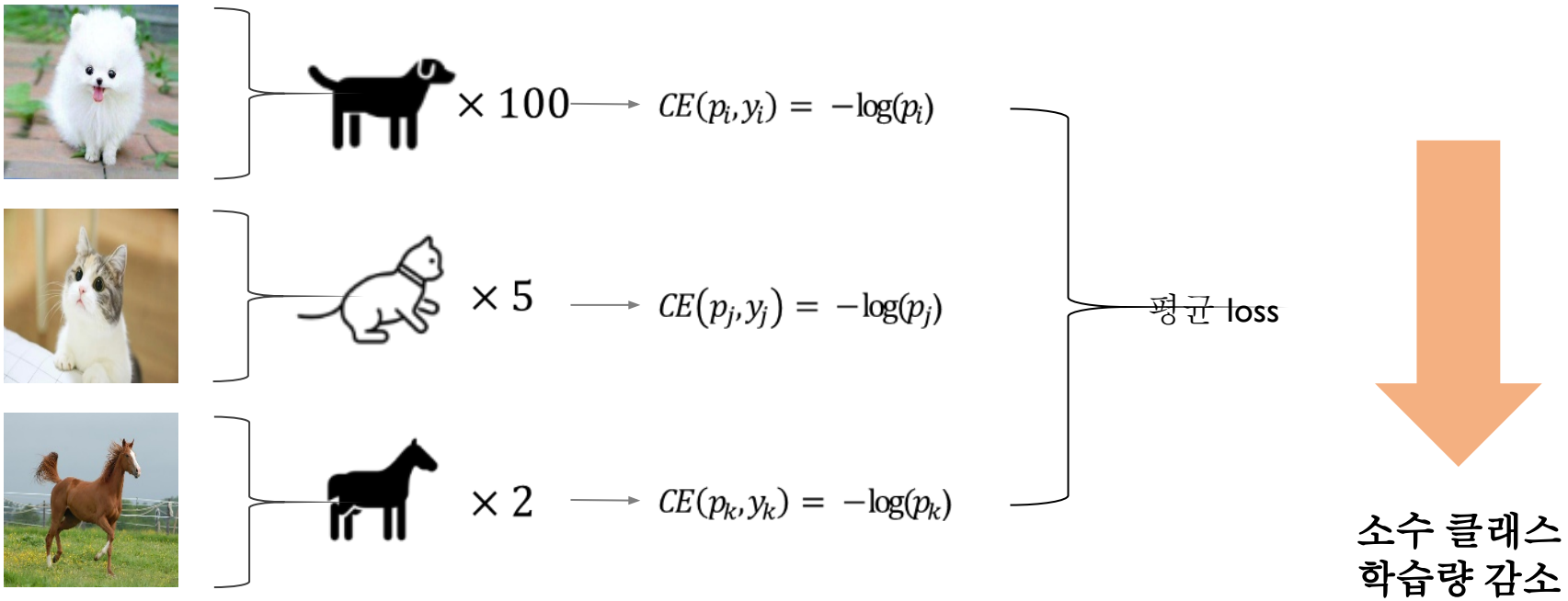


# Cost-sensitive learning methods

❖ **Cross Entropy Loss의 문제점**

- 다수 클래스와 소수 클래스 모두 같은 가중치를 가짐
- 다수 클래스 학습이 더 많이 될 것이고 소수 클래스에 대한 학습량이 줄어듬



# Cost-sensitive learning methods

❖ **Weighted Cross Entropy Loss**

- Cross Entropy Loss에서 클래스마다 비율이 다른 점을 개선
- 각 클래스 별 Loss 비율을 조절하는 가중치를 추가
- Inverse frequency weight : $\alpha_i = \frac{1}{n_i}$

$$Loss_{CE} = -log(p_i)$$

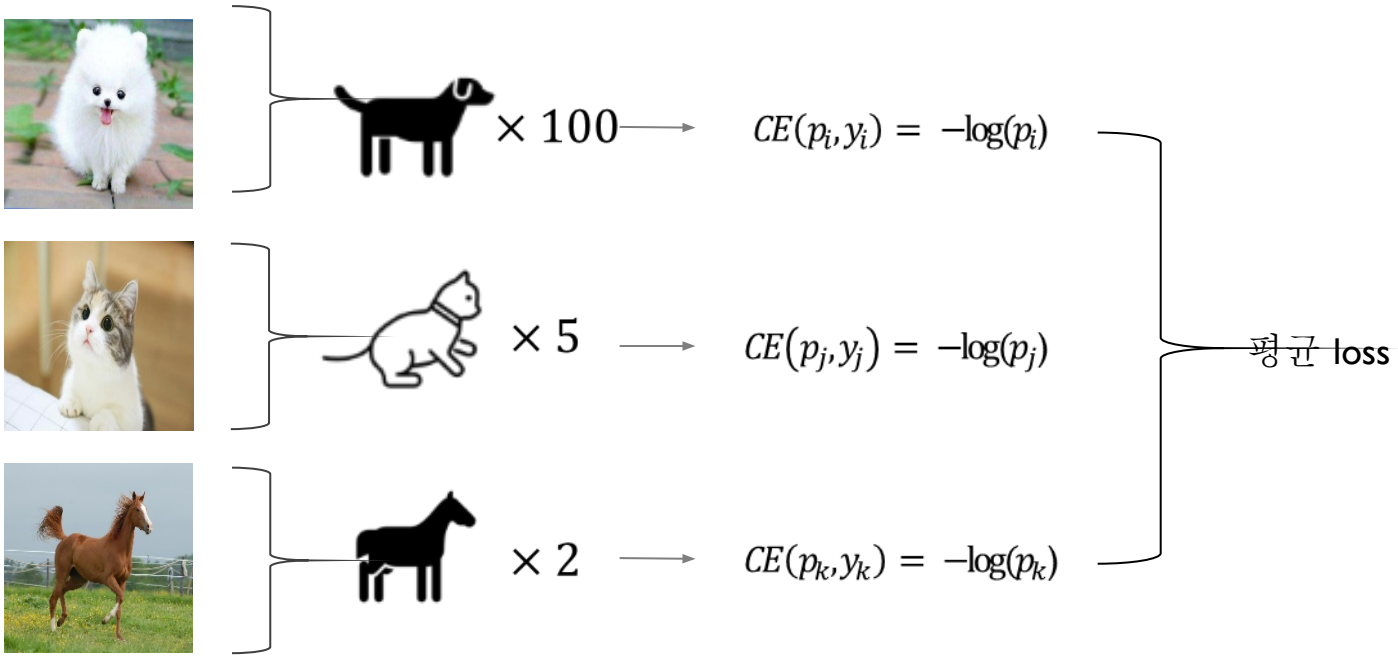
$$Loss_{WCE} = -\alpha_i \cdot log(p_i)$$

**클래스 별 샘플 수에 따른 균형**

# Cost-sensitive learning methods

❖ **Weighted Cross Entropy Loss**

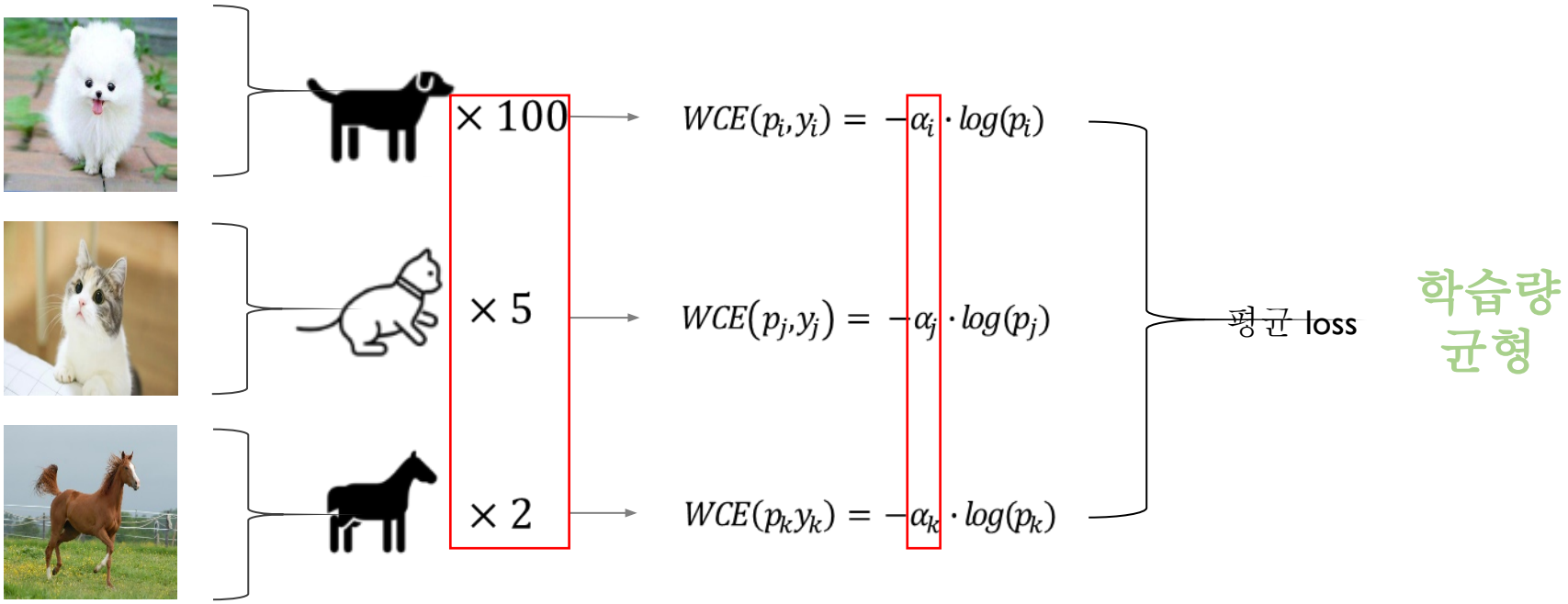
- Cross Entropy Loss에서 클래스마다 비율이 다른 점을 개선
- 각 클래스 별 Loss 비율을 조절하는 가중치를 추가
- Inverse frequency weight : $\alpha_i = \frac{1}{n_i}$



# Cost-sensitive learning methods

❖ **Weighted Cross Entropy Loss**

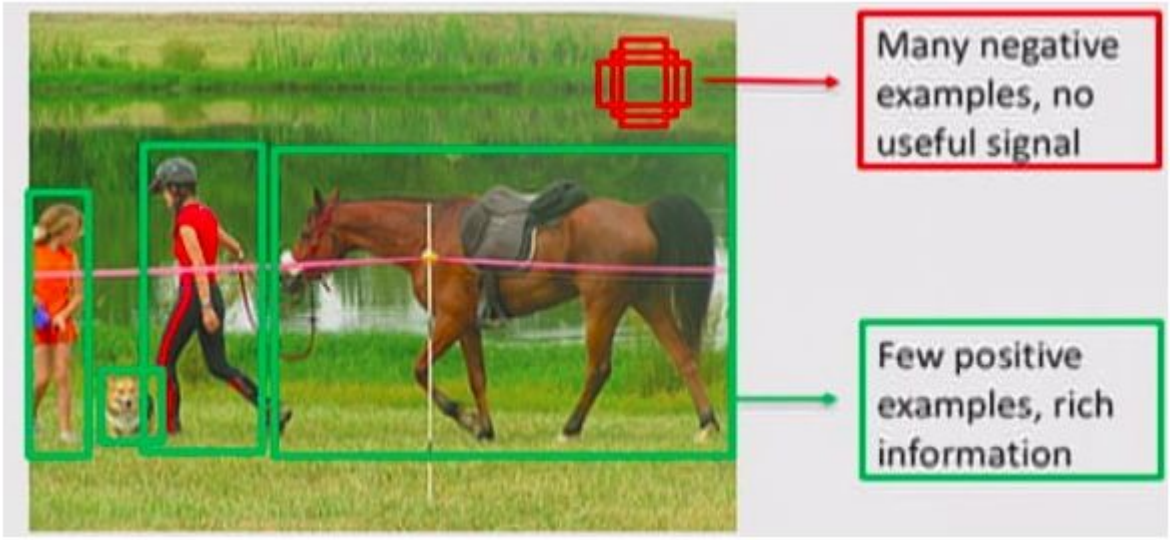
- Cross Entropy Loss에서 클래스마다 비율이 다른 점을 개선
- 각 클래스 별 Loss 비율을 조절하는 가중치를 추가
- Inverse frequency weight : $\alpha_i = \frac{1}{n_i}$



# Cost-sensitive learning methods

❖ **Focal Loss**

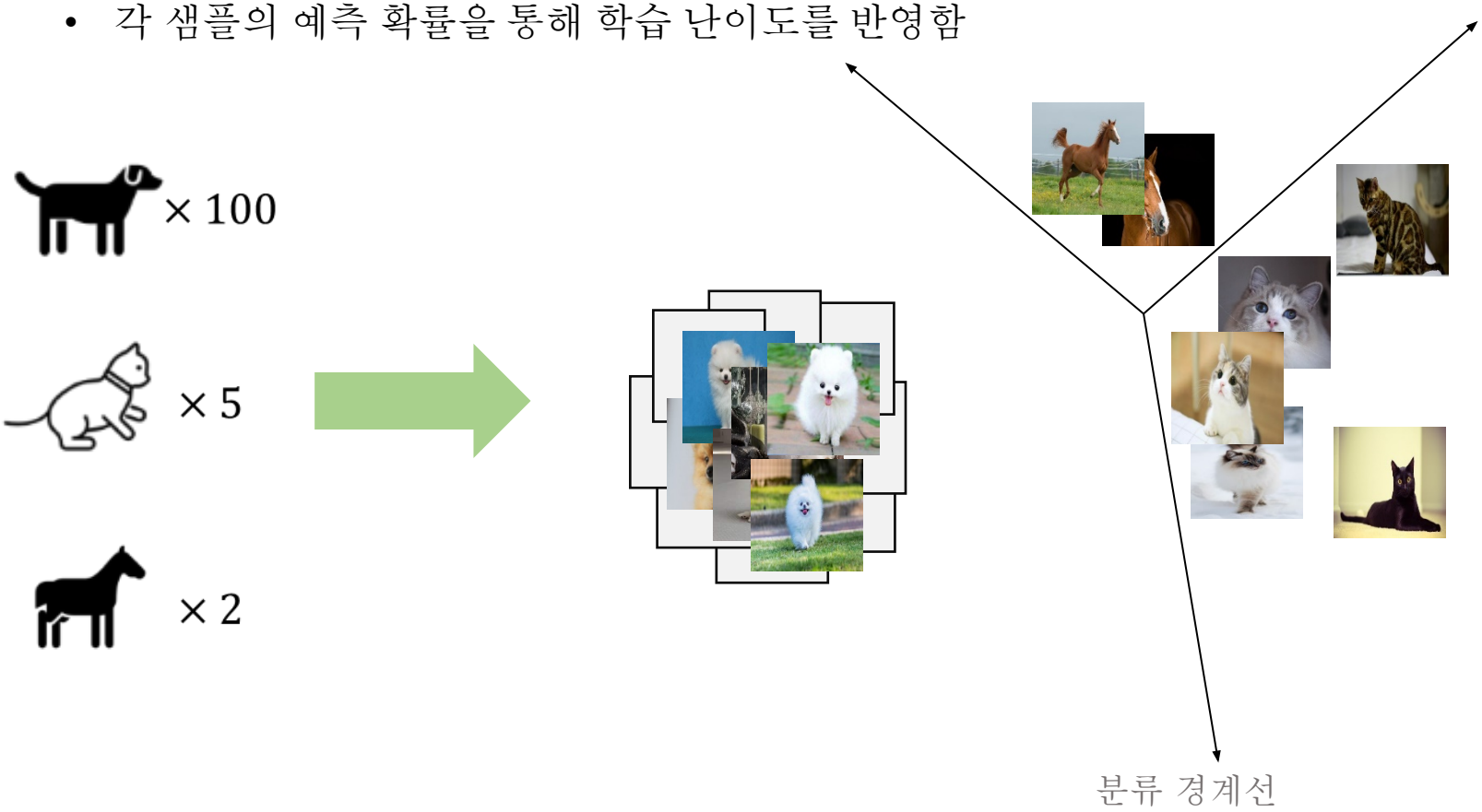
- Focal Loss는 Object Detection에서 사용된 개념
- 검출 대상에 비해 배경이 상대적으로 많아 클래스 불균형 문제 생김
- 분류하기 쉬운 샘플(배경)과 어려운 샘플(검출 대상)에 다른 가중치를 적용하여 해결




*Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection.*

# Cost-sensitive learning methods

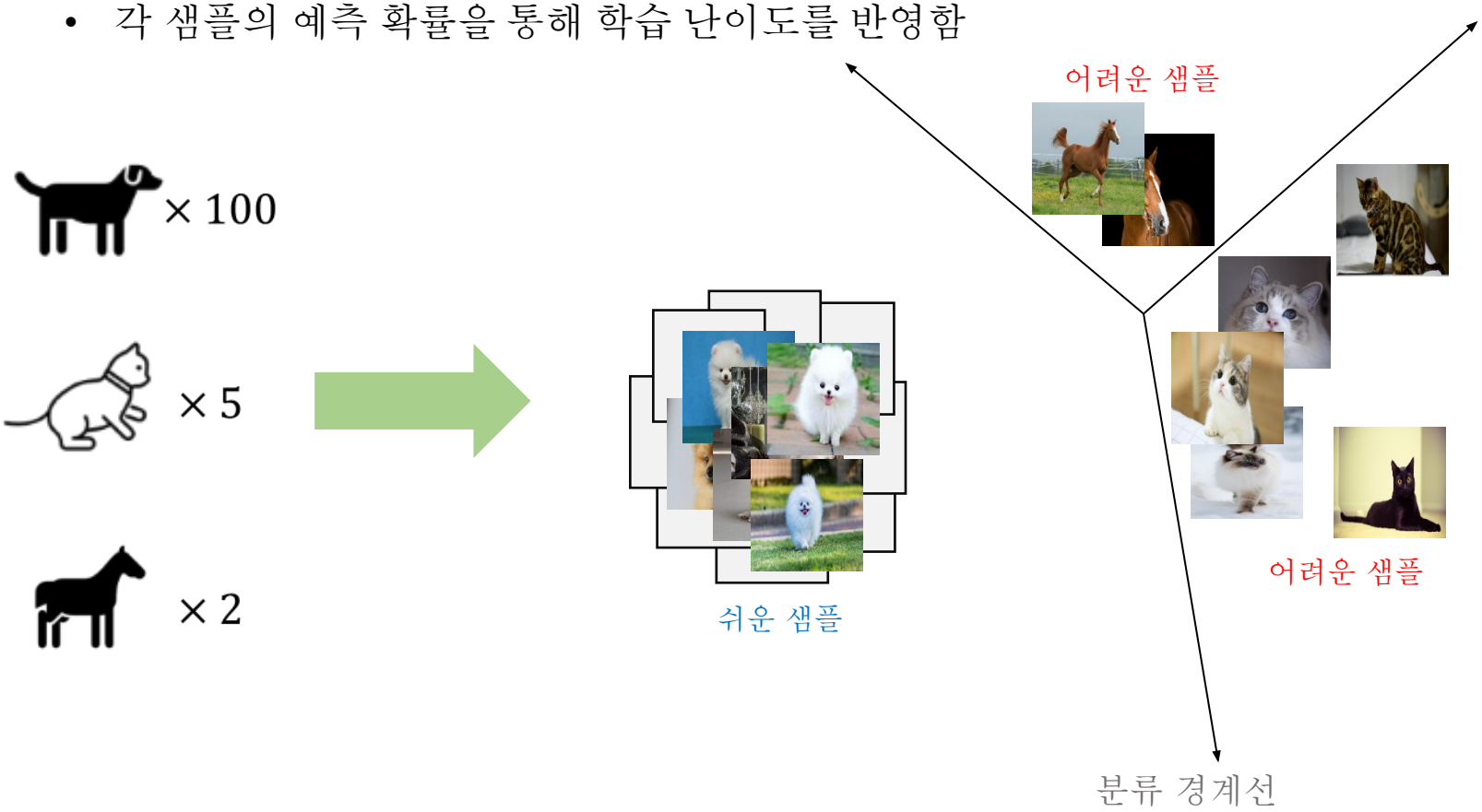
- **Focal Loss**
  - 학습 중 맞추기 쉬운 샘플에 대한 기여도를 줄이고 어려운 샘플의 기여도를 높이는 방식
  - 각 샘플의 예측 확률을 통해 학습 난이도를 반영함



# Cost-sensitive learning methods

❖ **Focal Loss**

- 학습 중 맞추기 쉬운 샘플에 대한 기여도를 줄이고 어려운 샘플의 기여도를 높이는 방식
- 각 샘플의 예측 확률을 통해 학습 난이도를 반영함



# Cost-sensitive learning methods

❖ **Focal Loss**

- 학습 중 맞추기 쉬운 샘플에 대한 기여도를 줄이고 어려운 샘플의 기여도를 높이는 방식
- 각 샘플의 예측 확률을 통해 학습 난이도를 반영함

$$Loss_{focal} = -\alpha_i \cdot (1 - p_i)^{\gamma} \cdot log(p_i)$$

- $\gamma$값이 커질수록 쉬운 샘플의 loss 기여도가 작아짐

쉬운 샘플

$p_i = 0.8 \longrightarrow Loss_{focal} = -\alpha_i \cdot (0.2)^{\gamma} \cdot \log(0.8)$

어려운 샘플

$p_j = 0.2 \longrightarrow Loss_{focal} = -\alpha_i \cdot (0.8)^{\gamma} \cdot \log(0.2)$

어려운 샘플

$p_k = 0.1 \longrightarrow Loss_{focal} = -\alpha_i \cdot (0.9)^{\gamma} \cdot \log(0.1)$

# Cost-sensitive learning methods

❖ **Focal Loss**

- 학습 중 맞추기 쉬운 샘플에 대한 기여도를 줄이고 어려운 샘플의 기여도를 높이는 방식
- 각 샘플의 예측 확률을 통해 학습 난이도를 반영함

$$Loss_{focal} = -\alpha_i \cdot (1 - p_i)^{\gamma} \cdot log(p_i)$$

- $\gamma$값이 커질수록 쉬운 샘플의 loss 기여도가 작아짐

쉬운 샘플

$p_i = 0.8 \longrightarrow Loss_{focal} = -\alpha_i \cdot (0.2)^{\gamma} \cdot \log(0.8)$

어려운 샘플

$p_j = 0.2 \longrightarrow Loss_{focal} = -\alpha_i \cdot (0.8)^{\gamma} \cdot \log(0.2)$

어려운 샘플

$p_k = 0.1 \longrightarrow Loss_{focal} = -\alpha_i \cdot (0.9)^{\gamma} \cdot \log(0.1)$

맞추기 어려운 샘플일수록
**Loss값이 커짐**

$\alpha_i$ $(1 - p_i)^{\gamma}$

**클래스 별 샘플 수에 따른 균형 + 샘플 별 예측 확률에 따른 균형**

# Cost-sensitive learning methods

## Class-Balanced Loss Based on Effective Number of Samples

❖ **Class-Balanced Loss Based on Effective Number of Samples (CVPR 2019)**

- Cornell 대학과 Google Brain에서 연구했으며 2022년 5월 26일 기준 862회 인용
- 데이터의 불균형 문제를 해결하기 위해 새로운 Loss Design 제안



**Class-Balanced Loss Based on Effective Number of Samples**

Yin Cui[1,2*] Menglin Jia[1] Tsung-Yi Lin[3] Yang Song[4] Serge Belongie[1,2]
[1]Cornell University [2]Cornell Tech [3]Google Brain [4]Alphabet Inc.

**Abstract**

*With the rapid increase of large-scale, real-world datasets, it becomes critical to address the problem of long-tailed data distribution (i.e., a few classes account for most of the data, while most classes are under-represented). Existing solutions typically adopt class re-balancing strategies such as re-sampling and re-weighting based on the number of observations for each class. In this work, we argue that as the number of samples increases, the additional benefit of a newly added data point will diminish. We introduce a novel theoretical framework to measure data overlap by associating with each sample a small neighboring region rather than a single point. The effective number of samples is defined as the volume of samples and can be calculated by a simple formula $(1-\beta^n)/(1-\beta)$, where $n$ is the number of samples and $\beta \in [0, 1)$ is a hyperparameter. We design a re-weighting scheme that uses the effective number of samples for each class to re-balance the loss, thereby yielding a* class-balanced loss. *Comprehensive experiments are conducted on artificially induced long-tailed CIFAR datasets and large-scale datasets including ImageNet and iNaturalist. Our results show that when trained with the proposed class-balanced loss, the network is able to achieve significant performance gains on long-tailed datasets.*

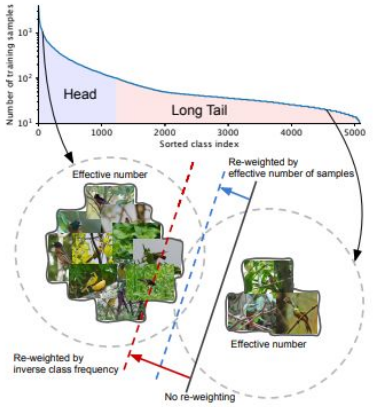



Figure 1. Two classes, one from the head and one from the tail of a long-tailed dataset (iNaturalist 2017 [41] in this example), have drastically different number of samples. Models trained on these samples are biased toward dominant classes (black solid line). Re-weighing the loss by inverse class frequency usually yields poor

# Cost-sensitive learning methods

Class-Balanced Loss Based on Effective Number of Samples

❖ **Effective Number of Samples**

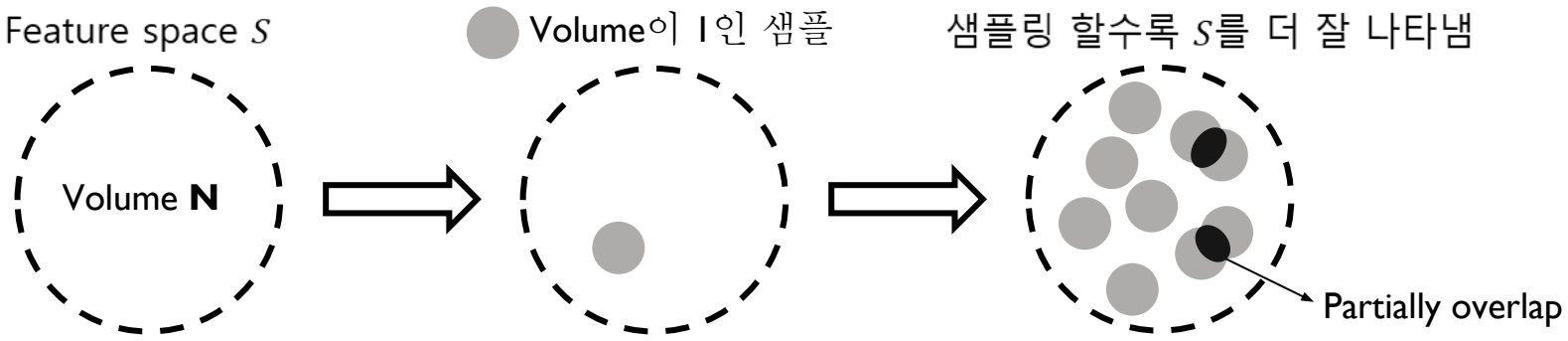
- 같은 클래스 데이터 사이의 정보 중복(Information overlap)을 고려한 값
- 가중치로 단순히 샘플 수를 이용하는 대신 Effective number of samples 개념 활용
- 다수 클래스보다 소수 클래스에 데이터가 추가될 때 새롭게 추가되는 정보가 많을 것임

# Cost-sensitive learning methods

Class-Balanced Loss Based on Effective Number of Samples

❖ **Effective Number of Samples**

- Effective number를 구하기 위해 Random covering 문제를 정의
- 샘플들은 랜덤 복원추출되고, 전체 set $S$를 Cover하는 것이 목표
- 샘플이 추출될 때마다 Overlap확률이 증가하므로 Volume의 기대값 감소
- Effective number는 샘플들의 Expected volume으로 정의



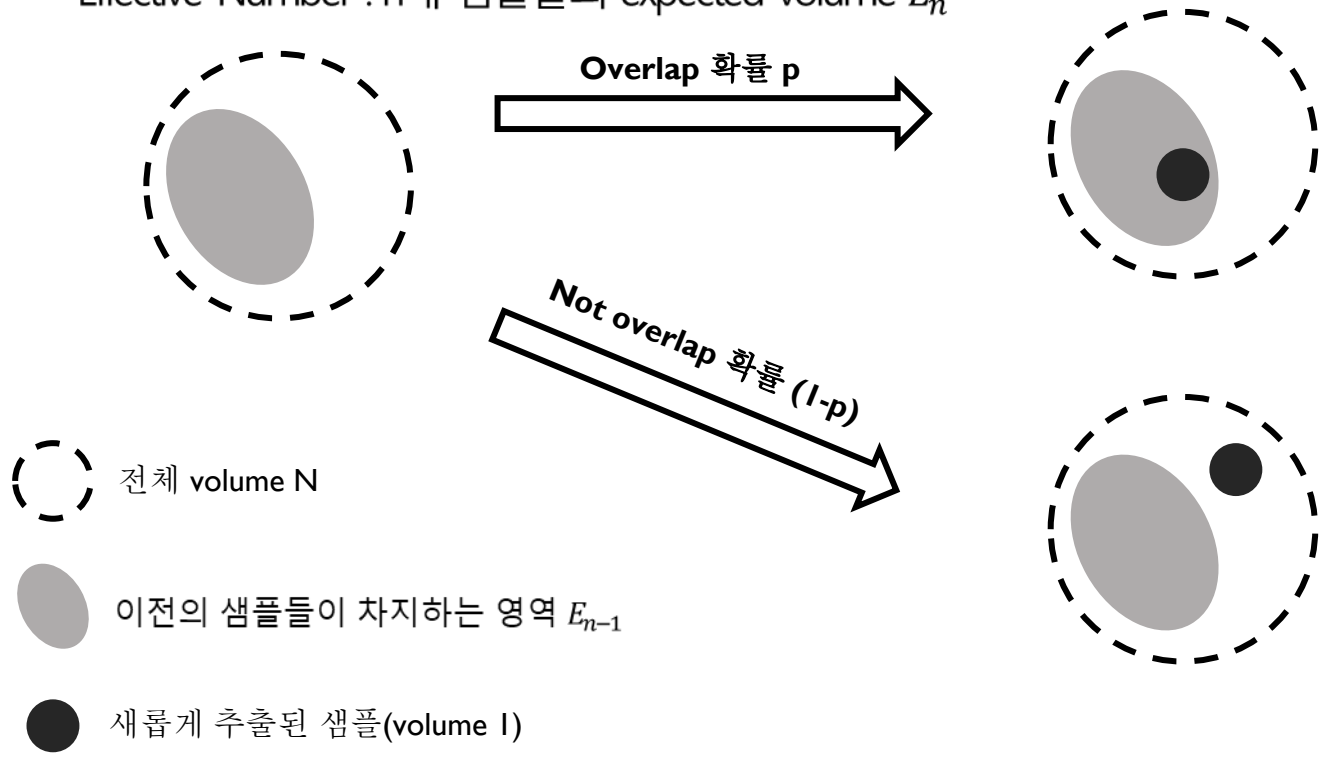

**Partially overlap을 고려하면 계산량이 많음**

# Cost-sensitive learning methods

Class-Balanced Loss Based on Effective Number of Samples

❖ **Effective Number of Samples**

- Partially overlap을 고려하지 않음
- 서로 다른 두 개의 샘플은 완전히 overlap하거나 Not overlap
- Effective Number : n개 샘플들의 expected volume $E_n$

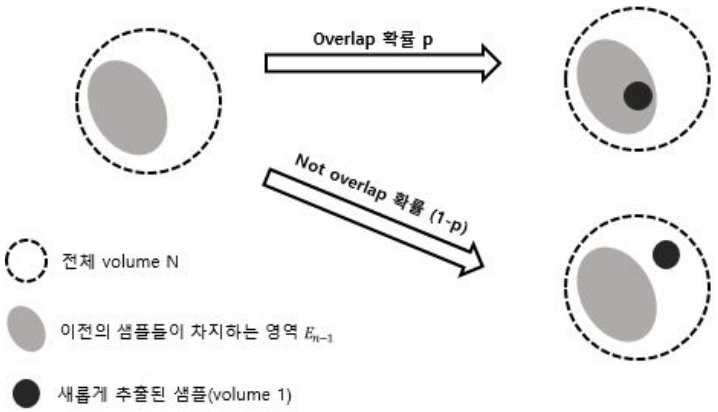


# Cost-sensitive learning methods

Class-Balanced Loss Based on Effective Number of Samples

❖ **Effective Number of Samples**

- Effective number $E_n = \frac{1-\beta^n}{1-\beta}$
- $\beta = \frac{N-1}{N}$ 이라 정의한 하이퍼파라미터
  - ✓ $\beta \in [0,1)$




$$E_1 = \frac{1-\beta^1}{1-\beta} = 1$$

$$E_n = pE_{n-1} + (1-p)(E_{n-1}+1)$$

$p = \frac{E_{n-1}}{N}$를 대입하면 $E_n = 1 + \frac{N-1}{N}E_{n-1}$

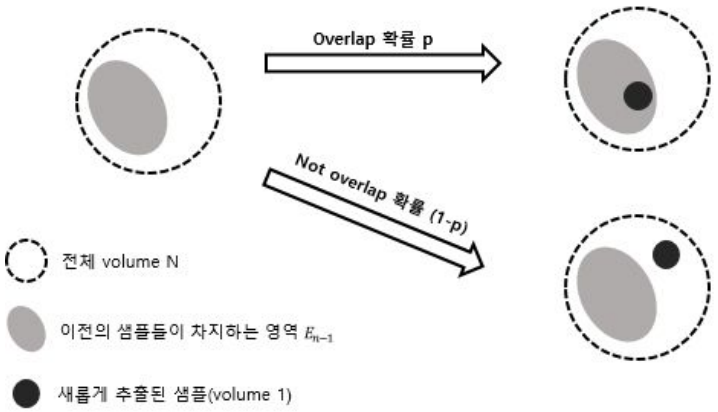
$E_{n-1} = \frac{1-\beta^{n-1}}{1-\beta}$이라고 가정하면 $E_n = \frac{1-\beta^n}{1-\beta}$

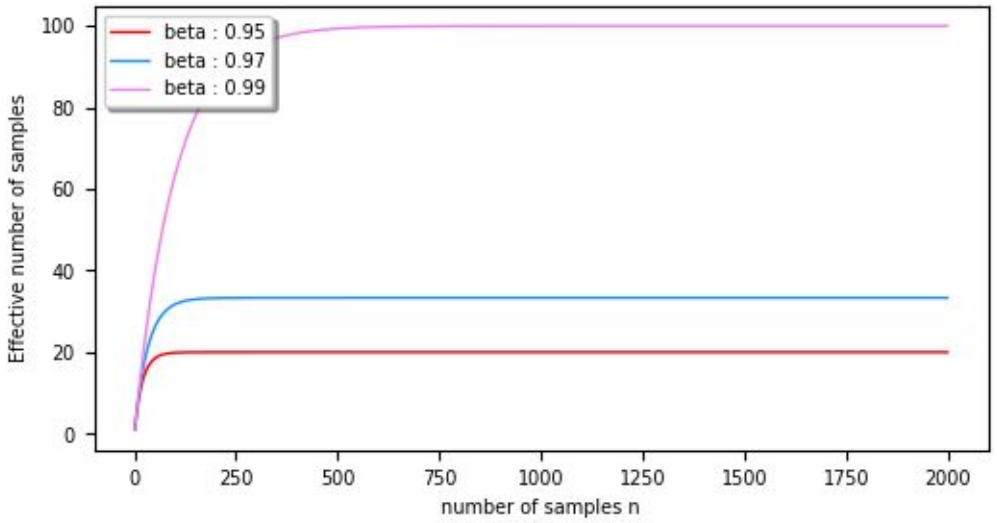
# Cost-sensitive learning methods

Class-Balanced Loss Based on Effective Number of Samples

❖ **Effective Number of Samples**

- Effective number $E_n = \frac{1-\beta^n}{1-\beta}$
- $\beta = \frac{N-1}{N}$ 이라 정의한 하이퍼파라미터
  - ✓ $\beta \in [0,1)$



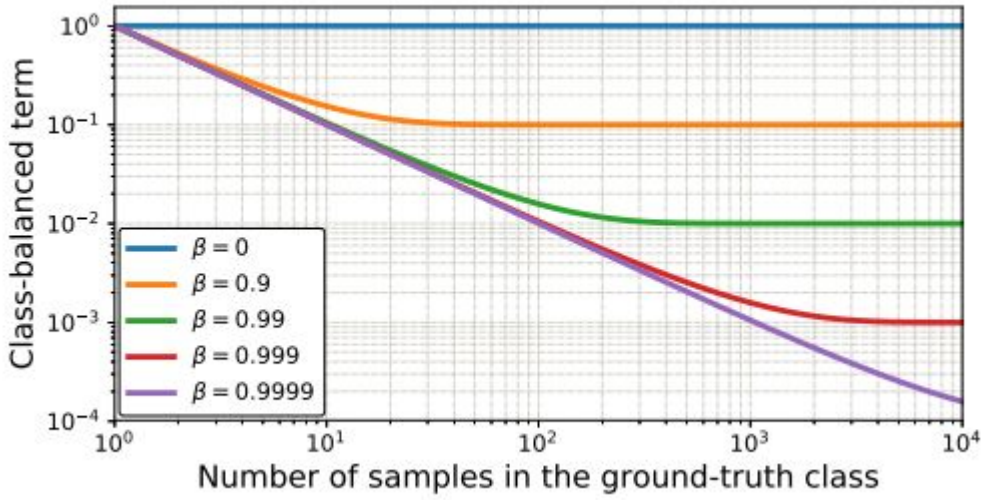

**Effective number**는 데이터 수뿐만 데이터 정보의 중복도 고려한 값

# Cost-sensitive learning methods

Class-Balanced Loss Based on Effective Number of Samples

❖ **Class-Balanced Loss**

- Class balanced term : $\frac{1}{E_{n_i}} = \frac{1-\beta}{1-\beta^{n_i}}$
  - ✓ $\beta = 0$ : no re-weighting
  - ✓ $\beta \to 1$ : re-weighting by inverse class frequency
- Inverse frequency weighted term : $\alpha_i = \frac{1}{n_i}$



# Cost-sensitive learning methods

Class-Balanced Loss Based on Effective Number of Samples

❖ **Class-Balanced Loss**

- Class balanced term : $\frac{1}{E_{n_i}} = \frac{1-\beta}{1-\beta^{n_i}}$
- Inverse frequency weighted term : $\alpha_i = \frac{1}{n_i}$

$$\text{WCE loss} = -\alpha_i \cdot log(p_i)$$

$$\text{Focal loss} = -\alpha_i \cdot (1-p_i)^{\gamma} \cdot log(p_i)$$

Class balanced term : $\frac{1-\beta}{1-\beta^{n_i}}$

$$CB_{CE}\ \text{loss} = -\frac{1-\beta}{1-\beta^{n_i}} \cdot log(p_i)$$

$$CB_{focal}\ \text{loss} = -\frac{1-\beta}{1-\beta^{n_i}} \cdot (1-p_i)^{\gamma} \cdot log(p_i)$$

# Cost-sensitive learning methods

Class-Balanced Loss Based on Effective Number of Samples

❖ **Class-Balanced Loss**

- Class balanced term : $\frac{1}{E_{n_i}} = \frac{1-\beta}{1-\beta^{n_i}}$
- $CB_{focal}$ loss $= -\frac{1-\beta}{1-\beta^{n_i}} \cdot (1-p_i)^{\gamma} \cdot log(p_i)$


× 100

$p_i = 0.8 \longrightarrow CB_{focal} = -\frac{1-\beta}{1-\beta^{100}} \cdot (0.2)^{\gamma} \cdot \log(0.8)$


× 5

$p_j = 0.2 \longrightarrow CB_{focal} = -\frac{1-\beta}{1-\beta^{5}} \cdot (0.8)^{\gamma} \cdot \log(0.2)$


× 2

$p_k = 0.1 \longrightarrow CB_{focal} = -\frac{1-\beta}{1-\beta^{2}} \cdot (0.9)^{\gamma} \cdot \log(0.1)$

각 샘플 마다 **Effective number**를 통해서 정보의 중복까지 고려함

# Conclusion

## ❖ Summary

- 클래스 불균형 문제 해결 방안은 크게 Data-level과 Algorithm-level로 나뉨
- Algorithm-level 중에서 Cost-sensitive learning을 다룸
  - ✔ Cross entropy loss : 불균형 상황에서 cross entropy loss만 사용하게 된다면 다수 클래스에 편향된 결과 가 나옴
  - ✔ Weighted Cross entropy loss : 불균형 상황에서 Cross entropy에 클래스 별 샘플 수를 고려함
  - ✔ Focal loss : 불균형 상황에서 다수 클래스의 맞추기 쉬운 샘플들과 소수 클래스의 맞추기 어려운 샘플 들에 대해 학습 난이도를 반영해주는 term을 추가하여 개선
  - ✔ Class balanced loss : 같은 클래스 내에서 샘플들 사이 정보의 중복을 손실 함수에 반영하기 위해서 Effective number of samples이라는 새로운 loss design term을 제안

# Thank you

**E-mail : jy_jeong@korea.ac.kr**

# 참고 문헌


- ✔ 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.

- ✔ Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980-2988).

- ✔ Cui, Y., Jia, M., Lin, T. Y., Song, Y., & Belongie, S. (2019). Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9268-9277).