Deep Neural Networks for Rank-Consistent Ordinal Regression Based On Conditional Probabilities

Xintong Shi, Wenzhi Cao, Sebastian Raschka

2022. 11. 22

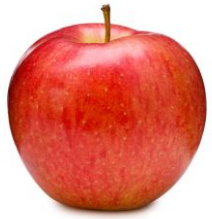
목차

1. Introduction

Classification (분류 문제)

output: 특정 클래스

- 스팸메일 분류 (spam, ham)
- 이미지 분류 (과일 이미지 분류 등)



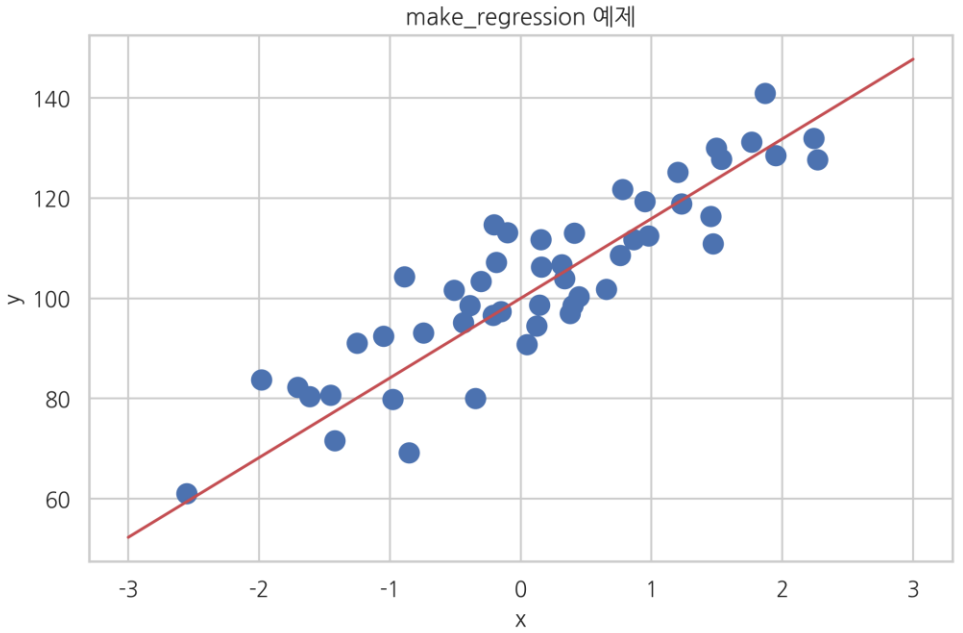
DL 모델

사과

Regression (회귀 문제)

output: 연속적인 값

- 집 값 예측
- 시험 점수 예측



1. Introduction

Classification (분류 문제)

+

Regression (회귀 문제)

Ordinal Regression (순서가 있는 분류 문제)
= Ordinal classification = Ranking learning

output: 순서형 클래스(Ranking)

- 질병 진행 정도 예측
- 얼굴 이미지로 나이 예측
- 리뷰 별점 예측

: 클래스 간의 **순서관계**가 있지만,
Regression처럼 **연속적인 값이 아니고**,
클래스 간의 거리를 측정할 수 없음.

=> Regression과 Classification 성질을
모두 가지고 있음.

1. Introduction

Ordinal Regression

① Label Extension

$Y = \{1,2,3,4,5\}$

$y^{[i]} \in Y$: Y의 i번째 데이터

K는 class 수

$y^{[i]}$ 를 K-1개의 vector 로 확장

$y^{[i]}$ = 1 -> {0,0,0,0}

$y^{[i]}$ = 5 -> {1,1,1,1}

$y^{[i]}$= 3 -> {1,1,0,0}

② Predicted Ordinal Label

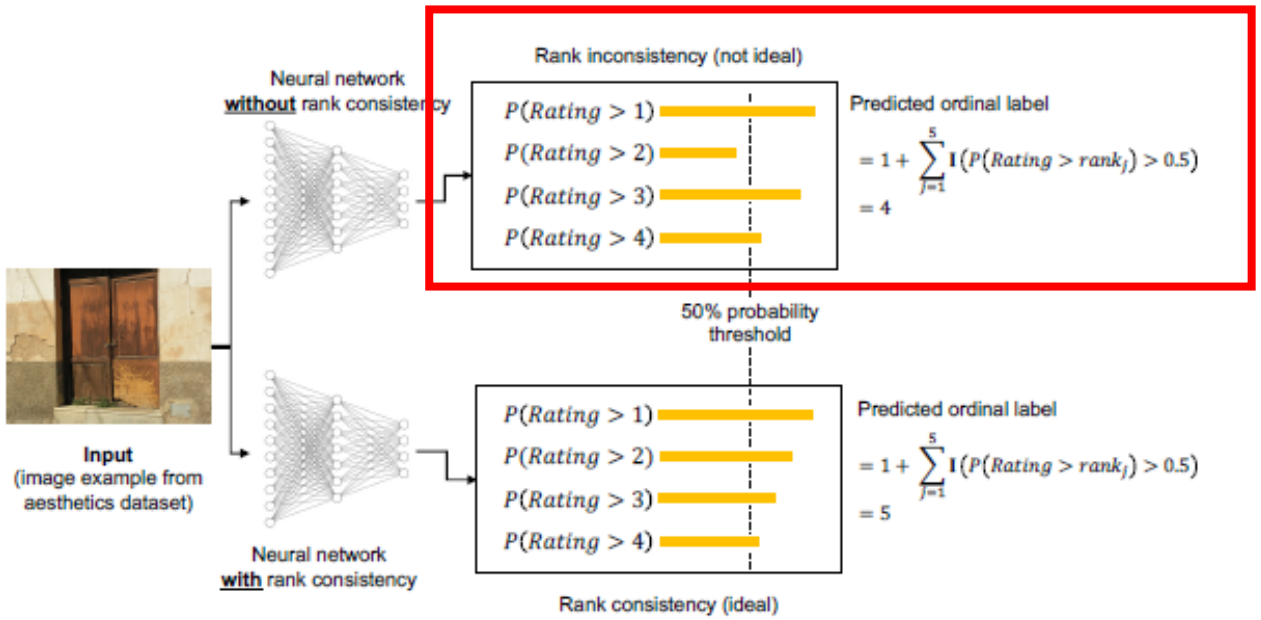


Figure 1: Illustration of the difference between rank-consistent and rank-inconsistent methods.

K-1 개의 binary classifier 에서 0 or 1 예측하고,
이를 축적하여 Ordinal label 예측.

1. Introduction

Ordinal Classification

① Label Extension

$Y = \{1,2,3,4,5\}$

$y_i \in Y$: Y의 i번째 데이터

y_i = 0 -> {0,0,0,0}

y_i = 5 -> {1,1,1,1}

y_i = 3 -> {1,1,1,0}

순서형 클래스 간의
순위 일관성을 지키고자
CORN 제안
Conditional Ordinal Regression for Neural Network

② Predicted Ordinal Label

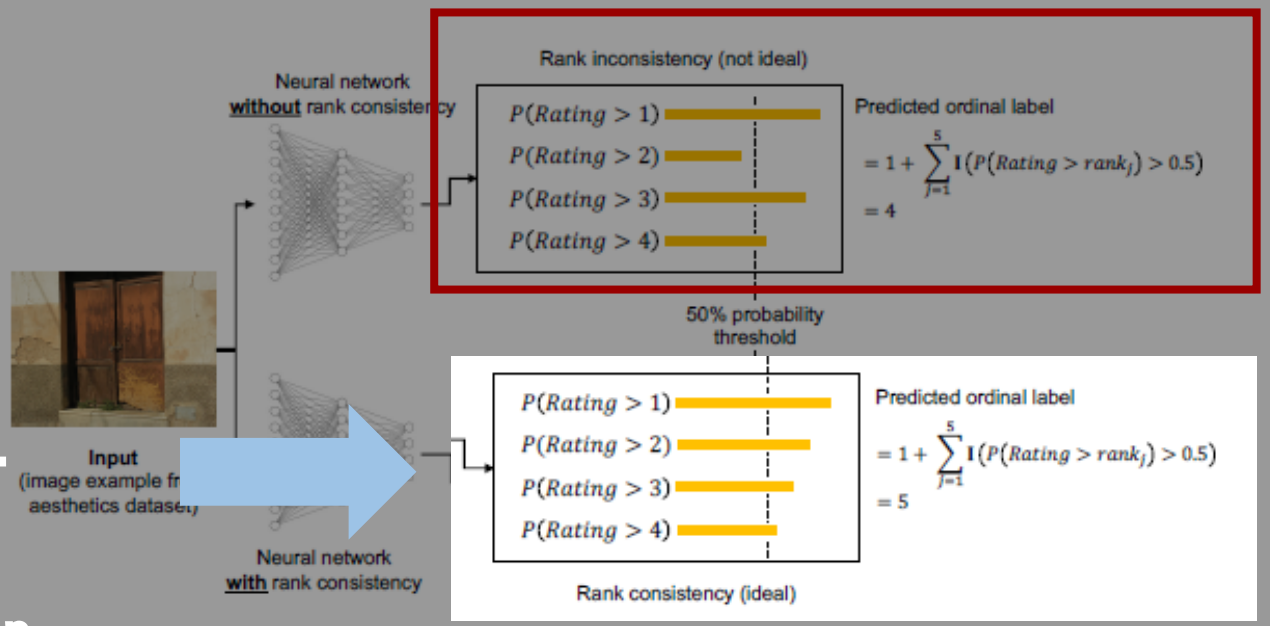


Figure 1: Illustration of the difference between rank-consistent and rank-inconsistent methods.

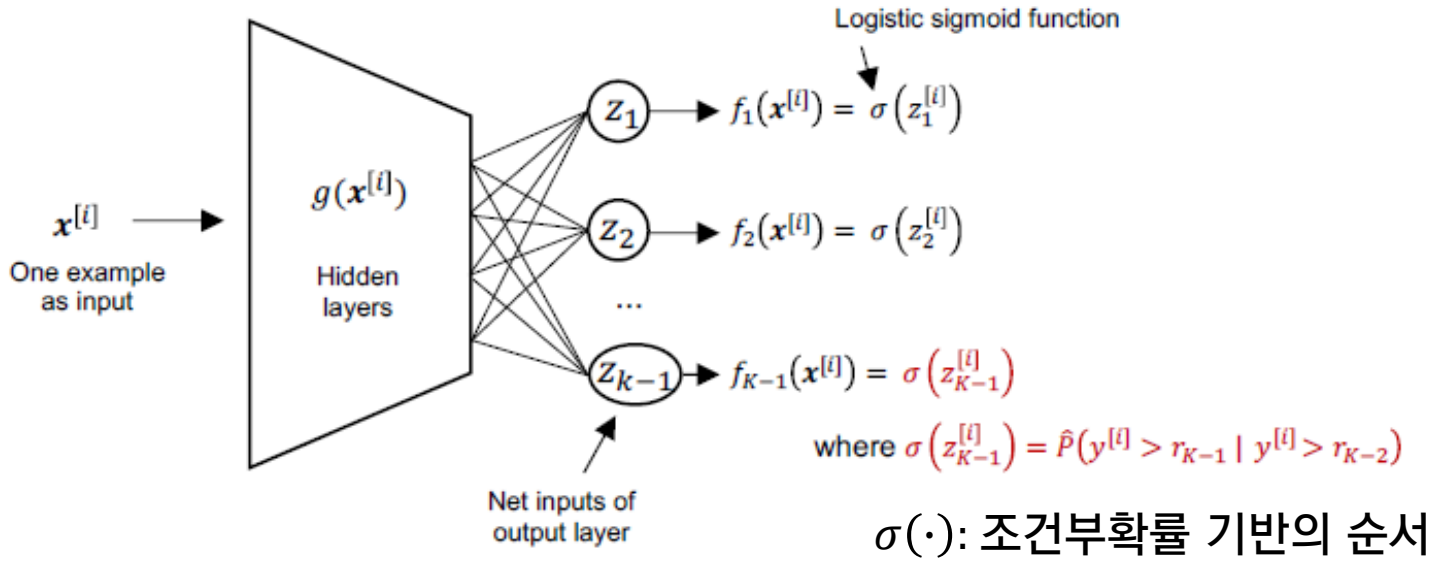
여러 개의 binary classifier 에서 0 or 1 예측하고,
이를 축적하여 Ordinal label 예측.

2. Preliminaries

Outline of the neural network architecture used for CORN

$y^{[i]} = 3 \rightarrow \{1,1,0,0\}$

$f_k(\cdot)$: k번째 벡터에 해당하는 확률을 구하는 함수



$\sigma(\cdot)$: 조건부확률 기반의 순서형 회귀 예측 함수

Figure 2: Outline of the neural network architecture used for CORN.

2. Preliminaries

1. Dataset

$D = \{x^{[i]}, y^{[i]}\}_{i=1}^{N}$ *(N training examples)*

$y^{[i]}$: *class label*

$y^{[i]} \in Y = \{r_1, r_2, \ldots, r_k\}$ $(r_K \succ r_{K-1} \succ \ldots \succ r_1)$

2. *Label Extension*

$y_k^{[i]} \in \{0, 1\}$

$y^{[i]}$ = 1 -> {0,0,0,0}

$y^{[i]}$ = 3 -> {1,1,0,0}

3. *Objective of an ordinal regression model*

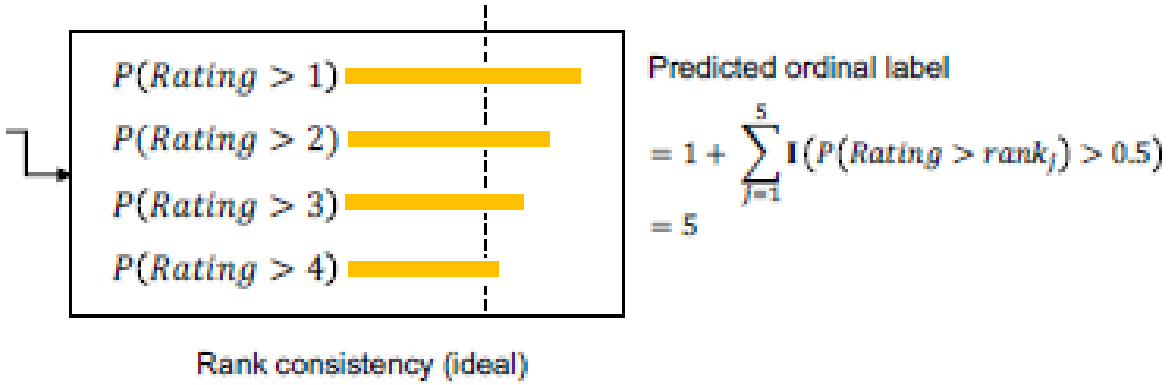
To find mapping $h : X \rightarrow Y$ *that minimizes a loss function* $L(h)$

3. CORN Framework

Label Prediction

$obtain\ via\ h(x^{[i]}) = r_q\ (q \in \{1, 2, \ldots, K\})$

$$q = 1 + \sum_{k=1}^{K-1} 1\{f_k(x^{[i]}) > 0.5\}$$



$f_k(x^{[i]}) \in [0, 1]$ (0에서 1사이의 값을 가지는 확률)

$1\{\cdot\}$: indicator function (Boolean function)

$\{\cdot\}$안의 조건이 참이면 1, 거짓이면 0

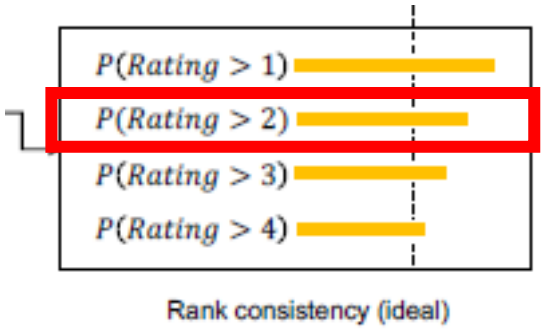
3. CORN Framework

Rank-consistent Ordinal Regression based on conditional Probabilities

확률 구하는 함수

$$f_k(x^{[i]}) = \hat{P}(y^{[i]} > r_k \mid y^{[i]} > r_{K-1}) = \frac{\hat{P}(y^{[i]} > r_k \cap y^{[i]} > r_{K-1})}{\hat{P}(y^{[i]} > r_{K-1})}$$

k = 2 일 때,



Rank consistency (ideal)

$$f_2(x^{[i]}) = \hat{P}(y^{[i]} > r_2 \mid y^{[i]} > r_1) = \frac{\hat{P}(y^{[i]} > r_2 \cap y^{[i]} > r_1)}{\hat{P}(y^{[i]} > r_1)}$$

3. CORN Framework

Rank-consistent Ordinal Regression based on conditional Probabilities

확률 구하는 함수

$$f_k(x^{[i]}) = \hat{P}(y^{[i]} > r_k \mid y^{[i]} > r_{K-1}) = \frac{\hat{P}(y^{[i]} > r_k \cap y^{[i]} > r_{K-1})}{\hat{P}(y^{[i]} > r_{K-1})}$$

when $k = 1, f_1(x^{[i]}) = \hat{P}(y^{[i]} > r_1)$.

where the events are nested

: $\{y^{[i]} > r_k\} \subseteq \{y^{[i]} > r_{k-1}\}$ (두 사건이 독립이 아니다.)

$$\hat{P}(y^{[i]} > r_k) = \prod_{j=1}^{k} f_j(x^{[i]})$$

since $\forall j, 0 \leq f_j(x^{[i]}) \leq 1, we\ have$

$$\hat{P}(y^{[i]} > r_1) \geq \hat{P}(y^{[i]} > r_2) \geq \cdots \geq \hat{P}(y^{[i]} > r_{K-1})$$

3. CORN Framework

Rank-consistent Ordinal Regression based on conditional Probabilities

$$\widehat{P}(y^{[i]} > r_1) \geq \widehat{P}(y^{[i]} > r_2) \geq \cdots \geq \widehat{P}(y^{[i]} > r_{K-1})$$

$$\hat{P}(y^{[i]} > r_1) = f_1(x^{[i]})$$

$$\hat{P}(y^{[i]} > r_2) = f_1(x^{[i]}) \cdot f_2(x^{[i]}) = \frac{\cancel{\hat{P}(y^{[i]}>r_1)}}{1} \cdot \frac{\hat{P}(y^{[i]}>r_2 \cap y^{[i]}>r_1)}{\cancel{\hat{P}(y^{[i]}>r_1)}} = \hat{P}(y^{[i]} > r_2 \cap y^{[i]} > r_1)$$

$$\hat{P}(y^{[i]} > r_3) = f_1(x^{[i]}) \cdot f_2(x^{[i]}) \cdot f_3(x^{[i]}) = \frac{\cancel{\hat{P}(y^{[i]}>r_1)}}{1} \cdot \frac{\hat{P}(y^{[i]}>r_2 \cap y^{[i]}>r_1)}{\cancel{\hat{P}(y^{[i]}>r_1)}} \cdot \frac{\hat{P}(y^{[i]}>r_3 \cap y^{[i]}>r_2)}{\cancel{\hat{P}(y^{[i]}>r_2)}} = \hat{P}(y^{[i]} > r_3 \cap y^{[i]} > r_2)$$

...

$$\hat{P}(y^{[i]} > r_k) = \prod_{j=1}^{k} f_j(x^{[i]}) = \hat{P}(y^{[i]} > r_k \cap y^{[i]} > r_{k-1})$$

...

$$\hat{P}(y^{[i]} > r_{K-1}) = \prod_{j=1}^{K-1} f_j(x^{[i]}) = \hat{P}(y^{[i]} > r_{K-1} \cap y^{[i]} > r_{k-2})$$

3. CORN Framework

Rank-consistent Ordinal Regression based on conditional Probabilities

$$\hat{P}(y^{[i]} > r_1) \geq \hat{P}(y^{[i]} > r_2) \geq \cdots \geq \hat{P}(y^{[i]} > r_{K-1})$$

$$\hat{P}(y^{[i]} > r_1) = f_1(x^{[i]})$$

$$\hat{P}(y^{[i]} > r_2) = f_1(x^{[i]}) \cdot f_2(x^{[i]}) = \frac{\cancel{\hat{P}(y^{[i]} > r_1)}}{1} \cdot \frac{\hat{P}(y^{[i]} > r_2 \cap y^{[i]} > r_1)}{\cancel{\hat{P}(y^{[i]} > r_1)}} = \hat{P}(y^{[i]} > r_2 \cap y^{[i]} > r_1)$$

$$\hat{P}(y^{[i]} > r_3) = f_1(x^{[i]}) \cdot f_2(x^{[i]}) \cdot f_3(x^{[i]}) = \frac{\cancel{\hat{P}(y^{[i]} > r_1)}}{1} \cdot \frac{\hat{P}(y^{[i]} > r_2 \cap y^{[i]} > r_1)}{\cancel{\hat{P}(y^{[i]} > r_1)}} \cdot \frac{\hat{P}(y^{[i]} > r_3 \cap y^{[i]} > r_2)}{\cancel{\hat{P}(y^{[i]} > r_2)}} = \hat{P}(y^{[i]} > r_3 \cap y^{[i]} > r_2)$$

...

$$\hat{P}(y^{[i]} > r_k) = \prod_{j=1}^{k} f_j(x^{[i]}) = \hat{P}(y^{[i]} > r_k \cap y^{[i]} > r_{k-1})$$

...

$$\hat{P}(y^{[i]} > r_{K-1}) = \prod_{j=1}^{K-1} f_j(x^{[i]}) = \hat{P}(y^{[i]} > r_{K-1} \cap y^{[i]} > r_{k-2})$$

Next *conditional Subset*

3. CORN Framework

Rank-consistent Ordinal Regression based on conditional Probabilities - conditional subsets

$f_1(x^{[i]}), f_2(x^{[i]}), \ldots, f_{K-1}(x^{[i]})$

$f_1(x^{[i]}) = \hat{P}(y^{[i]} > r_1)$ => *classic binary classification*

$f_2(x^{[i]}) = \hat{P}(y^{[i]} > r_2 \mid y^{[i]} > r_1)$

we focus only on the subset of the training data where $y^{[i]} > r_1$

...

$f_k(x^{[i]}) = \hat{P}(y^{[i]} > r_k \mid y^{[i]} > r_{k-1})$

we focus only on the subset of the training data where $y^{[i]} > r_{k-1}$

we minimize the binary cross-entropy loss on these conditional subsets.

3. CORN Framework

Rank-consistent Ordinal Regression based on conditional Probabilities

$f_1(x^{[i]}), f_2(x^{[i]}), \ldots, f_{K-1}(x^{[i]})$

$f_1(x^{[i]}) = \hat{P}(y^{[i]} > r_1)$ => *classic binary classification*

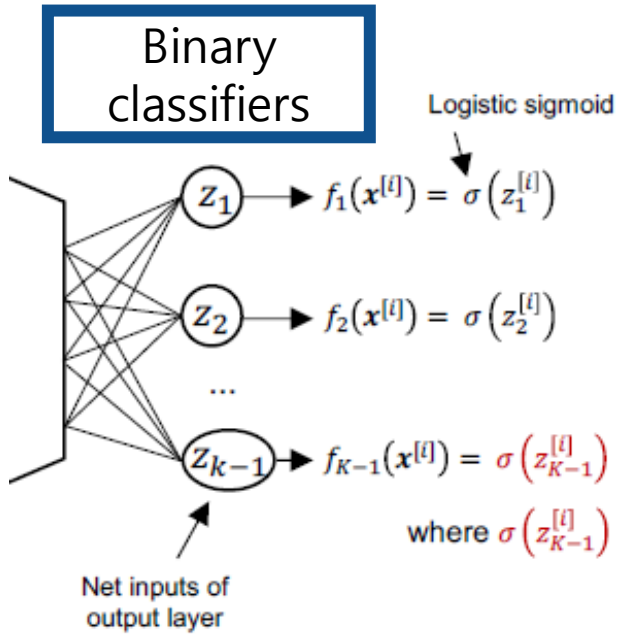
$f_2(x^{[i]}) = \hat{P}(y^{[i]} > r_2 \mid y^{[i]} > r_1)$

we focus only on the subset of the training data where $y^{[i]} > r_1$

...

$f_k(x^{[i]}) = \hat{P}(y^{[i]} > r_{K-2} \mid y^{[i]} > r_{K-1})$

we focus only on the subset of the training data where $y^{[i]} > r_k$



instead of minimizing the K – 1 loss functions corresponding to the K–1 conditional probabilities on each conditional subset separately, we minimize the binary cross–entropy loss on these conditional subsets.

3. CORN Framework

with out conditional subsets

$$f_k(x^{[i]}) = \hat{P}(y^{[i]} > r_k \mid y^{[i]} > r_{k-1})$$

$$=> f_k(x^{[i]}) = \hat{P}(y^{[i]} > r_k)\ (순위\ 일관성은\ 유지됨.)$$

since $\forall j, 0 \leq f_j(x^{[i]}) \leq 1, we\ have$

$$\hat{P}(y^{[i]} > r_1) \geq \hat{P}(y^{[i]} > r_2) \geq \cdots \geq \hat{P}(y^{[i]} > r_{K-1})$$

Table 4: MAE prediction errors on the test sets for the ResNet34 backbone. The class labels in all datasets were balanced. Best results are highlighted in bold.

	CORN	CORN w/o subsets
MORPH-2	2.98 ± 0.02	**2.93 ± 0.04**
AFAD	**2.81± 0.02**	3.06 ± 0.02
AES	**0.43 ± 0.01**	0.68 ± 0.01
Fireman	**0.76 ± 0.01**	0.81 ± 0.01

3. CORN Framework

Loss Function

$$\boldsymbol{L}(\boldsymbol{X}, \boldsymbol{y}) = -\frac{1}{\sum_{j=1}^{K-1} |S_j|} \sum_{j=1}^{K-1} \sum_{i=1}^{|S_j|} \Big[\log\left(f_j(\mathbf{x}^{[i]})\right) \cdot \mathbb{1}\left\{y^{[i]} > r_j\right\} + \log\left(1 - f_j\left(\mathbf{x}^{[i]}\right)\right) \cdot \mathbb{1}\left\{y^{[i]} \leq r_j\right\} \Big],$$

3. CORN Framework

Loss Function

$$L(X, y) = -\frac{1}{\sum_{j=1}^{K-1} |S_j|} \sum_{j=1}^{K-1} \sum_{i=1}^{|S_j|} \Big[\log\left(f_j(\mathbf{x}^{[i]})\right) \cdot \mathbb{1}\left\{y^{[i]} > r_j\right\} + \log\left(1 - f_j\left(\mathbf{x}^{[i]}\right)\right) \cdot \mathbb{1}\left\{y^{[i]} \leq r_j\right\} \Big],$$

Conditional Subsets

$$S_1 : \text{all } \left\{\left(\mathbf{x}^{[i]}, y^{[i]}\right)\right\}, \text{ for } i \in \{1, ..., N\},$$

$$S_2 : \left\{(\mathbf{x}^{[i]}, y^{[i]}) \mid y^{[i]} > r_1\right\},$$

$$\dots$$

$$S_{K-1} : \left\{(\mathbf{x}^{[i]}, y^{[i]}) \mid y^{[i]} > r_{k-2}\right\},$$

3. CORN Framework

Loss Function

$$\boldsymbol{L}(\boldsymbol{X}, \boldsymbol{y}) = -\frac{1}{\sum_{j=1}^{K-1} |S_j|} \sum_{j=1}^{K-1} \sum_{i=1}^{|S_j|} \left[\log\left(f_j(\mathbf{x}^{[i]})\right) \cdot \mathbb{1}\left\{y^{[i]} > r_j\right\} + \log\left(1 - f_j\left(\mathbf{x}^{[i]}\right)\right) \cdot \mathbb{1}\left\{y^{[i]} \leq r_j\right\} \right],$$

$$L(\mathbf{Z}, \mathbf{y}) = -\frac{1}{\sum_{j=1}^{K-1} |S_j|} \sum_{j=1}^{K-1} \sum_{i=1}^{|S_j|} \left[\log\left(\sigma\left(\mathbf{z}^{[i]}\right)\right) \cdot \mathbb{1}\left\{y^{[i]} > r_j\right\} + \left(\log\left(\sigma\left(\mathbf{z}^{[i]}\right)\right) - \mathbf{z}^{[i]}\right) \cdot \mathbb{1}\left\{y^{[i]} \leq r_j\right\} \right].$$

4. Experiments

DATASET

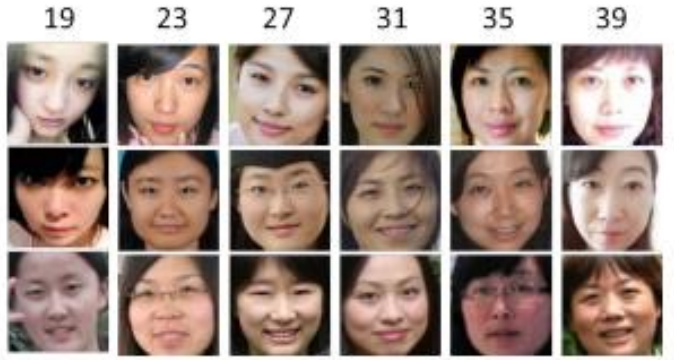
Each dataset was randomly divided into 75% training data, 5% validation data, and 20% test data.

MORPH-2 dataset

original MORPH-2 데이터셋은 16~70세까지의 연령이 포함되지만, 본 연구에서는 20,625개의 16~48까지의 33개의 연령이 골고루 포함된 balanced version 을 사용하였다.

Asian Face dataset (AFAD)

본 연구에서는 18~30까지 13개의 연령이 골고루 분포된 balanced version 을 사용하였다.



(a) female

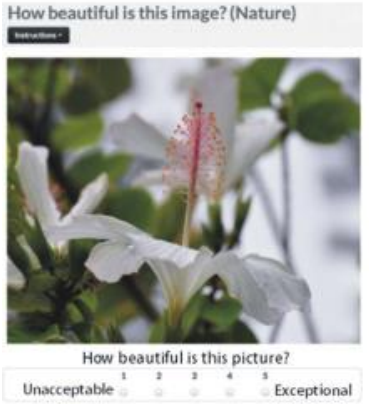
4. Experiments

DATASET

Each dataset was randomly divided into 75% training data, 5% validation data, and 20% test data.

Image Aesthetic dataset (AES)

Flickr 이미지 데이터셋의 사진들에
크라우드소싱 방법으로 beauty score 를 부여한 데이터셋.
13,868 장의 이미지 데이터셋으로, 다른 데이터셋들보다
수가 적어서 따로 balanced version 을 만들지 않고
모든 데이터를 사용하였다.



1	**Unacceptable**	Extremely low quality, out of focus, underexposed, badly framed images
2	**Flawed**	Low quality images with some technical flaws (slightly blurred, slightly over/underexposed, incorrectly framed) and without any artistic value
3	**Ordinary**	Standard quality images without technical flaws (subject well framed, in focus, and easily recognizable) and without any artistic value
4	**Professional**	Professional-quality images (flawless framing, focus, and lightning) or with some artistic value
5	**Exceptional**	Very appealing images, showing both outstanding professional quality (photographic and/or editing & techniques) and high artistic value

Table 2: Description of the five-level aesthetic judgment scale

Fireman dataset (Fireman)

Fireman dataset 은 40,768개의 인스턴스(instance), 10개의 숫자로 된 feature 및 16개의 순서형 변수를 포함하는 표 형식의 데이터셋.
16개의 ordinal class 마다 2,543개의 예제가 포함된 총 40,688 개의 예제로 구성된 이 데이터 세트의 균형 잡힌 버전을 만들어서 사용함.

response	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
12	0.487	0.072	0.004	0.833	0.765	0.6	0.132	0.886	0.073	0.342
8	0.223	0.401	0.659	0.528	0.843	0.713	0.58	0.473	0.572	0.528
15	0.903	0.913	0.94	0.979	0.561	0.744	0.627	0.818	0.309	0.51
13	0.791	0.857	0.359	0.844	0.155	0.948	0.114	0.292	0.412	0.991
16	0.326	0.593	0.085	0.927	0.926	0.633	0.431	0.326	0.031	0.73
16	0.562	0.89	0.006	0.691	0.72	0.208	0.279	0.283	0.116	0.882
12	0.481	0.613	0.499	0.572	0.914	0.783	0.204	0.428	0.828	0.487
8	0.625	0.197	0.725	0.628	0.541	0.481	0.46	0.021	0.765	0.392
13	0.21	0.519	0.029	0.61	0.724	0.515	0.371	0.731	0.575	0.73
6	0.084	0.496	0.486	0.813	0.406	0.491	0.418	0.344	0.978	0.409

4. Experiments

Evaluation Metric

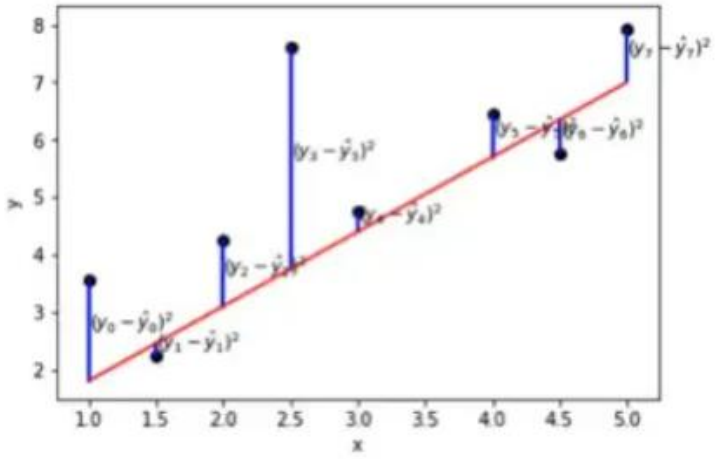
$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - h(\mathbf{x}_i)| \quad \text{and}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - h(\mathbf{x}_i))^2},$$

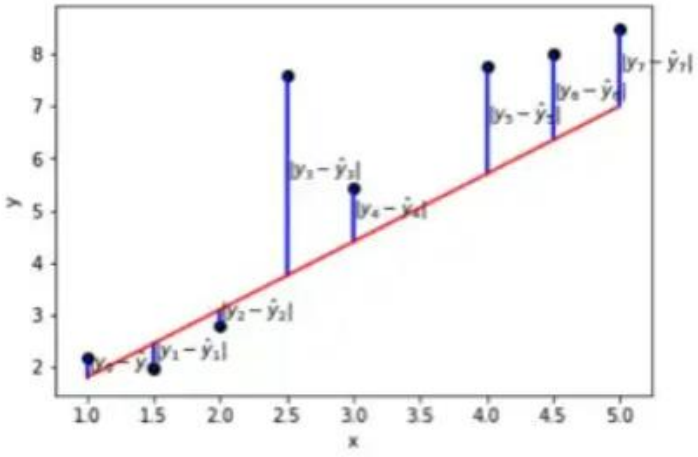
MAE: Mean Absolute Error

RMSE: Root Mean Squared Error

$$RMSE = \sqrt{\frac{1}{N}\sum_{i}^{N}(pred_i - target_i)^2}$$



$$MAE = \frac{1}{N}\sum_{i}^{N}|(pred_i - target_i)|$$



4. Experiments

IMAGE DATASET

Method	Metrics format	MORPH-2 (Balanced)		AFAD (Balanced)		Fireman	
		MAE	RMSE	MAE	RMSE	MAE	RMSE
CE-NN	AVG±SD	3.73 ± 0.12	5.04 ± 0.20	3.28 ± 0.04	4.19 ± 0.06	0.80 ± 0.01	1.14 ± 0.01
OR-NN [14]	AVG±SD	3.13 ± 0.09	4.23 ± 0.10	2.85 ± 0.03	3.48 ± 0.04	**0.76 ± 0.01**	**1.08 ± 0.01**
CORAL [1]	AVG±SD	2.99 ± 0.04	4.01 ± 0.03	2.99 ± 0.03	3.70 ± 0.07	0.82 ± 0.01	1.15 ± 0.01
CORN (ours)	AVG±SD	**2.98 ± 0.02**	**3.99 ± 0.05**	**2.81 ± 0.02**	**3.46 ± 0.02**	**0.76 ± 0.01**	**1.08 ± 0.01**

표 3: 테스트 세트의 예측 오류(낮은 것이 좋은 성능).

- 각 셀은 5번의 시험의 평균(AVG) 및 표준 편차(SD)를 나타냄.
- MORPH-2 및 AFAD 이미지 데이터셋에서는 ResNet34 가 backbone Network 로 사용됨.
- Fireman 데이터셋에서는 multilayer perceptron backbone 이 사용됨.
- 모든 데이터 세트는 균형 데이터셋임.

4. Experiments

Table S4: Prediction errors on the test sets for the RNN backbone (lower is better). The class labels for both the Coursera and TripAdvisor were balanced. Best results are highlighted in bold.

Method	Seed	TripAdvisor		Coursera	
		MAE	RMSE	MAE	RMSE
CE-RNN	0	1.13	1.56	1.01	1.48
	1	1.04	1.53	0.97	1.05
	2	1.05	1.54	1.12	1.65
	3	1.23	1.81	1.18	1.76
	4	1.03	1.52	0.84	1.26
	AVG±SD	1.10 ± 0.09	1.59 ± 0.12	1.02 ± 0.13	1.53 ± 0.19
OR-RNN [14]	0	1.06	1.53	0.98	1.34
	1	1.09	1.50	0.93	1.24
	2	1.11	1.53	1.12	1.47
	3	1.23	1.52	1.11	1.53
	4	1.07	1.40	0.85	1.16
	AVG±SD	1.11 ± 0.07	1.50 ± 0.06	1.00 ± 0.12	1.35 ± 0.15
CORAL [1]	0	1.15	1.58	0.99	1.29
	1	1.14	1.49	1.03	1.39
	2	1.16	1.46	1.14	1.40
	3	1.19	1.41	1.20	1.40
	4	1.13	1.47	0.82	1.11
	AVG±SD	1.15 ± 0.02	**1.48 ± 0.06**	1.04 ± 0.15	**1.33 ± 0.13**
CORN (ours)	0	1.09	1.55	0.95	1.37
	1	1.09	1.53	0.90	1.32
	2	1.01	1.45	1.07	1.49
	3	1.12	1.51	1.05	1.47
	4	1.03	1.46	0.78	1.14
	AVG±SD	**1.07 ± 0.05**	1.50 ± 0.04	**0.95 ± 0.12**	1.36 ± 0.14

TripAdvisor: 숙소에 대한 별점 리뷰 데이터
Coursera: 강의에 대한 별점 리뷰 데이터

이미지 데이터만큼 큰 차이는 없었지만,
CORAL, CORN 을 사용했을 때 조금이지만
개선된 성능을 보임.

Thank You