

Deep Imbalanced Regression

Yuzhe Yang¹ Kaiwen Zha¹ Ying-Cong Chen¹ Hao Wang²
Dina Katabi¹

¹MIT Computer Science & Artificial Intelligence Laboratory

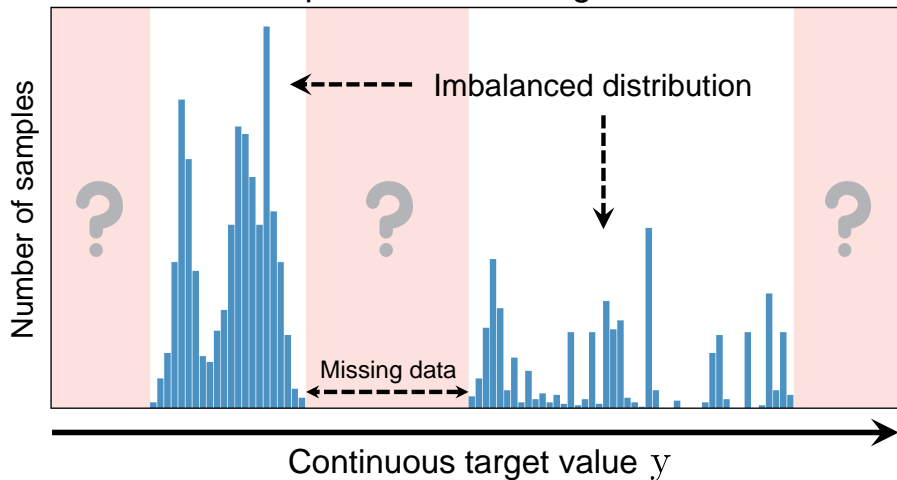
²Department of Computer Science, Rutgers University

ICML 2021

Presenter: Gianmarco Midena

26 November 2024

Deep Imbalanced Regression



Problem Settings

- $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$: training set
- $\mathbf{x}_i \in \mathbb{R}^d$: input
- $y_i \in \mathcal{Y}$: continuous label or target
- $b_i \in \mathcal{B}$: discrete label or target
- $\mathcal{Y} \subset \mathbb{R}$: continuous label space
- $\mathcal{B} = \{1, \dots, M\} \subset \mathbb{Z}^+$: index space
 - ▶ divides \mathcal{Y} into M groups (bins) with equal intervals $[t_j, t_{j+1})$
 - ▶ $\{[t_0, t_1), \dots, [t_{M-1}, t_M)\}$: discrete label space
 - ▶ $t_k \in \mathcal{Y}$
 - ▶ minimum resolution
 - ★ e.g., $\delta y \triangleq t_{j+1} - t_j = 1$ in age estimation
- $\hat{y}_i = g(\mathbf{z}_i) \in \mathbb{R}$: predicted continuous label
- $\mathbf{z}_i = f(\mathbf{x}_i; \theta) \in \mathbb{R}^{d'}$: learned representation
- θ : trainable model parameters

Evaluation

- Divide target space into disjoint regions (bins)

- ▶ *Many-shot*: > 100 training examples
- ▶ *Medium-shot*: 20-100 training examples
- ▶ *Few-shot*: < 20 training examples
- ▶ *Zero-shot*: 0 training examples
- Inspired by [Liu et al. 2019](#)

- Metrics

- ▶ Mean Absolute Error (MAE)
- ▶ Mean Squared Error (MSE)
- ▶ Pearson Correlation (PCC)
- ▶ Geometric Mean Error (GM)

$$GM = \sqrt[n]{\prod_{i=1}^n |y_i - \hat{y}_i|}$$

★ Pros: + fairness (uniformity) in prediction

Datasets - Overview

Dataset	Target type	Target range	Bin size	Max bin density	Min bin density	# Training set	# Val. set	# Test set
IMDB-WIKI-DIR	Age	0 - 186	1	7,149	1	191,509	11,022	11,022
AgeDB-DIR	Age	0 - 101	1	353	1	12,208	2,140	2,140
STS-B-DIR	Text similarity score	0 - 5	0.1	428	1	5,249	1,000	1,000
NYUD2-DIR	Depth	0.7 - 10	0.1	1.46×10^8	1.13×10^6	50,688 (3.51×10^9)	—	654 (8.70×10^5)
SHHS-DIR	Health condition score	0 - 100	1	275	0	1,892	369	369

Table credit: [Yang et al. \(2021\)](#)

(Training) Datasets - Label Distributions

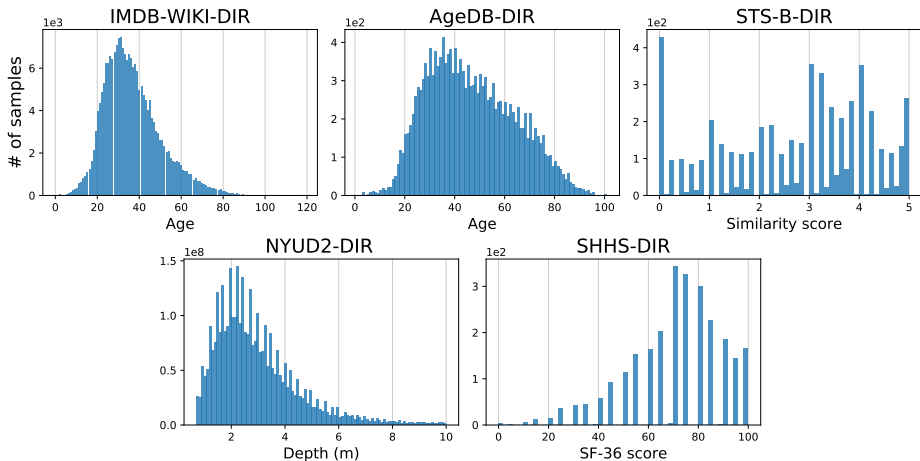
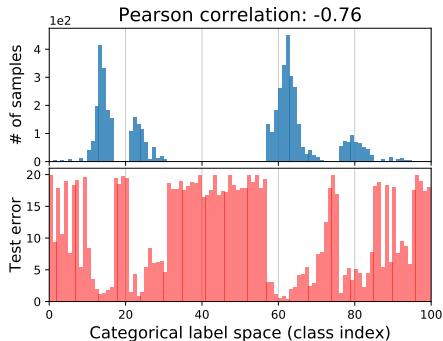


Image credit: [Yang et al. \(2021\)](#)

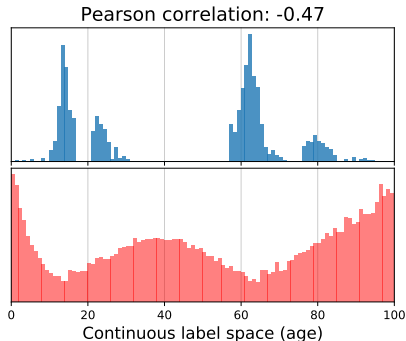
Label Distribution Smoothing (LDS)

Imbalanced Categorical vs. Continuous Label Space (1/3)



(a) Classification

- task: picture \longrightarrow class
- data source: CIFAR-100

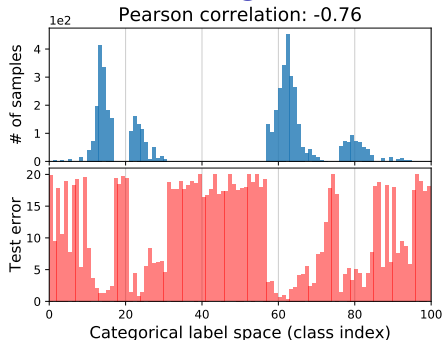


(b) Regression

- task:
person's picture \longrightarrow person's age
- age subrange: 0-99
- data source: IMDB-WIKI
- Simulated label imbalance
- Label density distributions forced to be equal
- Balanced test sets

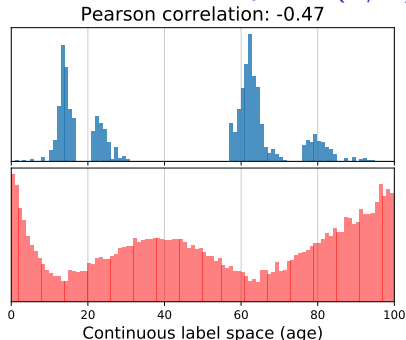
Image credit: [Yang et al. \(2021\)](#)

Imbalanced Categorical vs. Continuous Label Space (2/3)



(a) Classification

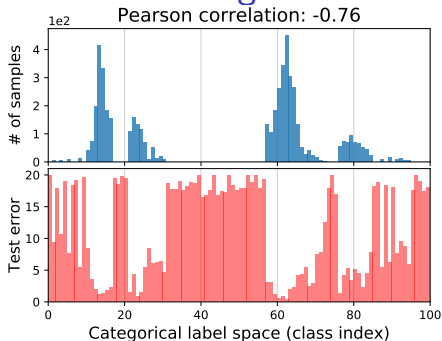
- the error distribution *correlates* with the label density distribution
- majority classes with more examples are better learned than minority classes



(b) Regression

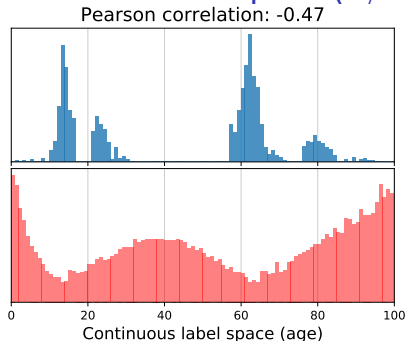
- the error distribution **DOES NOT** *correlate* well with the label density distribution
- smoother error distribution

Imbalanced Categorical vs. Continuous Label Space (3/3)



(a) Classification

- Compensating for the imbalance in the empirical label density distribution WORKS WELL.



(b) Regression

- The empirical density does not accurately reflect the imbalance as seen by the model.
- Compensating for the imbalance in the empirical label density distribution is INACCURATE.
- Proposed solution: **Label distribution smoothing**

Image credit: [Yang et al. \(2021\)](#)

Label Distribution Smoothing (LDS) - Overview

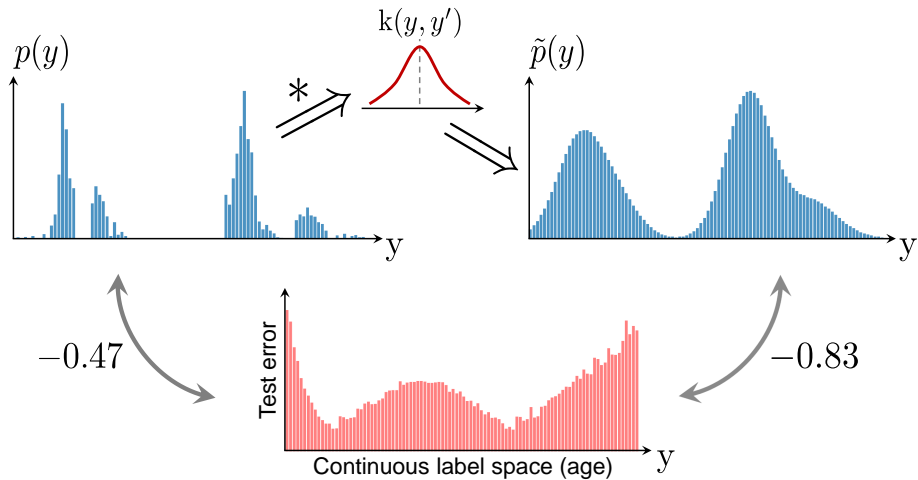


Image credit: Yang et al. (2021)

Feature Distribution Smoothing (FDS)

Feature Distribution Smoothing (FDS)

- Starting points

- ① Continuity in the **target** space \longleftrightarrow Continuity in the **feature** space

Feature Distribution Smoothing (FDS)

- Starting points
 - ① Continuity in the **target** space \longleftrightarrow Continuity in the **feature** space
 - ② Data balance \implies close feature statistics of nearby targets

Feature Distribution Smoothing (FDS)

- Starting points
 - ① Continuity in the **target** space \longleftrightarrow Continuity in the **feature** space
 - ② Data balance \implies close feature statistics of nearby targets
- Feature statistics: mean and variance w.r.t. each bin

$$\{\boldsymbol{\mu}_b, \boldsymbol{\sigma}_b\}_{b=1}^B$$

Feature Distribution Smoothing (FDS)

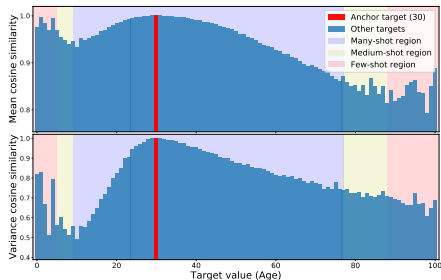
- Starting points
 - 1 Continuity in the **target** space \longleftrightarrow Continuity in the **feature** space
 - 2 Data balance \implies close feature statistics of nearby targets
- Feature statistics: mean and variance w.r.t. each bin

$$\{\mu_b, \sigma_b\}_{b=1}^B$$

- (next slides) Feature statistics similarity: cosine similarity of feature statistics between one anchor bin b_0 and all other bins
 - ▶ $b_0 = 0, 30, 60, 90$ (age): chosen anchor bins
 - ▶ different target densities:
many (>100), medium (20-100), few (<20) examples
 - ▶ task: person's picture \longrightarrow person's age
 - ▶ data source: IMDB-WIKI

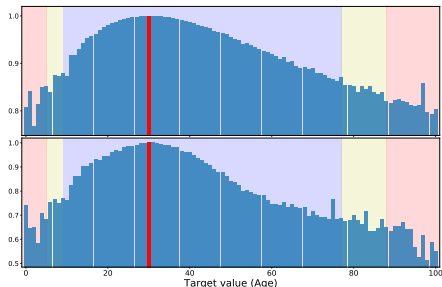
Feature statistics similarity (1/4)

Anchor age 30



(a) Baseline

- High similarity in neighbourhood
- High similarities with further regions
- Lower similarities with some closer regions



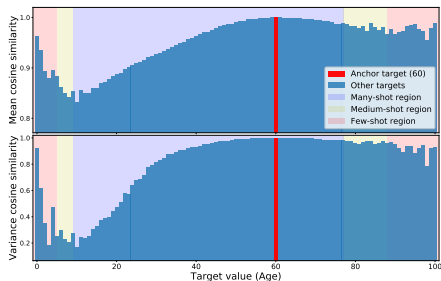
(b) FDS

- Improved feature statistics calibration:
 - ▶ High similarity only in neighbourhood
 - ▶ “The further the region the lower the similarity”
 - ▶ More gradual similarity change

Image credit: Yang et al. (2021)

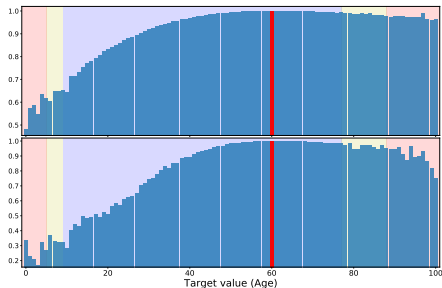
Feature statistics similarity (2/4)

Anchor age 60



(a) Baseline

- High similarity in neighbourhood
- High similarities with further regions
- Lower similarities with some closer regions



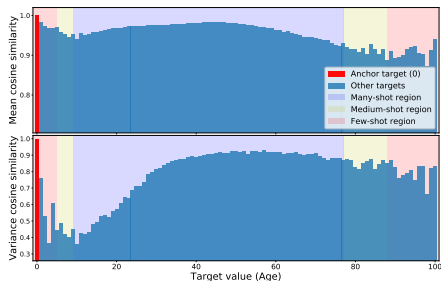
(b) FDS

- Improved feature statistics calibration:
 - ▶ High similarity only in neighbourhood
 - ▶ “The further the region the lower the similarity”
 - ▶ More gradual similarity change

Image credit: Yang et al. (2021)

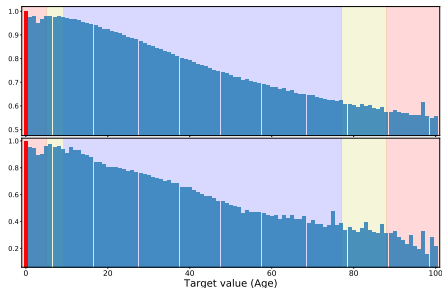
Feature statistics similarity (3/4)

Anchor age 0



(a) Baseline

- High similarity in neighbourhood for mean
- High similarities with further regions
- Lower similarities with some closer regions, e.g., variance neighbourhood



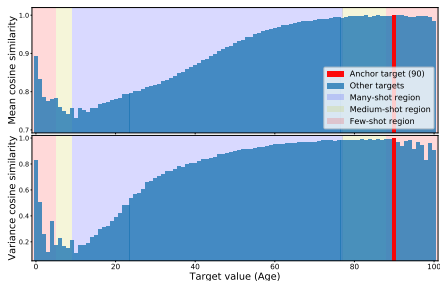
(b) FDS

- Improved feature statistics calibration:
 - ▶ High similarity only in neighbourhood
 - ▶ “The further the region the lower the similarity”
 - ▶ More gradual similarity change

Image credit: Yang et al. (2021)

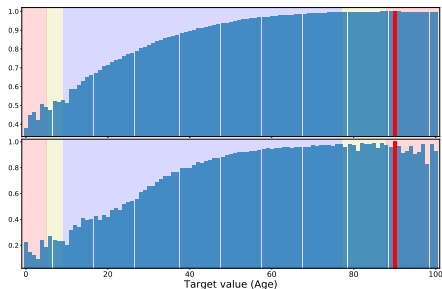
Feature statistics similarity (4/4)

Anchor age 90



(a) Baseline

- High similarity in neighbourhood, esp. for mean
- High similarities with further regions
- Lower similarities with some closer regions

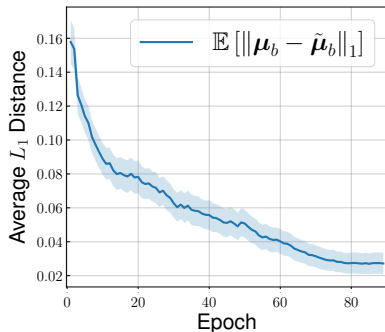


(b) FDS

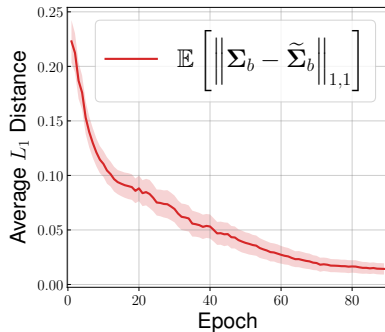
- Improved feature statistics calibration:
 - ▶ High similarity only in neighbourhood
 - ▶ “The further the region the lower the similarity”
 - ▶ More gradual similarity change

Image credit: Yang et al. (2021)

Change of feature statistics w.r.t. epoch



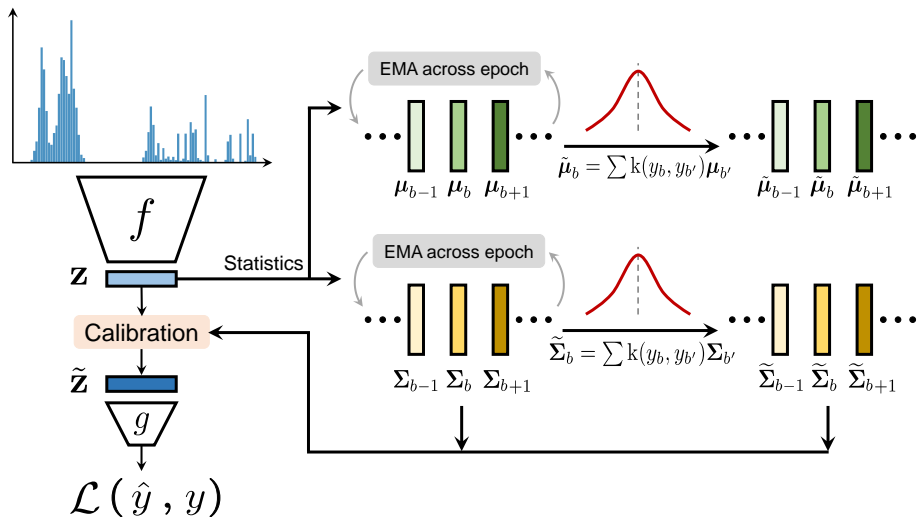
(a) Mean



(b) Variance

- μ, Σ : Running mean and variance
- $\tilde{\mu}, \tilde{\Sigma}$: Smoothed mean and variance

The FDS Algorithm



Baselines (1/2)

- Vanilla: neglects data imbalance
- Synthetic samples
 - ▶ SMOTER ([Torgo et al. 2013](#))
 - ① Defines frequent and rare regions using label density.
 - ② Creates synthetic samples for pre-defined rare regions by linearly interpolating both inputs and labels.
 - ▶ SMOGN ([Branco et al. 2017](#)): augments SMOTER with Gaussian noise
- Focal-R

$$\frac{1}{n} \sum_{i=1}^n \sigma(|\beta e_i|)^{\gamma} e_i$$

- ▶ Error-aware loss
- ▶ Maps the absolute error into $[0, 1]$.
- ▶ e_i : L_1 error for the i -th sample
- ▶ β, γ : hyper-parameters
- ▶ Inspired by Focal Loss ([Lin 2017](#)) for classification

Baselines (2/2)

- Regressor re-training (RRT)
 - ▶ Two-stage training
 - ① Train encoder
 - ② Re-train regressor with inverse re-weighting and frozen encoder.
 - ▶ Inspired by [Kang et al. 2019](#)
- Cost-sensitive re-weighting: re-weighting schemes based on label distribution
 - ▶ Inverse-frequency weighting (INV)
 - ▶ Square-root weighting variant (SQINV)

Results

Could LDS + FDS help when the label distribution is skewed with one or more Gaussian peaks?

- Experimental setup

- ▶ Curated skewed label distributions with 1-4 Gaussian peaks on IMDB-WIKI-DIR
- ▶ Compared with the vanilla model

- Findings

- ▶ Robustness to distribution change
- ▶ Brings improvement

Skewed label distribution with one Gaussian peak

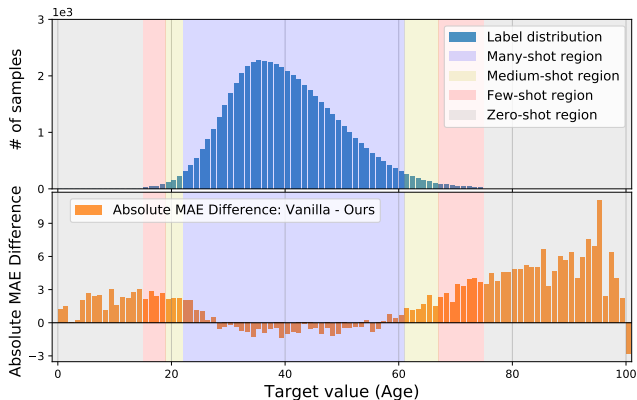


Figure: MAE gains of LDS + FDS over the vanilla model.

- Performance gains, esp. for extrapolation & interpolation

Skewed label distribution with two Gaussian peaks

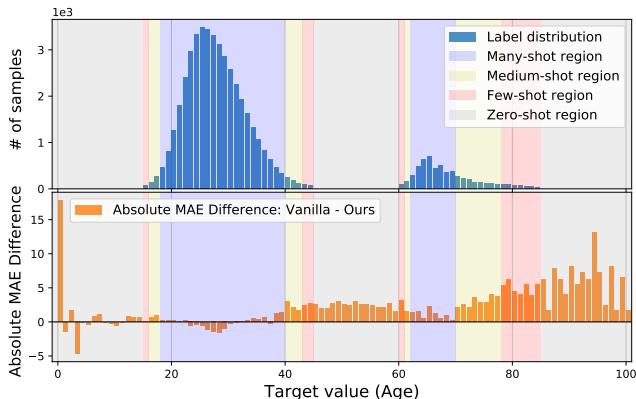


Figure: MAE gains of LDS + FDS over the vanilla model.

- Performance gains, esp. for extrapolation & interpolation

Skewed label distribution with three Gaussian peaks

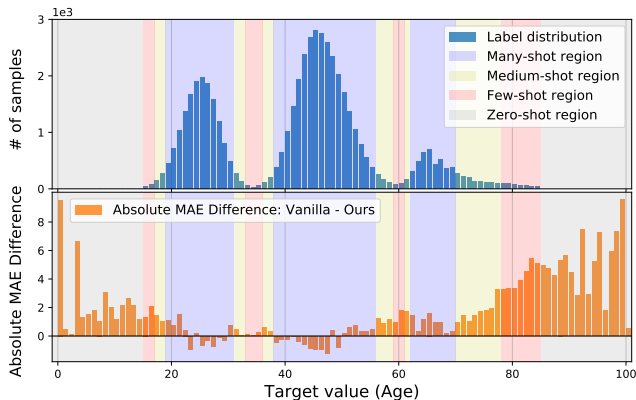


Figure: MAE gains of LDS + FDS over the vanilla model.

- Performance gains, esp. for extrapolation & interpolation

Skewed label distribution with four Gaussian peaks

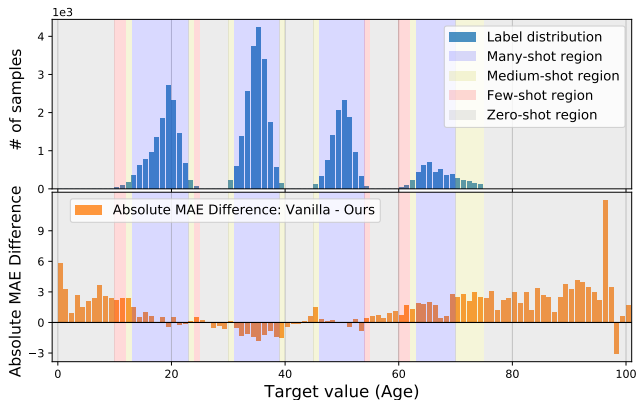


Figure: MAE gains of LDS + FDS over the vanilla model.

- Performance gains, esp. for extrapolation & interpolation

Skewed label distribution with two Gaussian peaks on IMDB-WIKI-DIR

Metrics	MAE ↓				GM ↓			
Shot	All	w/ data	Interp.	Extrap.	All	w/ data	Interp.	Extrap.
VANILLA	11.72	9.32	16.13	18.19	7.44	5.33	14.41	16.74
VANILLA + LDS	10.54	8.31	14.14	17.38	6.50	4.67	12.13	15.36
VANILLA + FDS	11.40	8.97	15.83	18.01	7.18	5.12	14.02	16.48
VANILLA + LDS + FDS	10.27	8.11	13.71	17.02	6.33	4.55	11.71	15.13
Ours (best) vs. VANILLA	+1.45	+1.21	+2.42	+1.17	+1.11	+0.78	+2.70	+1.61

Table: Interpolation & extrapolation results

- Best results by smoothing both label & feature distributions

Table credit: Yang et al. (2021)

Different skewed label distributions on IMDB-WIKI-DIR

Metrics	MAE ↓							GM ↓						
Shot	All	Many	Med.	Few	Zero	Interp.	Extrap.	All	Many	Med.	Few	Zero	Interp.	Extrap.
1 peak:														
VANILLA	11.20	6.05	11.43	14.76	22.67	—	22.67	7.02	3.84	8.67	12.26	21.07	—	21.07
VANILLA + LDS	10.09	6.26	9.91	12.12	19.37	—	19.37	6.14	3.92	6.50	8.30	16.35	—	16.35
VANILLA + FDS	11.04	5.97	11.19	14.54	22.35	—	22.35	6.96	3.84	8.54	12.08	20.71	—	20.71
VANILLA + LDS + FDS	10.00	6.28	9.66	11.83	19.21	—	19.21	6.09	3.96	6.26	8.14	15.89	—	15.89
2 peaks:														
VANILLA	11.72	6.83	11.78	15.35	16.86	16.13	18.19	7.44	3.61	8.06	12.94	15.21	14.41	16.74
VANILLA + LDS	10.54	6.72	9.65	12.60	15.30	14.14	17.38	6.50	3.65	5.65	9.30	13.20	12.13	15.36
VANILLA + FDS	11.40	6.69	11.02	14.85	16.61	15.83	18.01	7.18	3.50	7.49	12.73	14.86	14.02	16.48
VANILLA + LDS + FDS	10.27	6.61	9.46	11.96	14.89	13.71	17.02	6.33	3.54	5.68	8.80	12.83	11.71	15.13
3 peaks:														
VANILLA	9.83	7.01	9.81	11.93	20.11	—	20.11	6.04	3.93	6.94	9.84	17.77	—	17.77
VANILLA + LDS	9.08	6.77	8.82	10.48	18.43	—	18.43	5.35	3.78	5.63	7.49	15.46	—	15.46
VANILLA + FDS	9.65	6.88	9.58	11.75	19.80	—	19.80	5.86	3.83	6.68	9.48	17.43	—	17.43
VANILLA + LDS + FDS	8.96	6.88	8.62	10.08	17.76	—	17.76	5.38	3.90	5.61	7.36	14.65	—	14.65
4 peaks:														
VANILLA	9.49	7.23	9.73	10.85	12.16	8.23	18.78	5.68	3.45	6.95	8.20	9.43	6.89	16.02
VANILLA + LDS	8.80	6.98	8.26	10.07	11.26	8.31	16.22	5.10	3.33	5.07	7.08	8.47	6.66	12.74
VANILLA + FDS	9.28	7.11	9.16	10.88	11.95	8.30	18.11	5.49	3.36	6.35	8.15	9.21	6.82	15.30
VANILLA + LDS + FDS	8.76	7.07	8.23	9.54	11.13	8.05	16.32	5.05	3.36	5.07	6.56	8.30	6.34	13.10

Table credit: Yang et al. (2021)

References

- Branco, Paula, Luís Torgo, and Rita P Ribeiro (2017). "SMOGLN: a pre-processing approach for imbalanced regression". In: *First international workshop on learning with imbalanced domains: Theory and applications*. PMLR, pp. 36–50.
- Kang, Bingyi et al. (2019). "Decoupling representation and classifier for long-tailed recognition". In: *arXiv preprint arXiv:1910.09217*.
- Lin, T (2017). "Focal Loss for Dense Object Detection". In: *arXiv preprint arXiv:1708.02002*.
- Liu, Ziwei et al. (2019). "Large-scale long-tailed recognition in an open world". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2537–2546.
- Torgo, Luís et al. (2013). "Smote for regression". In: *Portuguese conference on artificial intelligence*. Springer, pp. 378–389.
- Yang, Yuzhe et al. (2021). "Delving into deep imbalanced regression". In: *International conference on machine learning*. PMLR, pp. 11842–11851.