github.com /ckaestne/seai/blob/S2022/assignments/I4_fairness.md

seai/I4_fairness.md at S2022 · ckaestne/seai

10-13 minutes

Individual Assignment 4: Fairness

(11-695/17-445/17-645 Machine Learning in Production / AI Engineering)

Overview

In this assignment, you will get practice on how to systematically identify fairness issues in AI-enabled systems and think through potential fairness problems in a credit scoring scenario and in the movie streaming scenario. In particular, you will learn to (1) think about potential harms that can be caused by an unfair AI system, (2) identify potential sources of bias, (3) analyze and improve fairness of a classifier, and (4) discuss possible fairness practices throughout the system's life cycle.

Learning goals:

- Understand the potential negative impact of a biased AI-driven system on society
- Select a suitable measure for fairness given a fairness goal
- Discuss the tradeoffs and limitations of the different notions of fairness
- Measure bias of a machine learning model
- Measure the fairness–accuracy tradeoff in machine learning
- Suggest fairness practices throughout a system's life cycle

Dataset

For the first tasks of this assignment, you will work with a dataset from a credit card scoring system used by Schufa, a German private credit bureau. Schufa scores are similar to FICO scores in the US; most German citizens have a Schufa score, and these scores are used to inform financial decisions in various contexts, from banking and insurance to real estate rentals.

Despite its significant impact on the lives of German citizens, the algorithm used by the Schufa scoring system has been kept opaque from the public (not surprisingly, since Schufa is a private company). This lack of transparency also means that it is difficult to determine whether the system may be (inadvertently) making unfair decisions against certain groups of people. In response to this concern, there have been attempts to unearth the inner workings of the system and identify potential bias (most notable one being the OpenSCHUFA project).

A sample dataset from Schufa is available for download (https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29). It is kind of a "hello world example" of fairness research and you can see it widely discussed if you search online. It contains information about 1000 loan applications and includes 20 attributes that describe various characteristics of applicants, including their credit history, employment, martial status, gender, age, and job. In addition, each row in the dataset is labeled with a classification decision that states whether the application is considered "good" or "bad". More detailed information about the format of the dataset and attribute values can be found in the provided link.

Tasks and Questions

Task 1: Train Credit Scoring Model: Train a credit scoring model based on part of the labeled sample data (technically, this would be a *global surrogate* model, but for the purpose of this assignment we are assuming that the labels correspond to people actually defaulting on the loan) – pretty much any supervised classification model that comes with customizable thresholds will do. Evaluate the quality of your model with the remaining data and plot an ROC curve. Typical solutions will achieve an ROC_AUC of around 0.7.

We intend this to be easy; it is just a necessary step to produce a model you can work with. You may reuse existing code you find online for this task. For example, we recommend the following Kaggle notebook, although you are free to use any other code, as long as you credit the source.

Task 2: Measuring fairness: Consider gender and age as protected attributes and assume that the labels in the validation data are an accurate reflection of that person's risk of defaulting on a loan. For each protected attribute evaluate the fairness *of your model* using (a) anti-classification, (b) group fairness, and (c) separation as the fairness criteria. If you need to make assumptions (e.g., select thresholds) state your assumptions.

You may use existing tools like IBM's AI Fairness 360 toolkit, but it might be easier to implement this yourself and derive measures directly from tests and the confusion matrix.

Task 3: Improve model fairness: For *one* protected attribute of your choice and *each* of the three fairness measures, try to improve the fairness of your classifier and evaluate the impact on model accuracy. You will likely want to try different interventions depending on the fairness measure you try to improve. Examples of simple interventions include, eliminating the protected attribute for anti-classification, and tweaking thresholds for group fairness and separation. If you like, you can also explore more advanced interventions such as augmenting training data, preprocessing, or tweaking learning techniques, possibly with existing tools like IBM's AI Fairness 360 toolkit.

Task 4: Fairness properties: Discuss which fairness property may be suitable in the credit rating setting and why. If you could only enforce one fairness property, which one would it be?

Task 5: Fairness in movie recommendation: Consider whether the recommendation algorithm of the group project (in its current for or with possible extensions) has potential fairness issues. Assume that movie ratings users make are public to all other users. Consider not only legal issues but also potential ethical issues.

- Discuss possible harms (harms of allocation and harms of representation) the model may cause for users or other stakeholders, or discuss why such harms are unlikely. Be explicit about the groups or protected attributes you are considering, which may include attributes of users or of movies.
- Discuss possible sources of bias (at least: skewed samples, tainted examples, limited features, sample size disparity, and proxies) that may lead to unfair recommendations, and whether they are relevant in the movie recommendation setting.
- Make a suggestion for fairness-related practices you might adopt throughout the life cycle of this project. In particular, your suggestion should describe practices pertaining to at least *one* of the following activities: (1) data collection and pre-processing for fairness, (2) selection of a fairness critera to train the model, and (3) monitoring the system for biased outcome and deploying an escalation plan. Briefly justify your answer. If you recommend not to take any action, justify this decision too. You do not need to actually implement any practices within this course, now or later, but attempt to make recommendations that could *realistically* be considered in a startup competing with Netflix.

Deliverables

Submit your analysis code to a previous GitHub repository from this class and submit a report (PDF) to Gradescope. The report should have the following clearly marked sections:

- **Credit scoring model:** Without any required textual description, simply provide a link to your implementation (e.g. notebook), mention the learning technique used, report the accuracy of your baseline model.
- **Fairness measures:** In a few sentences describe how you measure the three fairness criteria and provide separate measurement results for the two protected attributes gender and age.
- **Improved fairness:** In a few sentences describe the steps taken for improving all three fairness criteria. Separately for each fairness criteria, report fairness measure and model accuracy before and after this step.
- **Credit rating discussion (0.5–1 page):** Discuss the advantages and disadvantages of all three fairness criteria and recommend a single criteria to enforce. Justify your recommendation. Mention assumptions relevant for the decision (e.g., whether the decision is made by Schufa itself or by a regulator and what their goals are).

- **Movie recommendation discussion (max 1.5 pages):** Discuss possible harms and corresponding groups and possible sources of bias in the movie recommendation scenario. Make a recommendation for which fairness property to monitor for which protected attributes, if any.

Grading

The focus of this assignment is on understanding fairness measures and considering fairness at the system level. The accuracy of the classifier does not matter, nor will we focus on the quality of the pipeline to produce it.

The assignment is worth 100 points. For full credit, we expect:

- 15 points: Credit scoring model learned and evaluated, ROC curve reported, corresponding code provided or referenced.
- 25 points: Reasonable operationalization of the three fairness criteria, clear description of the measure each (clear enough they could be independently reimplemented), and provided 6 measurement results (for the three fairness criteria and two protected attributes).
- 20 points: Improvements attempted separately for all three fairness criteria, with sufficient description to understand what was attempted. Fairness and accuracy reported before and after for each.
- 10 points: Plausible tradeoff discussion between the three fairness measures and their corresponding goals. Well-argued recommendation for which measure to adopt that is grounded in the credit scoring scenario.
- Plausible discussion of fairness issues in the movie recommendation scenario that includes:
    - 10 points: A discussion of harms of representation and harms of allocation and corresponding groups/protected attributes in the context of the movie recommendation example is included.
    - 10 points: A discussion of the five sources of bias in the context of the movie recommendation example is included.
    - 10 points: The report includes a justified recommendation of at least *one* engineering practice (which can be **one** of (1) data collection & pre-processing, (2) fairness criteria selection and (3) monitoring & escalation) or a justification why no practice shall be adopted. The recommendation and justifications are grounded in the realism of the movie recommendation scenario.

Groupwork option

To encourage deeper engagement with the content and collaboration, we provide the option for this assignment to work together with one other student in the class. We suggest teams for the assignment on Canvas.

If you work together as a team, you can either submit a joint solution or separate solutions on Gradescope. If you submit a joint solution, both team members must have contributed to the solution and both team members will receive the same grade. If you submit separate solutions, those solutions may share text and you may discuss all aspects of the assignment, but we will grade them separately. Always make sure that you indicate with whom you worked together, even if just for part of the assignment.

Groupwork is optional. You may decide to work alone.