Feature-wise Bias Amplification

Klas Leino

CMU

Spring 2019
Overview

• **Background**
  • Motivation
  • Bias Amplification
  • Hypothesis

• **Analysis**
  • Formalization
  • Naïve Bayes case
  • Feature Asymmetry
  • Feature-wise bias amplification effect

• **Mitigation**
  • Feature parity & experts
  • Evaluation
Original Motivation

• Zhao et al. [1] show found that a model trained to predict actions and agents in images predicted that the agent in cooking scenes was a woman at a higher rate than in the training data
• This phenomenon is called bias amplification
• Setting of Zhao et al. is slightly different from our work [2], because they condition on a protected attribute, but our hypothesized mechanism is conceptually similar

[1] Zhao et al. “Men also Like Shopping: Reducing Gender Bias Amplification Using Corpus-level Constraints” ArXiv
Definition – Prior Distribution of Data

• Consider a dataset with 60% class “A” and 40% class “B”
  • We say the prior distribution of the data is (0.6, 0.4)

• Suppose a model labels 55% of its predictions “A” and 45% “B”
  • We say the prior distribution of the predictions is (0.55, 0.45)
Definition – Bias Amplification

We say a model exhibits *bias amplification* if the prior distribution of the model’s predictions does not match that of the data.

In particular, we don’t want the prior distribution of the predictions to be more disparate than that of the data.
Example

<table>
<thead>
<tr>
<th>Class</th>
<th>Man</th>
<th>Woman</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data prior</td>
<td>33%</td>
<td>67%</td>
</tr>
<tr>
<td>Pred. prior</td>
<td>16%</td>
<td>84%</td>
</tr>
</tbody>
</table>
Hypothesis

- Model learns to detect woman features, man features, and kitchen features
- Kitchen features contribute to woman class due to correlation in training data
Hypothesis

• Idea: distinguish between woman features and kitchen features, and prevent kitchen features from contributing to woman class
Recall – Experts

• An *expert* is a sub-model which has increased accuracy on a particular class from the original model
• Experts exploit the fact that performance can often be improved via the removal of “distracting” features
## Encouraging Preliminary Results

<table>
<thead>
<tr>
<th>Class</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data prior</td>
<td>20%</td>
<td>5%</td>
<td>5%</td>
<td>10%</td>
<td>10%</td>
<td>10%</td>
<td>10%</td>
<td>10%</td>
<td>10%</td>
<td>10%</td>
</tr>
<tr>
<td><strong>Original pred.</strong></td>
<td>19.6%</td>
<td>4.0%</td>
<td>3.3%</td>
<td>10.5%</td>
<td>10.2%</td>
<td>9.2%</td>
<td>11.1%</td>
<td>10.2%</td>
<td>11.2%</td>
<td>10.8%</td>
</tr>
<tr>
<td><strong>79.6% Accuracy</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Compressed pred.</strong></td>
<td>20.0%</td>
<td>4.6%</td>
<td>5.0%</td>
<td>10.1%</td>
<td>10.1%</td>
<td>10.0%</td>
<td>9.9%</td>
<td>9.9%</td>
<td>10.3%</td>
<td>10.2%</td>
</tr>
<tr>
<td><strong>79.8% Accuracy</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data prior</td>
<td>10%</td>
<td>10%</td>
<td>10%</td>
<td>10%</td>
<td>10%</td>
<td>10%</td>
<td>10%</td>
<td>10%</td>
<td>10%</td>
<td>10%</td>
</tr>
<tr>
<td><strong>Original pred.</strong></td>
<td>11.6%</td>
<td>7.8%</td>
<td>5.9%</td>
<td>10.8%</td>
<td>10.5%</td>
<td>9.4%</td>
<td>11.6%</td>
<td>10.2%</td>
<td>10.8%</td>
<td>11.3%</td>
</tr>
<tr>
<td><strong>76.9% Accuracy</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Compressed pred.</strong></td>
<td>11.9%</td>
<td>8.6%</td>
<td>7.9%</td>
<td>10.4%</td>
<td>10.5%</td>
<td>10.2%</td>
<td>10.1%</td>
<td>9.8%</td>
<td>10.1%</td>
<td>10.5%</td>
</tr>
<tr>
<td><strong>77.6% Accuracy</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Overview

• Background
  • Motivation
  • Bias Amplification
  • Hypothesis

• Analysis
  • Formalization
  • Naïve Bayes case
  • Feature Asymmetry
  • Feature-wise bias amplification effect

• Mitigation
  • Feature parity & experts
  • Evaluation
Formalization

- $D$: distribution of labeled points
- $S$: training set obtained via $n$ i.i.d. samples drawn from $D$
- $h_S$: binary classifier learned via some learning rule, $R$
Bias Amplification

Definition 1 (Bias amplification, systematic bias) Let $h_S$ be a binary classifier trained on $S \sim D^n$. The bias amplification of $h_S$ on $D$, written $B_D(h_S)$, is given by Equation 1.

$$B_D(h_S) = \mathbb{E}_{(x,y) \sim D} [h_S(x) - y]$$ (1)
Bias Amplification

Definition 1 (Bias amplification, systematic bias) Let $h_S$ be a binary classifier trained on $S \sim D^n$. The bias amplification of $h_S$ on $D$, written $B_D(h_S)$, is given by Equation 1.

$$B_D(h_S) = \mathbb{E}_{(x,y) \sim D} [h_S(x) - y]$$

Fraction of instances labeled "class 1" by the model

Fraction of instances with true label "class 1" in the data
Systematic Bias

We say that a learning rule exhibits systematic bias whenever it exhibits non-zero bias amplification on average over training samples, i.e. it satisfies Equation 2.

$$\mathbb{E}_{S \sim D^n}[B_D(h_S)] \neq 0$$
Systematic Bias

We say that a learning rule exhibits systematic bias whenever it exhibits non-zero bias amplification on average over training samples, i.e. it satisfies Equation 2.

$$\mathbb{E}_{S \sim D^n} [B_D(h_S)] \neq 0$$

(2)

i.e., given a random training set, we expect to get non-zero bias. We include this expectation since we may get bias on occasion simply by being unlucky with the training set we chose.
Simplified Formal Setting

Consider a special case of binary classification in which \( x \) are drawn from a multivariate Gaussian distribution with class means \( \mu_0^*, \mu_1^* \in \mathbb{R}^d \) and diagonal covariance matrix \( \Sigma^* \), and \( y \) is a Bernoulli random variable with parameter \( p^* \). Then \( D \) is given by Equation 3.

\[
D \triangleq \Pr[x|y] = \mathcal{N}(x|\mu_y^*, \Sigma^*), y \sim \text{Bernoulli}(p^*)
\] (3)
Simplified Formal Setting

Consider a special case of binary classification in which $x$ are drawn from a multivariate Gaussian distribution with class means $\mu_0^*, \mu_1^* \in \mathbb{R}^d$ and diagonal covariance matrix $\Sigma^*$, and $y$ is a Bernoulli random variable with parameter $p^*$. Then $D$ is given by Equation 3.

$$
D \triangleq \Pr[x|y] = \mathcal{N}(x|\mu_y^*, \Sigma^*), y \sim \text{Bernoulli}(p^*)
$$

(3)

i.e., we assume our data is Gaussian and distributed according to the Naïve Bayes assumption (features are independent conditioned on the class). This means that the Naïve Bayes classifier is the optimal classifier for our data.
Simplified Formal Setting

**Proposition 1** Let \( x \) be distributed according to Equation 3, \( y \) be Bernoulli with parameter \( p^* \), \( D \) be the Mahalanobis distance between the class means \( \mu_0^*, \mu_1^* \), and \( \beta = -D^{-1} \log(p^*/(1 - p^*)) \). Then the bias amplification of the Bayes-optimal classifier \( h^* \) is:

\[
B_D(h^*) = 1 - p^* - (1 - p^*) \Phi \left( \beta + \frac{D}{2} \right) - p^* \Phi \left( \beta - \frac{D}{2} \right)
\]

**Corollary 1** When \( x \) is distributed according to Equation 3 and \( p^* = 1/2 \), \( B_D(h^*) = 0 \).
Simplified Formal Setting

Proposition 1 Let $x$ be distributed according to Equation 3, $y$ be Bernoulli with parameter $p^*$, $D$ be the Mahalanobis distance between the class means $\mu_0^*, \mu_1^*$, and $\beta = -D^{-1} \log(p^*/(1 - p^*))$. Then the bias amplification of the Bayes-optimal classifier $h^*$ is:

$$B_D(h^*) = 1 - p^* - (1 - p^*)\Phi (\beta + \frac{D}{2}) - p^*\Phi (\beta - \frac{D}{2})$$

Corollary 1 When $x$ is distributed according to Equation 3 and $p^* = 1/2$, $B_D(h^*) = 0$.

i.e., if the classes of Naïve Bayes data are equally distributed, the optimal classifier will never exhibit bias amplification
Bias Amplification using LR & SGD

• In practice, we would be more likely to use LR models than NB models, since they make fewer assumptions about the data

• For Naïve Bayes data (Equation 3) LR converges to Bayes-optimal \textit{when given sufficient data}

• We show that prior to convergence, LR trained with SGD can exhibit systematic bias amplification \textit{even on data for which the Bayes-optimal classifier would not}
Strong & Weak Features

• At a high level, we consider “strong” features, which have high weight (high predictive power), and “weak” features, which have low weight (low predictive power)

• In practice, there is not a dichotomy, but we use a dichotomy for the sake of argument, and the general intuition holds for more nuanced cases
Direction of a Feature

We consider the direction of a feature to be the class for which the feature has positive weight. I.e., features are oriented towards whichever class they supply positive evidence for.
Feature Asymmetry

- SGD tends to overestimate the importance of weak features prior to convergence
- When there is an imbalance of weak features, the overestimation accumulates in favor of the class with more weak features
What is to Blame?

• Was this a consequence of the learning rule (SGD) or the target (LR)?

  ➢ Evidence shows effect is consistent across models trained with SGD, suggesting the effect is attributable to the learning rule.
Some Intuition

SGD effectively applies an implicit L2 regularization. This spreads out the influence of each feature rather than letting the influence go directly to the most important features. This leads weak features to be overestimated and strong features to be overestimated.

*Note that these claims are all intuitive rather than rigorous, so take them with a grain of salt*
Overview

• Background
  • Motivation
  • Bias Amplification
  • Hypothesis

• Analysis
  • Formalization
  • Naïve Bayes case
  • Feature Asymmetry
  • Feature-wise bias amplification effect

• Mitigation
  • Feature parity & experts
  • Evaluation
Possible Fixes

• Feature parity
  • Eliminate features such that the number of features in each direction is equal, removing lowest-weight features first
  • Generally too weak of an impact to make a large difference

• Experts
  • Select the expert that minimizes bias amplification subject to not harming accuracy

• L1 regularization
  • Encourages sparsity which might eliminate weak features
  • Tends not to be sufficient for desired effect
## Results

<table>
<thead>
<tr>
<th>dataset</th>
<th>$p^*(%)$</th>
<th>asymm. (%)</th>
<th>$B_D(h_S)(%)$</th>
<th>$B_D(h_S)(%)$ (post-fix)</th>
<th>acc. (%)</th>
<th>acc. (%) (post-fix)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\text{par}$</td>
<td>$\text{exp}$</td>
<td>$\ell_1$</td>
</tr>
<tr>
<td>CIFAR10</td>
<td>50.0</td>
<td>52.0</td>
<td>1.8</td>
<td>1.7</td>
<td>0.4</td>
<td>2.7</td>
</tr>
<tr>
<td>CelebA</td>
<td>50.4</td>
<td>50.2</td>
<td>7.7</td>
<td>7.7</td>
<td>0.2</td>
<td>n/a</td>
</tr>
<tr>
<td>arcene</td>
<td>56.0</td>
<td>57.7</td>
<td>2.7</td>
<td>0.6</td>
<td>1.2</td>
<td>1.7</td>
</tr>
<tr>
<td>colon</td>
<td>64.5</td>
<td>51.0</td>
<td>23.1</td>
<td>22.9</td>
<td>22.6</td>
<td>35.5</td>
</tr>
<tr>
<td>glioma</td>
<td>69.4</td>
<td>54.8</td>
<td>17.4</td>
<td>17.4</td>
<td>12.2</td>
<td>17.0</td>
</tr>
<tr>
<td>micromass</td>
<td>69.0</td>
<td>54.1</td>
<td>0.68</td>
<td>0.66</td>
<td>0.69</td>
<td>0.68</td>
</tr>
<tr>
<td>pc/mac</td>
<td>50.5</td>
<td>60.6</td>
<td>1.6</td>
<td>1.6</td>
<td>1.4</td>
<td>1.6</td>
</tr>
<tr>
<td>prostate</td>
<td>51.0</td>
<td>44.4</td>
<td>47.3</td>
<td>47.2</td>
<td>10.0</td>
<td>28.1</td>
</tr>
<tr>
<td>smokers</td>
<td>51.9</td>
<td>50.4</td>
<td>47.4</td>
<td>45.4</td>
<td>8.0</td>
<td>33.0</td>
</tr>
<tr>
<td>synthetic</td>
<td>50.0</td>
<td>99.9</td>
<td>24.1</td>
<td>17.2</td>
<td>23.6</td>
<td>5.7</td>
</tr>
</tbody>
</table>
Some Intuition

• L2 regularization spreads out influence, while L1 encourages sparsity, which eliminates weak features
• Experts are akin to a post-hoc L1/L0 regularization
• Adding L1 to training is not sufficient because SGD implicitly results in a form of L2 regularization

*Note that these claims are all intuitive rather than rigorous, so take them with a grain of salt*