Trade-offs and Guarantees of Adversarial Representation Learning for Information Obfuscation

Overview

Learning Representations that Obfuscate Sensitive Attributes:

Table: Empirical Results

<table>
<thead>
<tr>
<th></th>
<th>Adult-Gender</th>
<th>Adult-Age</th>
<th>Adult-Education</th>
<th>UTKFace-Race</th>
<th>UTKFace-Age</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Trade-offs and Guarantees of Adversarial Representation Learning for Information Obfuscation

Formal Guarantees against Attribute Inference

\[
\min_{h \in H} \max_{A \in \mathbb{A}} \mathbb{E}(h \circ f) - \lambda \left( \mathbb{P}(h_A(f(X)) = 0 \mid A = 1) + \mathbb{P}(h_A(f(X)) = 1 \mid A = 0) \right)
\]

In practice, we have:

\[
\min_{h \in H} \max_{A \in \mathbb{A}} \mathbb{C}(h \circ f) - \lambda \cdot \mathbb{C}(h_A \circ f)
\]

Theorem:

Let \( f^* \) be the optimal feature map such that \( f^* = \arg \min H(Y \mid Z = f(X)) - \lambda H(A \mid Z = f(X)) \) and define \( H^* := H(Y \mid Z = f^*(X)) \). Then for any adversary \( A \) such that \( (\hat{A}; A; Z) = 0 \), we have

\[
\mathbb{P}(\hat{A} \neq A) \geq H^*/2 \log(6/H^*).
\]

Implication: If the obfuscated representation \( Z \) contains little information on \( A \), then the inference error made by any adversary has to be large.

Inherent trade-off between Accuracy Maximization and Attribute Obfuscation

Theorem: Let \( H \subseteq 2^Z \) contains all the measurable functions from \( Z \) to \( \{0, 1\} \) and \( D_0^Y, D_1^Y \) be two distributions over \( Y \) conditioned on \( A = 0 \) and \( A = 1 \) respectively. Assume the Markov chain \( X \rightarrow Z \rightarrow A \rightarrow Y \) holds, if \( \text{Adv}(H_A) \leq d_{\text{Adv}}(D_0^Y, D_1^Y) \), then \( \forall h \in H \), we have

\[
\mathbb{E}(h \circ f) + \mathbb{E}(A \circ f) \geq \frac{1}{2} \left( d_{\text{Adv}}(D_0^Y, D_1^Y) - \sqrt{\text{Adv}(H_A \circ f)} \right)^2.
\]

Implication: if the label and the sensitive attribute are highly correlated, we cannot obfuscate the sensitive attribute while still maximizing the task accuracy simultaneously.

Preliminaries

Utility:

\[
\text{Acc}(h) := 1 - \mathbb{E}_D[|Y - h(X)|]
\]

Attribute Inference Advantage:

\[
\text{Adv}(h_A) := \max_{A \in \mathbb{A}} \left| \mathbb{P}(h_A(X) = 1 \mid A = 1) - \mathbb{P}(h_A(X) = 1 \mid A = 0) \right|
\]

- \( \text{Adv}_A(h) = 0 \) iff \( h(X); A = 0 \) and \( \text{Adv}_A(h) = 1 \) iff \( h(X) = A \) almost surely or \( h(X) = 1 - A \)
- \( \text{Adv}(H_A) = \min_{h_A \in H_A} \mathbb{P}(h_A(X) = 0 \mid A = 1) + \mathbb{P}(h_A(X) = 1 \mid A = 0) = 1 \) if \( H_A \) is symmetric: the larger the attribute inference advantage of \( H_A \), the smaller the minimum sum of Type-I and Type-II error under attacks from \( H_A \).