Optimization-Based Approaches for Enforcing Fairness in Machine Learning

Amil Merchant [*1] **Alexander Lin** [*1]

Abstract

1. Introduction

Over the past few years, machine learning (ML) and artificial intelligence (AI) have become increasingly more common for high-stakes decision making. Researchers have proposed machine learning algorithms for applications such as credit scoring (Huang et al., 2007), personalized medicine (Poplin et al., 2018), and redicivism prediction (Tollenaar & Van der Heijden, 2013).

In light of our increased adoption of ML/AI methods, it is important that we do not allow these technologies to foster unfairness within our society. Machine learning algorithms fundamentally rely on past data in order to function. They attempt to generalize patterns found in the data and apply these patterns to make predictions in future scenarios. However, in certain situations, historical injustices against presently protected subgroups of a population may have led to the recording of biased data. Naively training a model on this biased data may lead to a biased algorithm that discriminates against these protected subgroups. Subsequently using this algorithm for high-stakes decision making may lead to further injustices and bias the collection of future data, thereby leading to a dangerous positive feedback loop.

Thus, finding ways to enforce fair predictions for machine learning algorithms is a problem of utmost importance. In this paper, we propose some methods that strive to achieve this goal. These methods are primarily optimization-based, meaning that they each involve augmenting the objective function of machine learning methods in some manner and can be seen as a form of regularization. We employ our methods in neural networks, models that have garnered a great deal of popularity in recent years due to empirical success across many domains. Our empirical results are presented on the *adult income dataset*[1], which was collected from 1994 census data (Kohavi, 1996). We show that our proposed approaches can significantly reduce model bias defined in the form of *disparate impact* and uphold desired levels of *demographic parity* without sacrificing a prohibitive amount of accuracy.

1.1. Related Work

Talk about COMPAS, other work in fairness, etc.

2. Background

2.1. Adult Income Dataset

The adult income dataset (Kohavi, 1996) contains data from $N = 32,561$ respondents to the 1994 United States Census. Each person n is characterized by $J = 14$ attributes, denoted $\boldsymbol{x}^{(n)} = \{x_1^{(n)}, \ldots, x_J^{(n)}\}$, including education level, occupation type, capital gains, capital losses, and number of hours worked per week. The goal is to predict a binary variable $y^{(n)} \in \{0, 1\}$, which indicates whether or not person n makes over \$50,000 a year.

In this case, the protected attributes $\boldsymbol{z}^{(n)}$ for person n are their *sex* and their *race*. Historical inequities have led to groups such as women and African Americans having significantly lower fractions of individuals making over \$50,000 a year. Using a model naively trained on the adult income dataset for high stakes decision making in the present day – such as estimating a person's income for loan approval or determining how much to pay a new hire – may lead to heavily biased results. Thus, there is motivation to incorporate predictive fairness into the model training process.

2.2. Disparate Impact and Demographic Parity

Disparate impact is the notion in which a model's biased classification process leads to outcomes that disproportionately hurt (or benefit) people with sensitive attributes. It was first introduced by Zafar et al. (2015). Simply removing the sensitive attributes $\boldsymbol{z}$ from the dataset and training a model on the remaining attributes $\boldsymbol{x} \setminus \boldsymbol{z}$ may still yield biased predictions, because $\boldsymbol{z}$ may be correlated with the remaining

*Equal contribution [1]Applied Mathematics 221, Harvard University, Cambridge, Massachusetts, USA. Correspondence to: Amil Merchant <amilmerchant@college.harvard.edu>, Alexander Lin <alexanderlin01@college.harvard.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

[1]This dataset is publicly available at `https://archive.ics.uci.edu/ml/datasets/adult`.

subset (Agarwal et al., 2018).

To counter disparate impact, we wish to enforce *demographic parity*, which demands that the distribution of scores for any protected classes is the same. Let $\hat{p}(y=1)$ be a model's prediction of the probability of class 1 in binary classification. Formally, demographic parity is defined as:

$$\hat{p}(y=1 \mid \boldsymbol{z}=k_1) = \hat{p}(y=1 \mid \boldsymbol{z}=k_2), \tag{1}$$

where k_1 and k_2 are different realizations of the random variable $\boldsymbol{z}$. For example, if $\boldsymbol{z}$ is sex, k_1 could be `Male` and k_2 could be `Female`. Intuitively, this means that only changing the protected attribute $\boldsymbol{z}$ should not influence the predictions in any way.

Using demographic parity as a definition of machine learning fairness offers some advantages. First and foremost, there exists legal support for this definition in the United States. In 1978, four government agencies – including the EEOC, Department of Labor, Department of Justice, and the Civil Service Commission – proposed the four-fifths (or 80%) rule as a benchmark with assessing adverse disparate impact for protected classes (Bobko & Roth, 2004). Specifically, these agencies required that

$$\min\left\{\frac{\hat{p}(y=1 \mid \boldsymbol{z}=k_1)}{\hat{p}(y=1 \mid \boldsymbol{z}=k_2)}, \frac{\hat{p}(y=1 \mid \boldsymbol{z}=k_2)}{\hat{p}(y=1 \mid \boldsymbol{z}=k_1)}\right\} \geq \frac{q}{100} \tag{2}$$

where $q=80$ in the legal definition. Note that $q=100$ corresponds to zero disparate impact and complete demographic parity. Recently, Hu and Chen (2018) additionally argue that short-term enforcement of demographic parity has long-term benefits for countering discrimination against minorities in the labor market.

3. Methods for Enforcing Demographic Parity

We present two optimization-based methods for enforcing demographic parity in neural networks.

A neural network is a cascade of linear and nonlinear transformations of the input vector $\boldsymbol{x}$ to yield an output vector $\boldsymbol{h}_L$ (Goodfellow et al., 2016). An L-layer neural network can be described by the equations,

$$\begin{aligned} \boldsymbol{h}_1 &= f^{(1)}(W^{(1)}\boldsymbol{x} + b^{(1)}), & \dots \\ \boldsymbol{h}_\ell &= f^{(\ell)}(W^{(\ell)}\boldsymbol{h}_{\ell-1} + b^{(\ell)}), & \dots \\ \boldsymbol{h}_L &= f^{(L)}(W^{(L)}\boldsymbol{h}_{L-1} + b^{(L)}), \end{aligned} \tag{3}$$

where each pair $(W^{(\ell)}, b^{(\ell)})$ parameterizes an affine transformation (via matrix multiplication and bias addition), each $f^{(\ell)}$ is a nonlinear function applied element-wise, and each $\boldsymbol{h}_\ell$ denotes an intermediary hidden state representation of the input.

In binary classifiers, it is common to let $W^{(L)}$ be a row vector, $b^{(L)}$ be a single scalar, and $f^{(L)}$ be the sigmoid function $\sigma(a) = 1/(1+\exp(-a))$. Such constraints force the final output $\hat{p} = \boldsymbol{h}_L$ to be a scalar within the range $[0,1]$, which allows us to interpret it as the estimated probability of $y=1$. For selected nonlinearities $\{f^{(\ell)}\}_{\ell=1}^L$, the weights $\{W^{(\ell)}\}_{\ell=1}^L$ and biases $\{b^{(\ell)}\}_{\ell=1}^L$ are trained to minimize the *binary cross-entropy loss* Q over the entire dataset, which is defined as

$$Q = \sum_{n=1}^N y^{(n)} \cdot \log \hat{p}^{(n)} + (1-y^{(n)}) \cdot \log(1-\hat{p}^{(n)}), \tag{4}$$

where each $\hat{p}^{(n)}$ is generated by passing $\boldsymbol{x}^{(n)}$ through the neural network.

3.1. Regularizing Decision Boundary Covariance

Zafar et al. (2015) propose regularizing the covariance between the distance to the decision boundary of a classifier and the protected classes $\boldsymbol{z}$ to enforce demographic parity. They apply their framework to logistic regression and support vector machines. We generalize this method to working with neural networks.

Using the neural network binary classifier of Equation 3, we define the *decision boundary distance* $d^{(n)}$ of training example n as the value obtained before the final nonlinearity, i.e.

$$d^{(n)} = W^{(L)}\boldsymbol{h}_{L-1}^{(n)} + b^{(L)}. \tag{5}$$

To see why $d^{(n)}$ is related to the decision boundary of the neural network classifier, observe that the estimated probability of $y^{(n)}=1$ is $\hat{p}^{(n)} = \sigma(d^{(n)})$. Thus, if $d^{(n)} > 0$, then $\hat{p}^{(n)} > 1/2$ (so it makes more sense to classify n as class 1) and if $d^{(n)} < 0$, then $\hat{p}^{(n)} < 1/2$ (so it makes more sense to classify n as class 0). Thus, the variable d encodes a scale centered at zero and characterizes the confidence of the classifier to classify as class 0 or class 1.

If the covariance between the decision boundary distance d and the protected attribute $\boldsymbol{z}$ is zero, then knowing $\boldsymbol{z}$ should have no impact on knowing $p(y \mid \boldsymbol{x})$, which is the definition of satisfying demographic parity. We can empirically estimate this covariance by observing the following:

$$\begin{aligned} \mathbb{C}\mathrm{ov}(\boldsymbol{z}, d) &= \mathbb{E}[(\boldsymbol{z}-\bar{\boldsymbol{z}}) \cdot (d-\bar{d})] \\ &= \mathbb{E}[(\boldsymbol{z}-\bar{\boldsymbol{z}}) \cdot d] - \mathbb{E}[(\boldsymbol{z}-\bar{\boldsymbol{z}})] \cdot \bar{d} \\ &= \mathbb{E}[(\boldsymbol{z}-\bar{\boldsymbol{z}}) \cdot d] - 0 \\ &\approx \frac{1}{N}\sum_{n=1}^N (\boldsymbol{z}^{(n)} - \hat{\boldsymbol{z}}) \cdot d^{(n)}, \end{aligned} \tag{6}$$

where $\hat{\boldsymbol{z}} = 1/N \cdot \sum_{n=1}^N \boldsymbol{z}^{(n)}$. Since Zafar et al. (2015) work with only convex classifiers, they simply add the following convex constraint to their logistic regression and support vector machine settings:

$$\left| \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{z}^{(n)} - \hat{\boldsymbol{z}}) \cdot d^{(n)} \right| \leq \boldsymbol{c}, \tag{7}$$

for some constant $\boldsymbol{c}$ corresponding to the level of desired demographic parity. In our neural network setting, we instead directly add the empirical covariance as a penalized regularization term to the binary cross entropy objective function of Equation 4. Thus, the full objective function is

$$\begin{aligned} Q_1 = &\sum_{n=1}^{N} y^{(n)} \log \hat{p}^{(n)} + (1 - y^{(n)}) \log(1 - \hat{p}^{(n)}) \\ &+ \lambda \cdot \left| \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{z}^{(n)} - \hat{\boldsymbol{z}}) \cdot d^{(n)} \right|, \end{aligned} \tag{8}$$

where λ controls the degree of regularization. Increasing λ will increase the penalty of the covariance and ideally lead to greater demographic parity. We wish to adjust λ so that it is large enough to satisfy fairness constraints, yet small enough to not prohibitively affect classifier accuracy.

3.2. Regularizing Representation Space Bias

4. Results

Our empirical results are evaluated on the adult income dataset. We first naively train a vanilla neural network and show how it suffers from disparate impact. Then, we apply our methods for enforcing demographic parity to exhibit how this disparate impact can be mitigated. All experiments are implemented using the PyTorch deep learning library (Paszke et al., 2017).

4.1. Vanilla Neural Network

We train a simple neural network with $L = 2$ layers that performs well on the adult income dataset. The input is $\boldsymbol{x} \setminus \boldsymbol{z}$, the set of all attributes minus sex and race. The single hidden layer $\boldsymbol{h}^{(1)}$ has 300 hidden units. We let $f^{(1)}$ be the rectified linear (ReLU) function and $f^{(2)}$ be the sigmoid function. Weights and biases $\{W^{(1)}, W^{(2)}, b^{(1)}, b^{(2)}\}$ are initialized as $\mathcal{N}(0, 1)$ random variables.

We divide the dataset of $N = 48,4842$ individuals into a training set $\mathcal{D}_{\text{train}}$ of $26,048$ people and test set $\mathcal{D}_{\text{test}}$ of $6,513$ people, which roughly corresponds to an 80%-20% split. The network is trained using the binary cross entropy loss function of Equation 4 on $\mathcal{D}_{\text{train}}$. For optimization, we use the ADAM stochastic optimizer (Kingma & Ba, 2014) with a minibatch of 1024 examples. The network is trained for 20 epochs, which are defined as passes through the entire training set.

Evaluation is performed on the test set. Test set accuracy is 85.00%, which is decent. However, there are gross violations of demographic parity.

If we observe the distributions over estimated probabilities of making over 50K divided by sex (i.e. `Male` vs. `Female`), we see that there are significant discrepancies. Figure 1 presents histograms of $\hat{p}^{(n)} \mid \boldsymbol{z}^{(n)} = \texttt{Male}$ and $\hat{p}^{(n)} \mid \boldsymbol{z}^{(n)} = \texttt{Female}$ for all $n \in \mathcal{D}_{\text{test}}$. The shapes are quite different. Let $\mathcal{D}_{\text{test}}^{\texttt{Male}}$ and $\mathcal{D}_{\text{test}}^{\texttt{Female}}$ be partitions of $\mathcal{D}_{\text{test}}$ based on sex. We see that the largest possible q that satisfies Equation 2 is $q = 40.42\%$, where q is found empirically in this example as

$$q = \frac{|\mathcal{D}_{\text{test}}^{\texttt{Female}}|^{-1} \sum_{n \in \mathcal{D}_{\text{test}}^{\texttt{Female}}} \hat{p}^{(n)}}{|\mathcal{D}_{\text{test}}^{\texttt{Male}}|^{-1} \sum_{n \in \mathcal{D}_{\text{test}}^{\texttt{Male}}} \hat{p}^{(n)}}. \tag{9}$$

This model exhibits significant bias against females, likely because it was trained on a biased dataset. Thus, it is unsuitable for use in future high-stakes decision making, such as determining how much a female should make or estimating a female's income for loan approval.

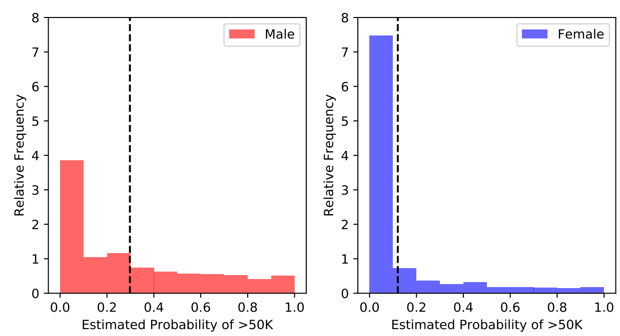

Figure 1. For the vanilla neural network, normalized histograms of estimated probabilities that males (left) and females (right) make over 50K a year. Dotted black lines indicate the means of each distribution, which is 0.298 for males and 0.120 for females.

4.2. Regularizing Decision Boundary Covariance

4.3. Regularizing Representation Space Bias

5. Discussion and Conclusion

Paper Deadline: The deadline for paper submission that is advertised on the conference website is strict. If your full, anonymized, submission does not reach us on time, it will not be considered for publication.

Anonymous Submission: ICML uses double-blind review: no identifying author information may appear on the title page or in the paper itself. Section 6.3 gives further details.

Simultaneous Submission: ICML will not accept any paper which, at the time of submission, is under review for another conference or has already been published. This policy also applies to papers that overlap substantially in technical content with conference papers under review or previously published. ICML submissions must not be submitted to other conferences during ICML's review period. Authors may submit to ICML substantially different versions of journal papers that are currently under review by the journal, but not yet accepted at the time of submission. Informal publications, such as technical reports or papers in workshop proceedings which do not appear in print, do not fall under these restrictions.

Authors must provide their manuscripts in **PDF** format. Furthermore, please make sure that files contain only embedded Type-1 fonts (e.g., using the program `pdffonts` in linux or using File/DocumentProperties/Fonts in Acrobat). Other fonts (like Type-3) might come from graphics files imported into the document.

Authors using **Word** must convert their document to PDF. Most of the latest versions of Word have the facility to do this automatically. Submissions will not be accepted in Word format or any format other than PDF. Really. We're not joking. Don't send Word.

Those who use **LaTeX** should avoid including Type-3 fonts. Those using `latex` and `dvips` may need the following two commands:

```
dvips -Ppdf -tletter -G0 -o paper.ps paper.dvi
ps2pdf paper.ps
```

It is a zero following the "-G", which tells dvips to use the config.pdf file. Newer TeX distributions don't always need this option.

Using `pdflatex` rather than `latex`, often gives better results. This program avoids the Type-3 font problem, and supports more advanced features in the `microtype` package.

Graphics files should be a reasonable size, and included from an appropriate format. Use vector formats (.eps/.pdf) for plots, lossless bitmap formats (.png) for raster graphics with sharp lines, and jpeg for photo-like images.

The style file uses the `hyperref` package to make clickable links in documents. If this causes problems for you, add `nohyperref` as one of the options to the `icml2019` usepackage statement.

5.1. Submitting Final Camera-Ready Copy

The final versions of papers accepted for publication should follow the same format and naming convention as initial submissions, except that author information (names and affiliations) should be given. See Section 6.3.2 for formatting instructions.

The footnote, "Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute." must be modified to "*Proceedings of the* 36^{th} *International Conference on Machine Learning*, Long Beach, USA, 2019. Copyright 2019 by the author(s)."

For those using the **LaTeX** style file, this change (and others) is handled automatically by simply changing `\usepackage{icml2019}` to

`\usepackage[accepted]{icml2019}`

Authors using **Word** must edit the footnote on the first page of the document themselves.

Camera-ready copies should have the title of the paper as running head on each page except the first one. The running title consists of a single line centered above a horizontal rule which is 1 point thick. The running head should be centered, bold and in 9 point type. The rule should be 10 points above the main text. For those using the **LaTeX** style file, the original title is automatically set as running head using the `fancyhdr` package which is included in the ICML 2019 style file package. In case that the original title exceeds the size restrictions, a shorter form can be supplied by using

`\icmltitlerunning{...}`

just before `\begin{document}`. Authors using **Word** must edit the header of the document themselves.

6. Format of the Paper

All submissions must follow the specified format.

6.1. Length and Dimensions

Submitted papers can be up to eight pages long, not including references, and up to twelve pages when references and acknowledgments are included. Acknowledgements should be limited to grants and people who contributed to the paper. Any submission that exceeds this page limit, or that diverges significantly from the specified format, will be rejected without review.

The text of the paper should be formatted in two columns, with an overall width of 6.75 inches, height of 9.0 inches, and 0.25 inches between the columns. The left margin should be 0.75 inches and the top margin 1.0 inch (2.54 cm). The right and bottom margins will depend on whether you print on US letter or A4 paper, but all final versions must be produced for US letter size.

The paper body should be set in 10 point type with a vertical spacing of 11 points. Please use Times typeface throughout the text.

6.2. Title

The paper title should be set in 14 point bold type and centered between two horizontal rules that are 1 point thick, with 1.0 inch between the top rule and the top edge of the page. Capitalize the first letter of content words and put the rest of the title in lower case.

6.3. Author Information for Submission

ICML uses double-blind review, so author information must not appear. If you are using LaTeX and the `icml2019.sty` file, use `\icmlauthor{...}` to specify authors and `\icmlaffiliation{...}` to specify affiliations. (Read the TeX code used to produce this document for an example usage.) The author information will not be printed unless `accepted` is passed as an argument to the style file. Submissions that include the author information will not be reviewed.

6.3.1. Self-Citations

If you are citing published papers for which you are an author, refer to yourself in the third person. In particular, do not use phrases that reveal your identity (e.g., "in previous work (**?**), we have shown ...").

Do not anonymize citations in the reference section. The only exception are manuscripts that are not yet published (e.g., under submission). If you choose to refer to such unpublished manuscripts (**?**), anonymized copies have to be submitted as Supplementary Material via CMT. However, keep in mind that an ICML paper should be self contained and should contain sufficient detail for the reviewers to evaluate the work. In particular, reviewers are not required to look at the Supplementary Material when writing their review.

6.3.2. Camera-Ready Author Information

If a paper is accepted, a final camera-ready copy must be prepared. For camera-ready papers, author information should start 0.3 inches below the bottom rule surrounding the title. The authors' names should appear in 10 point bold type, in a row, separated by white space, and centered. Author names should not be broken across lines. Unbolded superscripted numbers, starting 1, should be used to refer to affiliations.

Affiliations should be numbered in the order of appearance. A single footnote block of text should be used to list all the affiliations. (Academic affiliations should list Department, University, City, State/Region, Country. Similarly for industrial affiliations.)

Each distinct affiliations should be listed once. If an author has multiple affiliations, multiple superscripts should be placed after the name, separated by thin spaces. If the authors would like to highlight equal contribution by multiple first authors, those authors should have an asterisk placed after their name in superscript, and the term "*Equal contribution" should be placed in the footnote block ahead of the list of affiliations. A list of corresponding authors and their emails (in the format Full Name <email@domain.com>) can follow the list of affiliations. Ideally only one or two names should be listed.

A sample file with author names is included in the ICML2019 style file package. Turn on the `[accepted]` option to the stylefile to see the names rendered. All of the guidelines above are implemented by the LaTeX style file.

6.4. Abstract

The paper abstract should begin in the left column, 0.4 inches below the final address. The heading 'Abstract' should be centered, bold, and in 11 point type. The abstract body should use 10 point type, with a vertical spacing of 11 points, and should be indented 0.25 inches more than normal on left-hand and right-hand margins. Insert 0.4 inches of blank space after the body. Keep your abstract brief and self-contained, limiting it to one paragraph and roughly 4–6 sentences. Gross violations will require correction at the camera-ready phase.

6.5. Partitioning the Text

You should organize your paper into sections and paragraphs to help readers place a structure on the material and understand its contributions.

6.5.1. Sections and Subsections

Section headings should be numbered, flush left, and set in 11 pt bold type with the content words capitalized. Leave 0.25 inches of space before the heading and 0.15 inches after the heading.

Similarly, subsection headings should be numbered, flush left, and set in 10 pt bold type with the content words capitalized. Leave 0.2 inches of space before the heading and 0.13 inches afterward.

Finally, subsubsection headings should be numbered, flush left, and set in 10 pt small caps with the content words capitalized. Leave 0.18 inches of space before the heading and 0.1 inches after the heading.

Please use no more than three levels of headings.

6.5.2. Paragraphs and Footnotes

Within each section or subsection, you should further partition the paper into paragraphs. Do not indent the first line of a given paragraph, but insert a blank line between succeeding ones.

You can use footnotes[2] to provide readers with additional information about a topic without interrupting the flow of the paper. Indicate footnotes with a number in the text where the point is most relevant. Place the footnote in 9 point type at the bottom of the column in which it appears. Precede the first footnote in a column with a horizontal rule of 0.8 inches.[3]

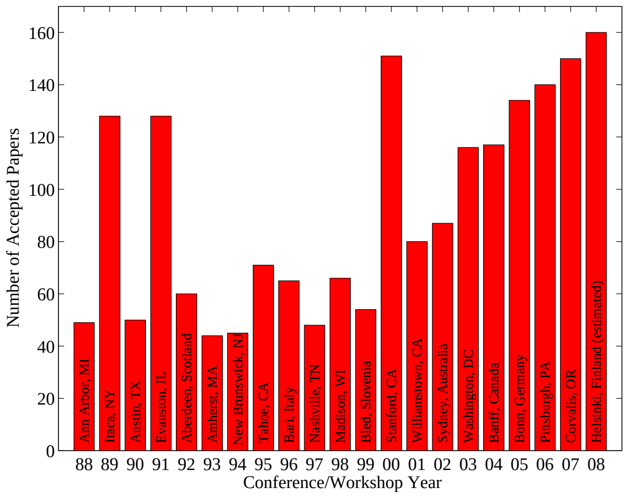

Figure 2. Historical locations and number of accepted papers for International Machine Learning Conferences (ICML 1993 – ICML 2008) and International Workshops on Machine Learning (ML 1988 – ML 1992). At the time this figure was produced, the number of accepted papers for ICML 2008 was unknown and instead estimated.

6.6. Figures

You may want to include figures in the paper to illustrate your approach and results. Such artwork should be centered, legible, and separated from the text. Lines should be dark and at least 0.5 points thick for purposes of reproduction, and text should not appear on a gray background.

Label all distinct components of each figure. If the figure takes the form of a graph, then give a name for each axis and include a legend that briefly describes each curve. Do not include a title inside the figure; instead, the caption should serve this function.

Number figures sequentially, placing the figure number and caption *after* the graphics, with at least 0.1 inches of space

[2] Footnotes should be complete sentences.

[3] Multiple footnotes can appear in each column, in the same order as they appear in the text, but spread them across columns and pages if possible.

Algorithm 1 Bubble Sort

Input: data x_i, size m
repeat
  Initialize $noChange = true$.
  for $i = 1$ **to** $m - 1$ **do**
    if $x_i > x_{i+1}$ **then**
      Swap x_i and x_{i+1}
      $noChange = false$
    end if
  end for
until $noChange$ is $true$

Table 1. Classification accuracies for naive Bayes and flexible Bayes on various data sets.

DATA SET	NAIVE	FLEXIBLE	BETTER?
BREAST	95.9$\pm$ 0.2	96.7$\pm$ 0.2	$\surd$
CLEVELAND	83.3$\pm$ 0.6	80.0$\pm$ 0.6	$\times$
GLASS2	61.9$\pm$ 1.4	83.8$\pm$ 0.7	$\surd$
CREDIT	74.8$\pm$ 0.5	78.3$\pm$ 0.6	
HORSE	73.3$\pm$ 0.9	69.7$\pm$ 1.0	$\times$
META	67.1$\pm$ 0.6	76.5$\pm$ 0.5	$\surd$
PIMA	75.1$\pm$ 0.6	73.9$\pm$ 0.5	
VEHICLE	44.9$\pm$ 0.6	61.5$\pm$ 0.4	$\surd$

before the caption and 0.1 inches after it, as in Figure 2. The figure caption should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left. You may float figures to the top or bottom of a column, and you may set wide figures across both columns (use the environment `figure*` in LaTeX). Always place two-column figures at the top or bottom of the page.

6.7. Algorithms

If you are using LaTeX, please use the “algorithm” and “algorithmic” environments to format pseudocode. These require the corresponding stylefiles, algorithm.sty and algorithmic.sty, which are supplied with this package. Algorithm 1 shows an example.

6.8. Tables

You may also want to include tables that summarize material. Like figures, these should be centered, legible, and numbered consecutively. However, place the title *above* the table with at least 0.1 inches of space before the title and the same after it, as in Table 1. The table title should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left.

Tables contain textual material, whereas figures contain graphical material. Specify the contents of each row and column in the table's topmost row. Again, you may float tables to a column's top or bottom, and set wide tables across both columns. Place two-column tables at the top or bottom of the page.

6.9. Citations and References

Please use APA reference format regardless of your formatter or word processor. If you rely on the LaTeX bibliographic facility, use `natbib.sty` and `icml2019.bst` included in the style-file package to obtain this format.

Citations within the text should include the authors' last names and year. If the authors' names are included in the sentence, place only the year in parentheses, for example when referencing Arthur Samuel's pioneering work (**?**). Otherwise place the entire reference in parentheses with the authors and year separated by a comma (**?**). List multiple references separated by semicolons (**???**). Use the 'et al.' construct only for citations with three or more authors or after listing all authors to a publication in an earlier reference (**?**).

Authors should cite their own work in the third person in the initial version of their paper submitted for blind review. Please refer to Section 6.3 for detailed instructions on how to cite your own papers.

Use an unnumbered first-level section heading for the references, and use a hanging indent style, with the first line of the reference flush against the left margin and subsequent lines indented by 10 points. The references at the end of this document give examples for journal articles (**?**), conference publications (**?**), book chapters (**?**), books (**?**), edited volumes (**?**), technical reports (**?**), and dissertations (**?**).

Alphabetize references by the surnames of the first authors, with single author entries preceding multiple author entries. Order references for the same authors by year of publication, with the earliest first. Make sure that each reference includes all relevant information (e.g., page numbers).

Please put some effort into making references complete, presentable, and consistent. If using bibtex, please protect capital letters of names and abbreviations in titles, for example, use {B}ayesian or {L}ipschitz in your .bib file.

6.10. Software and Data

We strongly encourage the publication of software and data with the camera-ready version of the paper whenever appropriate. This can be done by including a URL in the camera-ready copy. However, do not include URLs that reveal your institution or identity in your submission for review. Instead, provide an anonymous URL or upload the material as "Supplementary Material" into the CMT reviewing system. Note that reviewers are not required to look at this material when writing their review.

Acknowledgements

Do not include acknowledgements in the initial version of the paper submitted for blind review.

If a paper is accepted, the final camera-ready version can (and probably should) include acknowledgements. In this case, please place such acknowledgements in an unnumbered section at the end of the paper. Typically, this will include thanks to reviewers who gave useful comments, to colleagues who contributed to the ideas, and to funding agencies and corporate sponsors that provided financial support.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*, 2018.

Bobko, P. and Roth, P. L. The four-fifths rule for assessing adverse impact: An arithmetic, intuitive, and logical analysis of the rule and implications for future research and practice. In *Research in personnel and human resources management*, pp. 177–198. Emerald Group Publishing Limited, 2004.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. MIT press, 2016.

Hu, L. and Chen, Y. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pp. 1389–1398. International World Wide Web Conferences Steering Committee, 2018.

Huang, C.-L., Chen, M.-C., and Wang, C.-J. Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications*, 33(4):847–856, 2007.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kohavi, R. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pp. 202–207. Citeseer, 1996.

Paszke, A., Gross, S., Chintala, S., and Chanan, G. Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration. *PyTorch: Tensors and dynamic neural networks in Python with strong GPU acceleration*, 6, 2017.

Poplin, R., Varadarajan, A. V., Blumer, K., Liu, Y., McConnell, M. V., Corrado, G. S., Peng, L., and Webster, D. R. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, 2(3):158, 2018.

Tollenaar, N. and Van der Heijden, P. Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(2):565–584, 2013.

Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*, 2015.

A. Do *not* have an appendix here

Do not put content after the references. Put anything that you might normally include after the references in a separate supplementary file.

We recommend that you build supplementary material in a separate document. If you must create one PDF and cut it up, please be careful to use a tool that doesn't alter the margins, and that doesn't aggressively rewrite the PDF file. pdftk usually works fine.

Please do not use Apple's preview to cut off supplementary material. In previous years it has altered margins, and created headaches at the camera-ready stage.