Learning Fair Representations

COMP 551, Applied Machine Learning
McGill University

Babatunde Onadipe, Anton Gladyr, Saleh Bakhit

December 14, 2019

Abstract

This project is a study of Learning Fair Representations paper which is concerned with presenting a framework for achieving both types of fairness in learning models, namely, group and individual fairness. This is formed as an optimization problem of finding a good representation of the data. In particular, we reproduce the proposed baseline using two of three datasets presented in the paper. We also propose an alternative methods that achieves better results on both datasets. Since the original paper does not provide the detailed steps on reproducing the baseline, it is hard to tune a model which would perform totally the same. We obtained very similar baseline results on the German dataset and Adult dataset.

1 Introduction

In recent years, fair classification have been an active area of extensive research in the data science eld. This was sparked by the growing concern that datasets might be discriminatory against protected groups. The Canadian Human Rights Act defines discrimination as an action or a decision that treats a person or a group badly for reasons such as their race, age or disability. There has been an increased reliance on statistical inference and learning to render all sorts of decisions, including the setting of insurance rates, the allocation of police, the targeting of advertising, the issuing of bank loans, the provision of health care and the admission of students (Zemel et al., 2013).

In 2014, US President, Barack Obama set up a committee to review the data collection and analysis practices in the United States. An important outcome from the resulting report is that "Big data technologies have the tendency to cause societal damages beyond damages to privacy, such as discrimination against individuals and groups" (Podesta et al., 2014). In May 2016, the Executive Ofce of the President made the recommendation to "support research into development and fast implementation of mitigating algorithmic discrimination, building systems that support fairness and accountability, and developing strong data ethics frameworks" (Munoz et al., 2016).

Systems trained to make decisions based on historical data will inherently acquire biases from previous data. These may be mitigated by attempting to make the automated decision-making system regularize some attributes. This however, is a challenging task due to the fact that many attributes may be correlated with the protected one. The fundamental aim then is to make fair decisions, i.e., ones that are not unduly biased for or against protected subgroups in the population.

This project attempt to reproduce some aspects of a recent machine learning paper that aims to address the issue of fairness in learning models. Learning Fair Representations paper (Zemel et al., 2013) considers two major fairness classification methods, namely: group and individual fairness. Group fairness, also referred to as statistical parity, is a requirement that protected groups (groups of people qualified for special protection by a law, policy, or similar authority) should be treated similarly to the rest of the population. Individual fairness advocates that similar individuals should be treated similarly. Group fairness does not consider the individual merits and may result in choosing the less qualified members of a group, whereas individual fairness assumes a similarity metric of the individuals for the classification task. While statistical parity appears to be an important classification metric, it may still lead to outcomes that are undesirable and unfair to individuals.

Fairness was presented as an optimization problem of finding an intermediate representation of the data that best encodes the data while simultaneously making modifications to some aspects of it, i.e. removing information about membership with respect to the protected group. The developed model maps each individual, represented as a data point in a given input space, to a probability distribution in a new representation space. This representation can help make effective and reliable membership classifications. The model was developed using three datasets. A key aim in this work is to systematically come up with a measure or criteria for model selection that will lead to a good balance between fairness discrimination and prediction accuracy. Two approaches were adopted namely: selection based on minimizing the discrimination criteria as well as selection based on maximizing the difference between accuracy and discrimination. Upon examining the performance of the two approaches on the test dataset, the results showed that the developed model is capable of pushing the discrimination to very low values, while maintaining a good level of accuracy. The results of are illustrated in Table 1 and 2 of the supplementary material (Zemel et al., 2013).

In this paper, we aim to reproduce the proposed baseline, Logistic Regression, and present an improvement on it. We do so using the two UCI ML-repository datasets (German and Adult) to address the issue of group fairness.

2 Related Work

Before conducting our experiments, we believed it was in our best interest to look at other papers and reports that delved into related topics as well as also review the ones considered in Learning Fair Representations paper. Previous machine learning research into fair classification was carried out inline with two strategies. One is modifying the labels of the examples so that the proportion of positive labels are equal in the protected and unprotected groups. One can then learn a classifier with the new labels with the aim that the even labelling will aid the model to generalize to unseen dataset (Pedreschi et al., 2008; Kamiran & Calders, 2009; Luong et al., 2011). The second strategy is one that attempts to add a regularization term during the training phase which helps estimate the degree of bias as well as maximize accuracy while minimizing discrimination (Calders & Verwer, 2010; Kamishima et al.,2011).

The importance of fair representation was also expressed in the work of Gupta et al (2019), where they extended the notion of individual fairness to account for the time at which a decision is made, in settings where there exists a notion of conduciveness of decisions as perceived by the affected individuals. They introduced two definitions namely: (i) fairness-across-time (FT) and (ii) fairness-in-hindsight (FH). With FT, the treatment of individuals is required to be individually fair relative to the past as well as future. In FH, a one-sided notion of individual fairness that is defined relative to only the past decisions is required. Their work was able to show that the two definitions can have drastically different implications in the setting where the main need is to learn a utility model. Based on this, they came up with a new algorithm: Cautious Fair Exploration (CAFE), which satisfies FH and achieves sub-linear regret guarantees for a broad range of settings.

We also drew insights from the work presented by Kamishima et al (2011) where they employed a regularization strategy to quantify the degree of prejudice based on mutual information and was implemented as a regularization term in a Logistic regression model.

In the work of Dwork et al (2011), a mapping to an intermediate representation was obtained by optimizing the classification decision criteria while satisfying a Lipschitz condition on individuals, which states that nearby individuals should be mapped similarly. A key difference between the model presented in the referenced paper and the one presented in this paper is that the later's approach naturally produces out-of-sample representations, whereas the former's work left open the question of how to utilize this fair mapping for future unseen examples.

An interesting insight was presented by Joseph et al (2016) in their research which was motivated by concerns that automated decision-making procedures can unintentionally lead to discriminatory behavior. They studied a technical definition of fairness modeled after John Rawls' notion of "fair equality of opportunity". They developed an online decision making based model which anchored on an algorithm that satisfies fairness constraint, while still being able to learn at a rate that is comparable to that of the best algorithms without a fairness constraint. The gave a critical analysis of their algorithms both theoretically and experimentally and finally came up with a "discrimination index" which submitted that standard algorithms exhibit structured discriminatory behavior, whereas the "fair" algorithms do not based on the examined problem.

3 Datasets and Preprocessing

3.1 Description

As discussed in the introduction, Learning Fair Representations paper uses three different datasets to evaluate the performance of the proposed models: German, Adult and Health datasets. It is important to note that the Health dataset is unavailable as the challenge from which it was derived has ended. As a result, we use the German and Adult datasets to reproduce and improve the baseline.

3.1.1 German

The German Credit dataset is a publicly available UCI ML-repository dataset that classifies bank account holders into credit class *Good* or *Bad*. The dataset has 1000 samples, each sample (i.e. person) is described by 20 attributes (7 numerical, 13 categorical). Following the assumptions of the "Learning Fair Representations" paper, we consider *Age* as a sensitive feature. In particular, samples with age ≤ 25 are part of the protected group and samples with age > 25 are not (Kamiran et al., 2009).

3.1.2 Adult

Similar to the German dataset, the Adult income dataset is a publicly available UCI ML-repository. It shows whether a person's income exceeds 50K dollars per year (Kohavi, 1996). The dataset has 45,222 samples which are split into 2/3 train and 1/3 test sets. Each sample has 14 attributes (6 numerical, 8 categorical). The sensitive attribute is *Gender* as described in (Kohavi, 1996; Kamishima et al., 2011)). In particular, samples with $sex = Female$ are part of the protected group and samples with $sex = Male$ are not.

3.2 Setup

There are three main concerns we need to take care of in order to setup the datasets for learning, namely representing discrete attributes, continues attributes and the sensitive attribute. In order to have meaningful comparison between different models, we transform all attributes to binary attributes as described in the next subsections. For a complete description of the datasets and their representations see the supplementary material of Learning Fair Representations (Zemel et al., 2013).

3.2.1 Discrete Attributes

This is achieved by one-hot encoding all discrete attributes. This involves transforming each attribute to its N discrete values. Pandas library has a convenient function that implements one-hot encoding, *get_dummies()*.

3.2.2 Continuous Attributes

Quantization, also known as Binning, is the process of transforming continuous values to discrete ones. It does so by creating bins of ranges that the continuous values are projected onto. There are two main approaches; fixed-width and adaptive binning. fixed-width defines fixed ranges of values that are independent of the data. This is not ideal for binarizing the continuous columns as we will end up with irregularly sized bins. Instead, adaptive binning determines the ranges based on the distribution of data using what is called quantile binning. This approach partitions the data into q-quantile bins of equal sizes. In order to binarize, we use 2-quantile bins which is equivalent to the mean of the data (Sarkar, 2018).

Appendix A shows the distribution of continuous columns in both datasets along with their means. Please note that the supplementary material provided only mentions the use of quantization. The specifics described here are what we implemented after researching the topic.

3.2.3 Sensitive Attributes

The sensitive features in the German and Adult datasets, Age and Gender respectively, were transformed to binary as follows:

- Continuous (Age): threshold based on the presented cut-off. In particular, samples with age ≤ 25 take the value 1. Otherwise the value is set to 0.
- Discrete (Gender): The protected group samples ($Sex = Female$) take the value 1. Otherwise set to 0.

The value 1 indicates protected group samples while 0 indicates unprotected group samples.

Our modified German Credit dataset has 61 binary features, while modified Adult Income dataset has 103 binary features. So, our modified datasets correspond to the modified datasets from the study paper.

4 Approach to Reproduce Baseline

To reproduce the baseline of the study paper, two important metrics should be considered – accuracy and discrimination. To calculate these values, we use Eq. (1), Eq. (2), and Eq. (3), which proposed in the original paper:

$$accuracy = 1 - \frac{1}{N}\sum_{n=1}^{N}|y_n - \hat{y}_n| \quad (1)$$

$$discrimination = |\frac{\sum_{n:s_n=1}\hat{y}_n}{\sum_{n:s_n=1}1} - \frac{\sum_{n:s_n=0}\hat{y}_n}{\sum_{n:s_n=0}1}| \quad (2)$$

$$delta = accuracy - discrimination, \quad (3)$$

where N represents the number of samples in the dataset, y_n is a target value, $\hat{y}_n$ is a prediction, and S_n is a sensitive feature. The discrimination value shows the bias with respect to the attribute S. The proposed baseline model is **unregularized logistic regression**. So the goal is to reproduce both baselines when optimizing discrimination (finding the least discrimination value) and optimizing delta (finding the maximum difference between accuracy and discrimination). The target baseline values are presented in Tables 1-2.

Dataset	Delta	Accuracy	Discrimination
Adult	0.4895	0.6787	0.1892
German	0.5517	0.6790	0.1273

Table 1. Baseline: Optimizing Discrimination

Dataset	Delta	Accuracy	Discrimination
Adult	0.5971	0.7931	0.1960
German	0.5517	0.6790	0.1273

Table 2. Baseline: Optimizing Delta

It is challenging to reproduce the exact baseline results, since the study paper does not discuss many parameters needed during preprocessing and training. So, we get different results when we change logistic regression optimizer and number of iterations.

Following the supplementary materials, we split the German Credit Dataset into 5 splits, each containing 50% of the data for training, 20% for validation, and the remaining 30% for testing the final model. In the Adult Income Dataset, we divide the training set into 5 subsets, where we take one-third of each split as a validation set and the rest as a training set. In order to find the most similar logistic regression model, we implement a pipeline, which evaluates a set of unregularized logistic regression models with the different hyperparameters. The estimator runs cross-validation and reports the performance of the model based on the validation set. During cross-validation we shuffle each split and slide the position of the validation set in order to get different values for validation. So, our model's accuracy and discrimination is less biased towards original dataset distribution. After running the pipeline, we found the hyperparameters that give us the most similar results to the original report. The hyperparameters are presented in Table 3.

Optimizer	max_iter	C	penalty
Linear	500	1e40	L1

Table 3. Hyperparameters to Reproduce Baseline

Since we use *scikit-learn* library for testing different logistic regression models, we choose hyperparameters according to its implementation. Thus, we use linear optimizer with 500 iterations, and turn off regularization by setting C to a high value. The results of the model are presented in Table 4.

Dataset	Delta	Accuracy	Discrimination
Adult	0.6174	0.8314	0.2140
German	0.6429	0.6866	0.0437

Table 4. The results of chosen hyperparameters

4.1 Final Results

To make sure that the model is generalized, we run the model with the same hyperparameters many times on differently shuffled datasets. We choose the run with the least validation value of discrimination, and the run with the maximum value of delta, and run these models on the test set. The results of finding the minimum discrimination and maximum delta are presented in Tables 5 and 6.

Dataset	Delta	Accuracy	Discrimination
Adult	0.6290	0.8320	0.2029
German	0.5941	0.6733	0.0791

Table 5. Reproduced Baseline: Optimizing Discrimination

Dataset	Delta	Accuracy	Discrimination
Adult	0.6371	0.8331	0.1959
German	0.5941	0.6733	0.0791

Table 6. Reproduced Baseline: Optimizing Delta

We see when comparing Tables 1 and 2 with Tables 5 and 6 that we were able to get close to the original proposed baseline. For the German dataset, we

were able to almost exactly reproduce the baseline for Delta and accuracy, but we got the lower discrimination value of 0.0791 compared to the original 0.1273. For the Adult dataset on the other hand, we obtained better Delta and Accuracy results, but worse discrimination. We attribute these differences to the shuffling process we used.

5 Approach to Improve Baseline

To improve the results, at the first step we decided to run different standard models (SVM, Decision Trees, Regularized Logistic Regression, etc.), including the ones we used to find the baselines. So, we use the same preprocessing pipeline from outlined earlier in the report and run the models. The results were very similar to the reproduced baseline with little to no improvements.

The second step was then to run the models from the first step without encoding the continuous columns in the datasets. So, we transform both the categorical values and sensitive attribute leaving continuous attributes without quantization and run the models through the pipeline. As a result, almost all the models had very high accuracy and very low discrimination, performing better than the baseline. This approach however is not consistent with the preprocessing proposed in the original paper, making our results incomparable with both the original and reproduced baselines.

In the third approach, we decided to revert back to the original preprocessing pipeline in order to make our results comparable with all other results outlined here and in the original paper. We then eliminate the sensitive attribute column and run the models. As a result we have no discrimination due to the sensitive attribute and high accuracy making this approach our best one.

5.1 Final Results

Tables 7 and 8 show the results of the third approach. We believe this approach is a better baseline because for any attempt to reduce discrimination, we need to at least have the same accuracy with totally eliminating the column.

Model	Accuracy
Logistic regression	0.8336
Linear Discriminant Analysis	0.8325
Quadratic Discriminant Analysis	0.6585
Neural Network	0.8132
LinearSVC	0.8342

Table 7. Eliminated Sensitive Feature: Adult Dataset

Model	Accuracy
Logistic regression	0.7696
Linear Discriminant	0.7727
Quadratic Discriminant Analysis	0.5212
Neural Network	0.7515
LinearSVC	0.7696

Table 8. Eliminated Sensitive Feature: German Dataset

6 Statement of Contribution

At the start of the project, we allocated some time for all group members to understand the paper to be studied. We then divided the work to three different sections, allowing for one possible division of labour. The first section is obtaining, analyzing and representing the datasets. This was the bottleneck for the next two parts as the datasets are needed. The next two sections are reproducing and experimenting with the German and Adult datasets. Each team member took the lead in one of the presented sections while still participating in other sections as needed.

References

[1] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013, February). *Learning fair representations.* In International Conference on Machine Learning (pp. 325-333).

[2] Podesta, J., Pritzker, P., Moniz, E.J., Holdren, J., Zients, J.: *Big data: seizing opportunities, preserving values.* Executive Office of the President (2014).

[3] Munoz, C., Smith, M., Patil, D.: *Big data: a report on algorithmic systems, opportunity, and civil rights.* Executive Ofce of the President (2016)

[4] Luong, B., Ruggieri, S., and Turini, F. : *k-NN as an implementation of situation testing for discrimination discovery and prevention.* In Proceedings of the 17th ACM KDD Conference, pp. 502-510, 2011.

[5] Kamishima, T., Akaho, S., and Sakuma, J.: *Fairness-aware learning through regularization approach.* In IEEE 11th International Conference on Data Mining, pp. 643-650, 2011.

[6] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R.: *Fairness through Awareness.* In Proceedings of Innovations of Theoretical Computer Science, 2011.

[7] Pedreschi, D., Ruggieri, S., and Turini, F. *Discrimination-aware data mining.* In Proceedings of the 14th ACM KDD Conference, pp. 560-568, 2008.

[8] Kamiran, F. and Calders, T. *Classifying without discriminating.* In 2nd International Conference on Computer, Control and Communication, pp. 1 6, 2009.

[9] Luong, B., Ruggieri, S., and Turini, F. *k-NN as an implementation of situation testing for discrimination discovery and prevention.* In Proceedings of the 17th ACM KDD Conference, pp. 502-510, 2011.

[10] Calders, T. and Verwer, S. *Three naive Bayes approaches for discrimination-free classification.* Data Mining and Knowledge Discovery, 21:277-292, 2010.

[11] Gupta, S., & Kamble, V. (2019, June). *Individual fairness in hindsight.* In Proceedings of the 2019 ACM Conference on Economics and Computation (pp. 805-806). ACM.

[12] Joseph, M., Kearns, M., Morgenstern, J., Neel, S., & Roth, A. (2016). *Rawlsian fairness for machine learning.* arXiv preprint arXiv:1610.09559, 1(2).

[13] Dua, D. and Graff, C. (2019). *UCI Machine Learning Repository [http://archive.ics.uci.edu/ml].* Irvine, CA: University of California, School of Information and Computer Science.

[14] Kohavi, R. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. *In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996.

[15] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013, February). *Learning Fair Representations: Supplementary Materials [http://proceedings.mlr.press/v28/zemel13-supp.pdf]* In International Conference on Machine Learning

[16] Sarkar, D (2018, January). *Continuous Numeric Data: Strategies for working with continuous, numerical data [https://towardsdatascience.com/understanding-feature-engineering-part-1-continuous-numeric-data-da4e47099a7b]*

A Distribution of Continuous Attributes

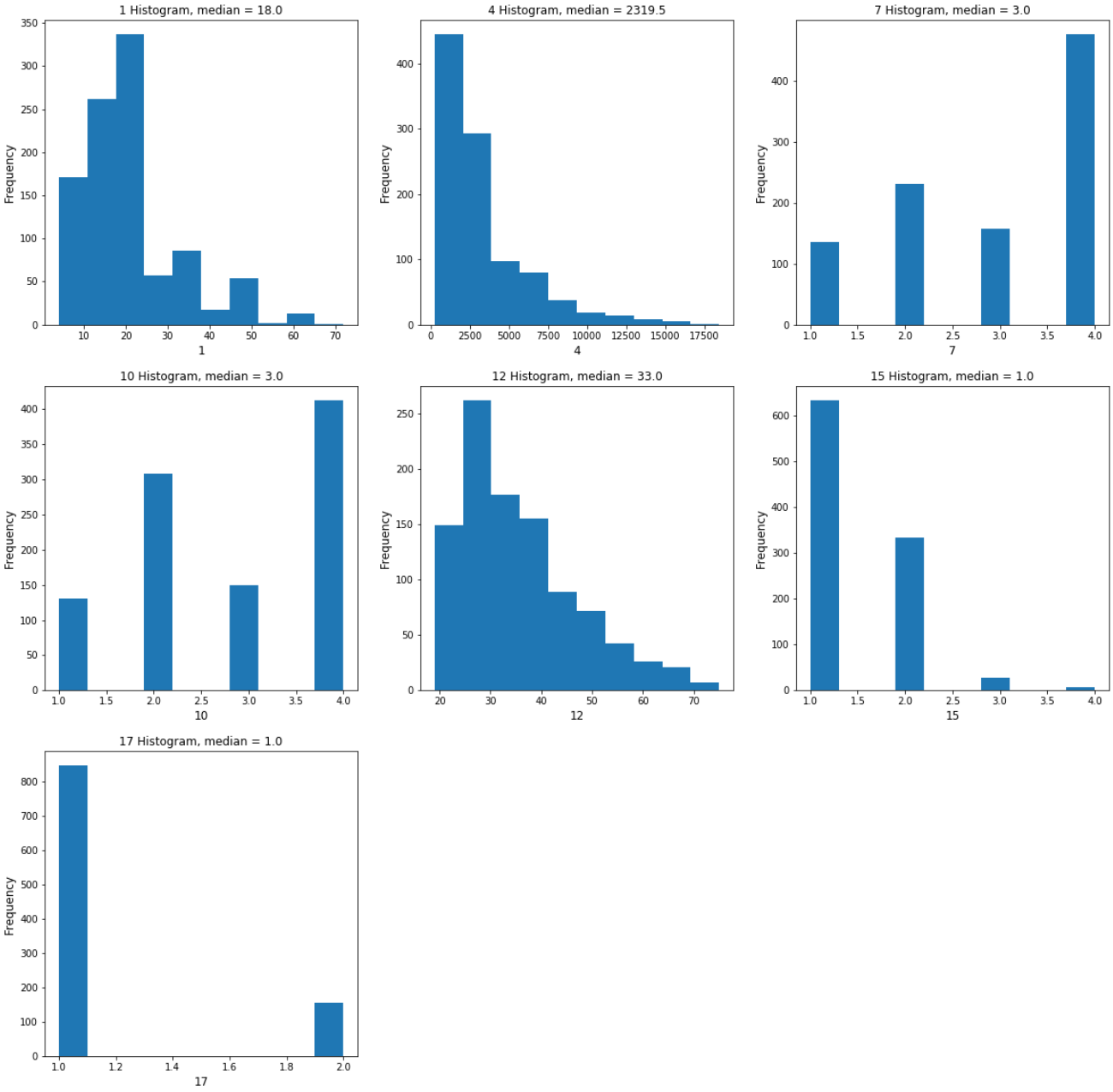

Figure 1: Distribution of Continuous Attributes - German Dataset

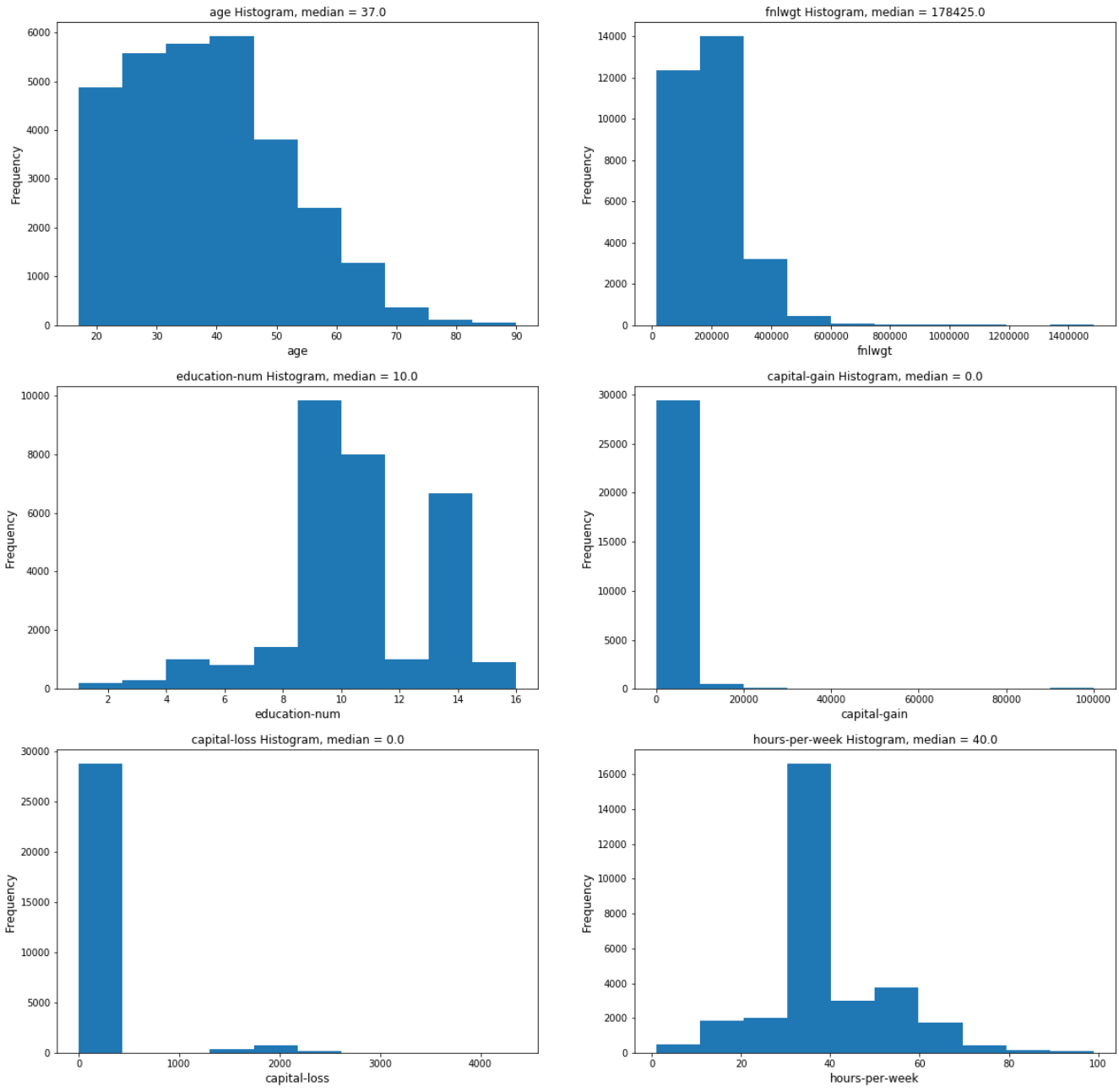

Figure 2: Distribution of Continuous Attributes - Adult Dataset