CSC4009: FIP-ML

Assignment 1: 30% Weighting for the Module

Analysing Bias in ML Tasks on the Census-Income Dataset

Released: 22nd January 2021

V1.0 (22nd January 2021)

V1.1 (This Version) (30th January 2021)

Please refer edit history at bottom to view the changes across versions.

Abstract

In this assignment, you will implement one basic analytics task over the Census-Income dataset that contains data about individuals. This will be followed by an analysis and evaluation of bias involved in the process. This is intended as an exploratory task, and you are expected to use the concepts of fairness discussed in the lectures to analyse bias in ways that you assess to be appropriate or useful. At the end of this assignment, you will be required to produce a 3-page report detailing the task performed, your major findings and analysis over them.

Submission Due: 5pm on February 19th, 2021

Dataset

The dataset that will be used for this assignment is the Census-Income dataset, sometimes referred to as the Adult dataset, available from http://archive.ics.uci.edu/ml/datasets/Census+Income . This dataset contains information about ~48k individuals across several attributes detailed in the webpage. Among the attributes are also ones that may be regarded as protected/sensitive (as discussed in the lectures), such as sex, race, relationship, marital-status etc. There are also other attributes that pertain to employment, education or finances such as workclass, education, capital-gain etc. The original intent of the dataset was to design a classifier to accurately predict whether an individual earns <=50k$ or >50k$, but the dataset has been used for a variety of other analytics tasks as well.

Analytics Task

You will choose, for this assignment, any one of the following tasks:

- **Classification**: The task of building a classifier (using an appropriate train/test split of the data) to predict whether an individual earns <=50k$ or >50k$. Popular clustering algorithms include AdaBoost, Decision Trees, Random Forest, SVM, or neural methods.
- **Clustering**: The task of grouping the individuals in the dataset into a specified number of clusters using an appropriate clustering algorithm. Popular clustering algorithms include K-Means, Hierarchical Agglomerative Clustering, DBSCAN etc.
- **Retrieval**: The task of developing a retrieval mechanism which, given a specific individual, will retrieve the top-k individuals that are regarded to be similar to the query individual. Implementing a retrieval method involves determining an appropriate notion of similarity between individuals' records, followed by choosing the top-k similar records to a query record.

Once you have decided on one of the above tasks, you will also need to identify an implementation of a classification, clustering or retrieval method. **You are required to choose any one technique** and are *not* expected to try multiple classifiers, multiple clustering algorithms, or multiple retrieval methods. You could use implementations of classifiers and clustering algorithms in the Weka toolkit, or within other Python-based machine learning frameworks. For retrieval, you may choose to use retrieval toolkits such as Apache Lucene. Alternatively, you may write your retrieval method from scratch (should be quite an easy task). When there are hyperparameters (e.g., fraction to be used as training data to train the classifier, number of clusters in the output for clustering, or the choice of k in retrieving the top-k results in retrieval), you are encouraged to experiment with several settings of hyperparameters and choose those that could be considered most appropriate.

Dealing with Protected Attributes: At this stage, you may encounter questions as to whether protected attributes should be used as features for classification/clustering, or as a source of similarity for retrieval i.e., whether these attributes are to be used in the model. **You may choose two protected attributes (from among marital-status, relationship, race, sex and native-country) as attributes for bias analysis**, and either (i) exclude all protected attributes from usage in building the model, or (ii) exclude just the two chosen protected attributes from usage in building the model. Either of the above choices are fine, and you can feel free to decide yourself after considering which is better.

Evaluation of Bias

This is a very important part of the assignment and relates to analysing and evaluating bias in the application of the chosen analytics task over the data. You are required to **choose any two** from the following three as analysis exercises:

- **Individual Fairness**: In analysing individual fairness, you may like to analyse questions like:
    1. How do we quantify individual fairness in the model? A simple pathway would be to measure correlation between individual pairwise similarities and see how well similar individuals are assigned similar outcomes (e.g., classification outcome, cluster membership or membership in the top-k result set across several queries).
    2. How well does the model agree to individual fairness? We could measure cumulative agreement to individual fairness across the dataset, and also examine whether there are some subsets of the data where the model falls short in individual fairness specifically.
- **Group Fairness**: As above, a similar analysis of group fairness would involve addressing questions such as:
    1. How can we quantify group fairness in the model? For classification, this may involve measuring the accuracy of classification across various groups defined over the protected attributes chosen for analysis (and measuring uniformity). For clustering, this may involve quantifying the extent to which there is mixing of demographic groups (defined over the protected attributes chosen for analysis) within clusters. For retrieval, this may involve quantifying the extent of representations of demographic groups (defined over the protected attributes chosen for analysis) among the top-k results across several queries.
    2. How well does the model agree to group fairness? We could measure cumulative agreement to individual fairness across the dataset, and also compare the extent of achievement of group fairness across groups defined over the two protected attributes (e.g., is group fairness violated more for gender than for race? etc.).
- **Cause of Unfairness**: Analyse, through experimenting and playing with the models extensively, as to what is the cause of unfairness. This may involve addressing questions such as:
    1. How much is the data responsible for the unfairness? Is the dataset already skewed when analysed over the chosen protected attributes for analysis? Discuss how you think such skew could bring about unfairness (either individual or group or both).

2. How much is the method responsible for the unfairness? Do you think that some assumptions in the model are responsible for the unfairness (either individual or group or both) that manifests in the results? How does unfairness vary with different settings for hyperparameters?

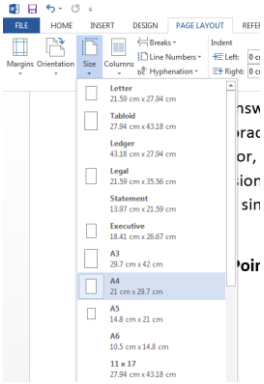
The output of the analysis exercises above could include tables indicating unfairness, charts indicating comparative analysis, and discussion as to your observations on the same. You are particularly encouraged to pay attention to the discussion part and ensure that you can illustrate reasonable depth and nuance in the analysis.

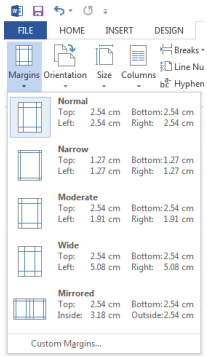
Note: While you have been provided sample questions to address in the analysis, you are encouraged to think of other meaningful analysis questions and address them as appropriate. Such chosen questions should not be exceedingly simple and should involve a fair amount of empirical and/or analytical work as the above suggested questions entail.

Format of the Submission

The 3-page report that will form the output of this assignment would be structured as below:

1. **Page 1**: This would contain information about the student (Name, Student Number, Email etc.) and also indicate the choice of the task (one of classification, clustering, or retrieval), technique used (e.g., K-Means clustering, Decision Tree etc.), toolkit used (e.g., Weka, Keras etc.) and the choice of two protected attributes for analysis. This may also include details of the chosen hyper-parameters or similar details.
2. **Pages 2-3**: **One page each** to be devoted to the two kinds of analysis you have chosen to do (among the three options detailed above). As a suggestion, you may structure this as three paragraphs: (1) Details of the analysis structure (e.g., metrics to measure unfairness, how you went about analysing the cause of unfairness), (2) quantitative results (e.g., in tables or charts) of the analysis, and (3) observations, inferences and insights from the analysis worth noting.
3. **Optional References Page (Page 4)**: If you need to include many references, you may use an optional extra page to include only a list of references.
4. **Report Formatting:** Please use the A4 page size (choosing it in Microsoft Word is shown in the screenshot below) and use the Times New Roman font with a minimum font size of 11pt (larger font sizes are fine, but they reduce the amount of text you can fit within your page limits). Additionally, you should use at least 1-inch (2.54cm) margins on all four sides, which maps to the "Normal" page margin setting in Microsoft Word (screenshot below). Any violations of recommended report formatting or non-adherence to task-wise page limits (mentioned earlier) may be penalized by up to 10% of total allocated marks.





Marking Criteria

The following would be the main marking criteria. Weightings are indicated totalling to 30% (the total weighting for this assignment):

- Evidence of successful completion of the implementation of the chosen method over the data (this is expected to be judged from the quantitative results in the analysis section). This will have a weighting of 6% (of 30% overall for the assignment)
- The two analysis tasks are assigned 12% each (thus, 24% across the two). These will be judged based on:
    - For *group fairness* and *individual fairness*: Appropriateness of the quantitative measure used, appropriate presentation of quantitative results, and depth of analysis. Among the above, the depth of analysis and insights will carry a higher weighting than the other two.
    - For *cause of unfairness*: Appropriateness of the analysis methodology on whether data or method is responsible for unfairness, depth of discussion based on the analysis.

Guidelines

- **Use of Web Resources**: You are encouraged to look up on the web as to how to accomplish the tasks above. While you may be inspired by ideas that you find from research papers or other literature, you are expected to be original in your analysis. Whenever you use something that is inspired heavily from something in the literature, you would be expected to cite that in footnotes. Using information without appropriately referencing may lead to loss of marks.
- **Individual Work**: This is strictly meant to be individual work (unlike the practical work where many of you worked in groups). While you may discuss broad ideas among your friends, you would be responsible for the decisions you make in your assignment.
- **Code**: The submission is just the report as detailed above. However, this assignment involves quite a bit of coding. The code developed is not necessary to be submitted along with the assignment. However, if it is found, during the evaluation, that the code is useful or necessary to peruse, I would request access to the code; you would be expected to respond to the request within reasonable time.
- **Integrity**: It is expected that you comply with university practices on academic integrity; this goes without saying.
- **Questions:** If you have any questions on the assignment, please enter it at the form: https://docs.google.com/forms/d/1f0dxLFhDO8dxnvzJDOzhdeYvWQYs6rPJ9EgXwjovSxM/viewform?edit_requested=true . If there are several questions, I will prepare an FAQ document which will be sent to you separately once prepared.

Edit History:

Changes from V1.0 to V1.1:

1. Within the **Dealing with Protected Attributes** section: "(ii) exclude just the protected attributes" -> "(ii) exclude just the two chosen protected attributes".
2. Added a provision of an extra page to list references. This has been added as an extra bullet point in the "**Format of the Submission**" section.