Variational Fair Information Bottleneck

We propose an information theoretically motivated objective for learning a fair representations to solve downstream classification tasks.

To formalize the approach we first introduce some notation:

- X: Observation
- S: Sensitive Attribute
- Y: Label of the classification task (target label)
- Z: Representation of data that we want to learn

We believe the objective of fair representation can be written as:

\[
\max \{ I(Y, Z; S) - \beta I(Z, X, S) \}
\]

Using variational methods we can define a lower bound for the above formulation, and the final objective for a given \( X, Y, S \) is:

\[
\max E_{p(x,y,z)}[\log p(y|x,z)] - \beta KL(p(z|x, s)||p(z))
\]

Due to similarity to Variational Information Bottleneck we named our method Variational Fair Information Bottleneck (VFIB).

The general framework for learning a fair representation is shown in the figure below. Our proposed method doesn’t have any decoder or adversary which makes it much simpler to train. One of the primary properties of our method is using sensitive attribute for training classifier (shown in pink line).

### Related Works

The proposed objective has direction resemblances with Unsupervised variant of Variational Fair Autoencoder [1] (VFAE) and Deep Variational Information Bottleneck [2] (VIB).

### Results on Toy Problem

To evaluate the proposed method we applied VFIB on the toy problem and compared it with VFAE. After training the encoder, we trained two classifiers for sensitive attribute and the image label. Moreover, we trained a decoder network to reconstruct the original data.

While the accuracy of predicting label for VFAE is higher, its accuracy for sensitive attribute is also higher which is a non-desirable property in fair classification.

### Results on Adult Dataset

We binarized the adult dataset based on the mean and considered the age as a sensitive attribute. The target label is predicting whether a person has an income above 50k or not.

<table>
<thead>
<tr>
<th>Method</th>
<th>Label ( Y )</th>
<th>S</th>
<th>Discrimination</th>
</tr>
</thead>
<tbody>
<tr>
<td>VFAE</td>
<td>84.05 ± 0.07</td>
<td>66.27 ± 0.2</td>
<td>0.2024</td>
</tr>
<tr>
<td>VFIB</td>
<td>84.06 ± 0.08</td>
<td>65.15 ± 0.5</td>
<td>0.1956</td>
</tr>
</tbody>
</table>

The fully supervised non-fair classification accuracy is 84.35. The proposed VFIB can keep the accuracy as high as VFAE, but remove more information about age from the latent which is our goal.

### Conclusion

We proposed a new framework for learning a fair representation. We showed the effectiveness of the method on both Toy dataset and real datasets.

### Future Works

- Apply this method on more complex datasets
- Use VampPrior for modeling the prior
- Extend the framework to semi-supervised problems
- Extend the framework to multi-label problems

### References

1. Louizos, Swersky, Li, Welling, & Zemel, The Variatinal Fair Autoencoder
2. Alemi, Fischer, Dillon & Kevin Murphy, Deep Variational Information Bottleneck

Code: https://github.com/sajadn/Variational-Fair-Information-Bottleneck