Homework 2 - Fairness in Healthcare

Due Date Sunday, March 1 at 11:59pm

Submission You will be submitting the following five files on Markus:

- `metrics.py`: Definitions of various fairness metrics.
- `s1.ipynb`: Jupyter Notebook for Section 1.
- `s2.ipynb`: Jupyter Notebook for Section 2.
- `model.pt`: Pytorch model trained in Section 2 Question 5b.
- `report.pdf`: All of your analyses, figures, and tables.

Late Submission Homeworks will be accepted up to 3 days late, but 10% will be deducted for each day late, rounded up to the nearest day. Submissions later than 3 days after the due date will not be accepted without a valid medical certificate or other documentation of an emergency.

Collaboration What you turn in must be your own work. You may not work with anyone else on any of the problems in this assignment. If you need assistance, ask in the Piazza board for this course, or contact the instructor or TA.

Grading This assignment will be graded out of 60. There is one bonus point available. The assignment is worth 13.3% of your final grade.

Overview

Read the following items *carefully*:

- This assignment consists of two sections.
- In the first section, you will implement several fairness metrics, and evaluate the bias in a simple logistic regression model trained on *tf-idf* features.
- In the second section, you will analyze the bias present in a BERT model that has been pretrained on clinical notes, in three different ways.
- We will be providing GCP credit to work on this problem set and others. For details, see Piazza.

- If you use code from an external source like Stack Overflow, you must note this as a python comment in your code. Otherwise, all code must be your own.
- Your code should be *readable*, meaning it should contain informative comments and appropriate variable names. If your code is not giving the correct output, and the TA is unable to debug it, you will receive a poor grade.
- We will be using an autograder to grade some of your functions. To ensure that your code is compatible with this, please adhere to the following rules:
  - Do not rename any of the functions or variables in the template code.
  - Ensure that all your functions depend *only* on the variables that are passed in, and do not reference any variables in the global context. Calls to helper functions and functions from imported libraries are okay.
  - Do not remove any of the cells that are in the template notebook.
  - You are welcome to add any helper functions.
  - You are welcome to add additional cells and code anywhere in the notebook. However, they must not give any errors when ran.
- You are allowed to use any function from the following external libraries in your code. No other libraries are allowed.
  - Numpy
  - Scipy
  - Pandas
  - Scikit-Learn
  - NLTK
  - PyTorch
  - Transformers [1]
- **Do not** post solutions to this problem set to a public GitHub repository.

[1] https://github.com/huggingface/transformers

- Before starting this assignment, ensure you have access to MIMIC-III, either locally, or through GCP. If you do not, see the instructions for Homework 1.
- Before starting this assignment, familiarize yourself with the following concepts:
  - Group fairness definitions (e.g.[2])
  - BERT[3]
  - Clinical BERT[4]
  - Log-probability bias scores[5]

[2] A. Beutel, J. Chen, Z. Zhao, *et al.*, "Data decisions and theoretical implications when adversarially learning fair representations," *CoRR*, vol. abs/1707.00075, 2017. arXiv: 1707.00075. [Online]. Available: http://arxiv.org/abs/1707.00075.

[3] J. Devlin, M. Chang, K. Lee, *et al.*, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. arXiv: 1810.04805. [Online]. Available: http://arxiv.org/abs/1810.04805.

[4] E. Alsentzer, J. Murphy, W. Boag, *et al.*, "Publicly available clinical BERT embeddings," in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 72–78. DOI: 10.18653/v1/W19-1909. [Online]. Available: https://www.aclweb.org/anthology/W19-1909.

[5] K. Kurita, N. Vyas, A. Pareek, *et al.*, "Measuring bias in contextualized word representations," *CoRR*, vol. abs/1906.07337, 2019. arXiv: 1906.07337. [Online]. Available: http://arxiv.org/abs/1906.07337.

Section 1 - Exploratory Data Analysis and Warmup (23pts)

1. [**0pts**] **Preliminiaries.**

   (a) Download and unzip the template code: local, BigQuery

   (b) In `create_data.py` (or `create_data_bigquery.ipynb` in Colab), update the database access credentials with the credentials for your database.

   (c) Run the `create_data.py` script (or `create_data_bigquery.ipynb` in Colab), it will output `cohort.h5` and `notes.h5` in a `mimic_data` folder. The script should take no more than 20 minutes to execute.

   (d) A detailed description of each of the files is as follows:
   `cohort.h5`:

   - Contains one record for each adult patient's first ICU stay over 48 hours in length within their first hospital admission.
   - The `mort_icu` column represents whether the patient died during their ICU stay.
   - The columns from *Acute Renal* to *Shock* correspond to each of the 25 CCS code groups, which are derived from ICD-9 codes assigned at the end of a patient's hospital stay.

     > *These definitions group ICD-9 billing and diagnostic codes into mutually exclusive, largely homogeneous disease categories, reducing some of the noise, redundancy, and ambiguity in the original ICD-9 codes. HCUP CCS code groups are used for reporting to state and national agencies, so they constitute sensible phenotype labels.*[6]

     The column names are shorthands for the group names. For the full name, see `mapping.csv`.
   - The *Any Acute* and *Any Chronic* columns are derived from whether the patient has *any* acute and chronic phenotypes respectively.

   `notes.h5`:

   - Contains, for each of the patients in `cohort`, all of the notes written during their hospital stay (along with the timestamp) for the following note types:
     - Discharge Summary
     - Nursing
     - Nursing/other
   - Note that discharge summaries only have a date stamp, but no timestamp.
   - The notes have been lightly preprocessed (ex: removing PHI identifiers, lower case).

[6] H. Harutyunyan, H. Khachatrian, D. C. Kale, *et al.*, "Multitask learning and benchmarking with clinical time series data," *Scientific data*, vol. 6, no. 1, pp. 1–18, 2019.

2. [**5pts**] **Data Exploration.** Using the `s1.ipynb` notebook, answer the following questions in your report.

   (a) [**2pts**] What is the distribution of the cohort (in percentages) for gender, ethnicity, language, and insurance?

   (b) [**1pt**] What is the distribution of the cohort (in percentages) for the intersection of gender and ethnicity?

   (c) [**2pts**] Explain, using one sentence each, what the following terms mean in an insurance context: *Medicare*, *private*, *Medicaid*, *self-pay*.

3. [**5pts**] **Fairness Metrics.**

   (a) [**3pts**] In `metrics.py`, implement the function `gap_metrics`. Ensure your functions pass the `binary_test` before continuing.

   (b) [**1pt**] Explain why the parity gap is almost never a useful metric in a healthcare context. What assumption would have to hold for the parity gap to become useful?

   (c) [**1pt**] Explain why individual fairness might be tricky to define in a healthcare setting.

4. [**13pts**] **TF-IDF Model for ICU Mortality.** In this question, we will build a logistic regression model for ICU mortality on TF-IDF note representations. We will then evaluate its fairness with regards to gender.

   (a) [**4pts**] Process the data in the following way:

      - For each of the patients in the cohort, select their first 48 hours of notes after their `intime`, excluding discharge summaries. Concatenate the notes into a single string.
      - Drop all patients who do not have any notes within the first 48 hours.
      - Report the following:
        - The number of patients remaining in the cohort.
        - The average length (in characters) of the notes (for those that remain).

   (b) [**1pt**] Report the average length (in characters) of the notes for men versus women. Use an appropriate statistical test to determine if there is a significant difference.

   (c) [**1pt**] Report the prevalence of `mort_icu` for men versus women.

   (d) [**3pts**] Encode the concatenated notes using TF-IDF, and train a logistic regression model on these features to predict ICU mortality. Feel free to re-use code from Homework 1 for this question only. Your model should achieve at least 85% test AUROC. Report your model AUROC and AUPRC on the test set.

   (e) [**4pts**] Calculate the fairness metrics implemented in Question 3 for your model, using 1,000 bootstrapped samples of the test set, with gender as the protected variable. Use a decision threshold of 0.3. Answer the following questions:

- Report each of the following (with 95% CIs): parity gap, specificity gap, recall gap. Use the convention that a positive gap denotes better performance in males.
- Looking at the recall gap, which gender does it favor? Is this gap statistically significant?

Section 2 - Biases in Clinical BERT (37+1pts)

1. **[0pts] Preliminiaries.** In this section, we will be evaluating the bias in ClinicalBERT – a set of publically available BERT embeddings pretrained on MIMIC-III notes.
   (a) Download the ClinicalBERT[7] weights from this repository.
   (b) Extract the `biobert_pretrain_output_all_notes_150000` directory – this is the model we will be analyzing for this section.
   (c) In `s2.ipynb`, update the `bert_path` variable to be the extracted folder.

2. **[7pts] Sentence Completion.** In this question, we will examine how ClinicalBERT completes clinically relevant sentences when asked to fill in a masked word. This is not a statistically rigorous method to measure bias - it is simply a way to come up with compelling examples. We will examine two much more rigorous approaches below.
   (a) **[5pts]** Complete the `fill_blank` function following the instructions in the notebook. Ensure your function passes the provided test before continuing.
   (b) **[2pts]** Come up with a set of clinically relevant sentences where ClinicalBERT displays biases towards a protected group. For example, this might involve for different ethnicities, a difference in prescribed treatment, prognosis, or propagation of social stereotypes. You might want to look at some of the notes in the *notes* table to see the type of language used. One example set is shown below:

[CLS] 40 yo *black* homeless man with h/o polysubstance abuse and recently released from <u>prison</u> [SEP]
[CLS] 40 yo *asian* homeless man with h/o polysubstance abuse and recently released from <u>home</u> [SEP]
[CLS] 40 yo *hispanic* homeless man with h/o polysubstance abuse and recently released from <u>home</u> [SEP]

3. **[9pts] Log-Probability Scores.** In this question, you will be implementing the algorithm to evaluate bias in contexual word embeddings proposed by Kurita et al[8]. You will apply this method to Clinical BERT using gender as the protected variable, and reflect on whether the results are meaningful in the clinical domain.
   (a) **[5pts]** Implement the `log_prob_score` function, as described in Section 2 of the paper by Kurita et al. Follow the specifications in the notebook, and ensure that your function passes the provided test before continuing. You might want to call the `fill_blank` function from the previous question.

[7] E. Alsentzer, J. Murphy, W. Boag, *et al.*, "Publicly available clinical BERT embeddings," in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 72–78. DOI: 10.18653/v1/W19-1909. [Online]. Available: https://www.aclweb.org/anthology/W19-1909.

[8] K. Kurita, N. Vyas, A. Pareek, *et al.*, "Measuring bias in contextualized word representations," *CoRR*, vol. abs/1906.07337, 2019. arXiv: 1906.07337. [Online]. Available: http://arxiv.org/abs/1906.07337.

(b) **[3pts]** Complete the `log_cats` variable by filling in clinical terms for the category related to mental health. Run the provided script to generate a summary table. Include this table in your report. Briefly interpret the results.

(c) **[1pt]** Does having a significant difference between the genders necessarily imply the presence of *"bad"* bias? Explain your answer.

4. [**4pts**] **Multi-Group Fairness Metric.**

(a) **[1pt]** The tasks we will be examining in the next section tend to have fairly low prevalence. Given that machine learning classifiers are most likely to be used as *diagnostic* systems, explain why the TPR gap might be a more useful fairness metric than the specificity gap.

(b) **[3pt]** Implement the `multigroup_tpr_gap` function in `metrics.py`, using the following method:

- Given a task with protected groups $z = \{z_1, ..., z_k\}$ each with TPR $\{t_1, ..., t_k\}$
- Calculate g_j, the TPR gap for class j, as follows:

$$i^* = arg \max_{i \in z} |t_j - t_i|$$
$$g_j = t_j - t_{i^*}$$

Ensure your function passes the provided test before continuing.

5. [**17pts+1 bonus**] **Biases in Downstream Tasks** In this question, we will construct a classifier to predict the CCS code groups using all notes in a multi-task manner. To build this classifier, we will be using a temporal model on static note embeddings extracted from clinical BERT[9]. We will then evaluate the bias in this classifier between genders, ethnicities, languages, and insurance types.

(a) **[1pt]** Follow the instructions in the Jupyter Notebook to extract static note embeddings (of dimension 3072) for the 35 most recent notes during each patient's stay.[10] Report the value requested in the notebook to gain the point for this question.

(b) **[8pts+1 bonus]** Build a temporal model[11] that takes as input all of the note embeddings for a particular patient during their hospital stay, and simultaneously outputs a vector of length 27, with each element corresponding to a classification

[9] Generally, training BERT on a downstream task involves finetuning over the entire BERT model. However, this is computationally intensive, and so we will be opting for static BERT representations in this assignment.

[10] Since BERT can only take inputs of size 512, our note embeddings are only based off of the first ∼510 tokens in each note. You might think of more elegant ways to integrate information in the entire note, but that is not required for this assignment.

[11] such as an LSTM or a CNN

target. After building the model, run the included code to generate a table summarizing your model performance per task. Include this table in your report. Five of the points in this section will be based on your code, model architecture, and report table. The remaining points will be assigned as shown in Table 1.

(c) **[2pts]** Use the code provided to calculate the bootstrapped multi-group fairness gaps for gender, ethnicity, language, insurance, and the intersection of gender and ethnicity. For each group, the code calculates the number of statistically significant gaps (out of 27 tasks), as well as, out of the statistically significant gaps, how many of them *favor* each group. Report each of the five tables generated.

(d) **[2pts]** For each protected variable, interpret your results from the previous table – which group(s) does the classifier favor? Which groups(s) does the classifier disfavor?

(e) **[1pt]** For one of the groups that are disfavored, give a conjecture based on epidemiology or societal norms for why they might be disfavored.

(f) **[2pts]** Describe two sources of bias that might responsible for these performance gaps.

(g) **[1pt]** Describe a method that you could use to "debias" the classifier we just created (i.e. bring the TPR gaps closer to zero). If this is a published method, provide a reference.

Average AUROC	**# Points**
Below 73%	0
73-76%	1
76-78%	2
78-80%	3
80%+	3+1

Table 1: Distribution of points for 5b