Exercise 11 - Solution

Individual Fairness

Reliable and Trustworthy Artificial Intelligence
ETH Zurich

Recall the standard notations from the lecture:

- $\phi\colon \mathbb{R}^n \times \mathbb{R}^n \to \{0,1\}$ – binary function specifying if its inputs, x and x', are similar;
- $f_\theta\colon \mathbb{R}^n \to \mathbb{R}^k$ – encoder (data producer) mapping the input x to a learned representation $f_\theta(x)$;
- $h_\psi\colon \mathbb{R}^k \to \mathbb{R}^o$ – classifier (data consumer) mapping the learnt representation to the logits corresponding to each label $y \in \mathcal{Y}$;

Problem 1 (Similarity Sets)**.** One of the goals for the data producer f_θ is to map similar individuals close to each other:

$$\phi\left(x, x'\right) \implies \|f_\theta\left(x\right) - f_\theta\left(x'\right)\|_\infty \le \delta, \tag{1}$$

where ϕ defines input similarity and δ is a hyperparameter. Assuming that x is given (e.g., as a sample from the dataset), let $\mathcal{L}(\phi \implies \omega)(x, x')$ be the standard DL2 loss of Formula (1) at x' (as discussed in the lectures). In this problem, you will consider different notions of input similarity.

1. Let $x \in \mathbb{R}^d$ represent samples from the following dataset (only the first 2 rows and a subset of the features are shown below):

Name	Gender	Age	Salary	
Alice	Female	30	\$130,000	…
Bob	Male	32	\$121,000	
⋮				

Assume that categorical features are encoded in a finite range of numbers, e.g., by enumeration. Two individuals x and x' are similar if all of their attributes are equal except for:

- Their name or gender;
- Their age can differ by at most 3;
- Their salary can differ by at most \$10,000.

Define ϕ formally and design a DL2 query which finds a counterexample of (1).

2. Let $x \in \mathbb{R}^d$ be an image of a person and $G = D \circ E$ be a generative model with an encoder E and a decoder D. Two individuals x and x' are similar if all of their attributes are the same apart from their skin color and hair color. Propose a formal definition of the similarity set $S_\phi(x)$ of individuals similar to x and design a DL2 query which can be used to find counterexamples of (1).

   Hint: Assume access to the vectors $\boldsymbol{a}_{\text{skin}}$ and $\boldsymbol{a}_{\text{hair}}$ which manipulate these attributes in the latent space of the generative model.

3. What algorithms can be used to solve the DL2 queries which you wrote in the previous subtasks?

Solution 1.

1.

$$\begin{aligned}\phi(x, x') \coloneqq & \bigwedge_{i \in \text{Cat} \setminus \{\text{name, gender}\}} (x_i = x'_i) \\ & \bigwedge_{j \in \text{Num} \setminus \{\text{age, salary}\}} (x_j = x'_j) \\ & \bigwedge \left| x_{\text{age}} - x'_{\text{age}} \right| \leq 3 \\ & \bigwedge \left| x_{\text{salary}} - x'_{\text{salary}} \right| \leq 10{,}000\end{aligned}$$

   Assuming that the feature columns are 0-indexed, we can write the following DL2 query:

```
find x'[d]
where x'[4:] = x[4:],
  |x[2] - x'[2]| ≤ 3,
  |x[3] - x'[3]| ≤ 10000,
  ‖fθ(x) - fθ(x')‖∞ > δ
```

2. One possible approach is to define similarity in the latent space of the generative model: $S_\phi(x) = \{D\left(E\left(x\right) + t_1 \cdot \boldsymbol{a}_{\text{skin}} + t_2 \cdot \boldsymbol{a}_{\text{hair}}\right) \mid t_1, t_2 \in [-\varepsilon, \varepsilon]\}$. Here, ε is a hyperparameter controlling the maximum perturbation level and without loss of generality can be set to 1 (e.g., after scaling the attribute vectors $\boldsymbol{a}_{\text{skin}}$ and $\boldsymbol{a}_{\text{hair}}$ appropriately).

We notice that since we parameterize the similar individuals by t (the generative model G, the input x and the attribute vectors are fixed), we can design the DL2 query to search for a t which results in a counterexample:

```
find t[2]
where t in [-1, 1],
  let x' = D(E(x) + t[0]·a_skin + t[1]·a_hair),
  $\|f_\theta$(x) - $f_\theta$(x')$\|_\infty$ > $\delta$
```

3. The DL2 queries aim to find counterexamples of Formula (1), i.e., x^* such that x and x^* are similar (that is, $\phi(x, x^*)$ holds) but their representations $f_\theta(x)$ and $f_\theta(x^*)$ are far away from each other. As you will see next, x^* will be used for adversarial training, so solving the queries with PGD is a natural fit. However, in cases where the underlying search domain is low-dimensional (as in subtask 2.), adversarial random sampling can be just as efficient or effective.

Problem 2 (Enforcing Individual Fairness)**.** A typical strategy in fair representation learning is to enforce fairness by casting it as an optimization problem. In this task we will design a loss function which can be used to train the data producer f_θ. You can assume a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ and a binary function ϕ defining similarity, with respect to which the end-to-end model should be individually fair.

1. Define a loss term which enforces individual fairness and explain how you can compute or approximate it.

2. Define a loss term which makes the data producer f_θ *aware* of the potential data consumer h_ψ during training and encourages high task utility.

3. Define a loss term which promotes the transferability of the learnt representations, i.e., making them useful to other downstream tasks as well.

4. Mention at least one issue which must be considered when combining all of the above loss terms together.

Solution 2.

1. We define a *fairness* loss term:

$$\mathcal{L}_F = \max_{x' \in S_\phi(x)} \mathcal{L}(\phi \implies \omega)(x, x')$$

Then, minimizing $\mathcal{L}_F$ becomes a min-max optimization problem. Therefore, we approximate $\mathcal{L}_F$ by first computing $x^* = \arg\min_{x' \in S_\phi(x)} \mathcal{L}(\neg(\phi \implies \omega))(x, x')$

(e.g., by executing the DL2 queries which you designed in the previous problem), and then setting $\mathcal{L}_F = \mathcal{L}(\phi \implies \omega)(x, x^*)$.

Note: A slightly different approach to enforcing fairness based on adversarial learning has been adopted by [1], although they focus on group fairness and not on individual fairness.

2. We can train f_θ jointly with an auxiliary classifier q (e.g., with the same architecture as h_ψ) and introduce the *classification* loss term:

$$\mathcal{L}_C = \text{cross\_entropy}(q(f_\theta(x)), y).$$

3. In order to learn representations which are potentially useful for different downstream tasks, we should aim to preserve as much of the original signal as possible, while obfuscating only the sensitive or biased features. To that end, we introduce a decoder g, which is optimized to reconstruct the original data. The *transferability* loss term can then be defined as:

$$\mathcal{L}_T = \|x - g(f_\theta(x))\|_2$$

where any other suitable distance metric other than ℓ_2 can be used.

4. Finally, we can combine all loss terms into a single total loss which is used to train f_θ (jointly with q and g). A simple strategy is to employ linear weighting: $\mathcal{L}_{total} = \lambda_1 \mathcal{L}_F + \lambda_2 \mathcal{L}_C + \lambda_3 \mathcal{L}_T$, where λ_1, λ_2 and λ_3 are hyperparameters controlling the trade-off between the different objectives. Multi-objective optimization is an active area of research and careful exploration of the optimization landscape is required in order to setup all hyperparameters appropriately.

Problem 3 (Certifying Individual Fairness)**.** Prove the following lemma which formalizes the compositional individual fairness certificate:

Lemma 1 (Individual fairness certificate)**.** *Suppose $M = h_\psi \circ f_\theta$ with data point x and similarity notion ϕ. Furthermore, let $z = f_\theta(x)$, $S_\phi(x) = \{x' \in \mathbb{R}^d \mid \phi(x, x')\}$ and $\epsilon = \max_{x' \in S_\phi(x)} \|z - f_\theta(x')\|_\infty$. If*

$$\max_{z' \in \mathbb{B}_\infty(z, \epsilon)} h_\psi^{(y')}(z') - h_\psi^{(y)}(z') < 0 \tag{2}$$

for all labels y' different from the predicted label $y = M(x)$, then for all $x' \in S_\phi(x)$ we have $M(x) = M(x')$.

Here, $h_\psi^{(y)}(z)$ is the value of the logit corresponding to label y computed by the classifier h_ψ for the input z. Note that both ϵ and the left-hand side of Eq. (2) can be computed by MILP or soundly approximated by other incomplete certification methods.

Solution 3. The data producer computes the latent representation $z = f_\theta(x)$ and certifies that

$$\epsilon = \max_{x' \in S_\phi(x)} \|z - f_\theta(x')\|_\infty. \tag{3}$$

Thus, it immediately follows that $f_\theta(S_\phi(x)) \subseteq \mathbb{B}_\infty(z, \epsilon)$, where $\mathbb{B}_\infty(z, \epsilon)$ is the ℓ_∞-bounding box with center z and radius ϵ. Consider any label y' different from the predicted label $y = M(x)$. If the data consumer certifies that

$$\max_{z' \in \mathbb{B}_\infty(z,\epsilon)} h_\psi^{(y')}\left(z'\right) - h_\psi^{(y)}\left(z'\right) < 0, \tag{4}$$

then the classifier will predict label y for all $z' \in \mathbb{B}_\infty(z, \epsilon)$. Combining this with $f_\theta(S_\phi(x)) \subseteq \mathbb{B}_\infty(z, \epsilon)$, we have

$$\forall x' \in S_\phi(x).\ M(x) = M(x'), \tag{5}$$

implying that the end-to-end classifier is individually fair for similarity notion ϕ at data point x.

Problem 4 (Properties of Fair Representation – *from a previous exam*)**.** The encoder f_θ and the classifier h_ψ are trained jointly to compose the end-to-end model $M = h_\psi \circ f_\theta$. Moreover, f_θ is trained (via DL2 and adversarial training) to satisfy the following condition

$$\phi(x, x') \implies \|f_\theta(x) - f_\theta(x')\|_\infty \leq \delta \tag{6}$$

for a given input x and a fixed hyperparameter δ.

1. Let $\delta = 0$. The model M is such that its encoder f_θ satisfies Condition (6) for all inputs $x \in \mathbb{R}^n$ and for all similarity functions ϕ. That is,

   $$\forall \phi, x, x' \ :\ \phi(x, x') \implies \|f_\theta(x) - f_\theta(x')\|_\infty \leq 0.$$

   Show that M is a constant classifier, i.e., it classifies all inputs $x \in \mathbb{R}^n$ the same.

Consider the case $\delta > 0$. The model $M = h_\psi \circ f_\theta$ was trained on a given dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^m$. Assume that the set of all valid inputs x is closed and bounded, e.g., $x, x' \in [0, 1]^n$, and therefore $L = \max_{x \in [0,1]^n} \|f_\theta(x)\|_\infty > 0$ exists.

2. Prove that the encoder $f'_\theta(x) \coloneqq \frac{\delta}{2L} f_\theta(x)$ satisfies Condition (6) for all valid inputs x and for all similarity functions ϕ. That is,

   $$\forall \phi, x, x' \ :\ \phi(x, x') \implies \|f'_\theta(x) - f'_\theta(x')\|_\infty \leq \delta.$$

3. Construct a model $M' = h'_\psi \circ f'_\theta$ with the same accuracy on $\mathcal{D}$ as M such that f'_θ satisfies Condition (6) for all valid inputs x, x' and for all similarity functions ϕ.

4. Which of the models, M or M' (from the previous subtask), is more individually fair?

Solution 4.

1. Consider the similarity function $\phi(x, x') = 1$ for all $x, x' \in \mathbb{R}^n$. Then, we have

$$\begin{aligned} & \forall x, x' \in \mathbb{R}^n.\ \|f_\theta(x) - f_\theta(x')\|_\infty = 0 \\ \Longrightarrow\ & \forall x, x' \in \mathbb{R}^n.\ f_\theta(x) = f_\theta(x') \\ \Longrightarrow\ & \forall x, x' \in \mathbb{R}^n.\ M(x) = h_\psi\left(f_\theta(x)\right) = h_\psi\left(f_\theta(x')\right) = M(x') \end{aligned}$$

as required.

2.

$$\begin{aligned} \|f'_\theta(x) - f'_\theta(x')\|_\infty &= \left\| \frac{\delta}{2L} f_\theta(x) - \frac{\delta}{2L} f_\theta(x') \right\|_\infty \\ &= \frac{\delta}{2L} \left\| f_\theta(x) - f_\theta(x') \right\|_\infty \\ &\leq \frac{\delta}{\cancel{2}L} \cdot \cancel{2} \max\left(\|f_\theta(x)\|_\infty, \left\|f_\theta(x')\right\|_\infty \right) \\ &= \frac{\delta}{\cancel{L}} \cdot \cancel{L} = \delta, \end{aligned}$$

as required.

3. Consider

$$\begin{aligned} f'_\theta(x) &= \frac{\delta}{2L} f_\theta(x) \\ h'_\psi(z) &= h_\psi\left(\frac{2L}{\delta} z\right) \end{aligned}$$

Then:

$$M'(x) = h'_\psi(f'_\theta(x)) = h'_\psi\left(\frac{\delta}{2L} f_\theta(x)\right) = h_\psi\left(\frac{\cancel{2L}}{\cancel{\delta}} \cdot \frac{\cancel{\delta}}{\cancel{2L}} f_\theta(x)\right) = h_\psi\left(f_\theta(x)\right) = M(x).$$

That is, M' makes the same classifications as M. Therefore, M and M' have the same accuracy on $\mathcal{D}$. The encoder f'_θ satisfies Condition (6) for all valid inputs x and for all similarity functions ϕ, as proved in the previous subtask.

4. Neither. M and M' make the same classifications, so if M is individually fair at x, so is M', and vice versa.

References

[1] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. “Learning Adversarially Fair and Transferable Representations”. In: *Proceedings of the 35th International Conference on Machine Learning*. 2018.

[2] Anian Ruoss, Mislav Balunovic, Marc Fischer, and Martin Vechev. “Learning Certified Individually Fair Representations”. In: *Advances in Neural Information Processing Systems 33*. 2020.