CSE474/574 Introduction to Machine Learning
Programming Assignment 3

Fairness in Machine Learning

Role: Volunteer for Non-Governmental Organization

Due Date: **May 7th 2021** by 11.59 EST
Maximum Score: 100

Note In this assignment your group takes the role of machine learning engineers who have been tasked with designing a new system for use in U.S. court systems. You are given 3 different ML models and must apply 5 postprocessing techniques onto each of them. Finally, your group must determine a single model and technique to submit, which will then be measured against the rest of the class for a chance at extra credit.

Submission You are required to submit a single file called ***pa3.zip*** using UBLearns. One submission per group is required. File ***pa3.zip*** must contain 3 files: ***report.pdf***, ***Postprocessing.py***, and ***marketsubmission_<classsection>_<groupnumber>.py***. So if you are a CSE474 student in group #22, you would submit ***marketsubmission_474_22.py***.

- Submit your report in a pdf format. Please indicate the **team members**, **group number**, and your **course number** on the top of the report.
- The file ***Postprocessing.py*** should contain all implemented functions. Please do not change the name of the file.
- The file ***marketsubmission_<classsection>_<groupnumber>.py*** should contain your market submission model. Please do not change the name of the file.

Please make sure that your group is enrolled in the UBLearns system: You should submit one solution per group through the groups page. *If you want to change the group, contact the instructors.*

This assignment has two main parts: implementation and evaluation. You will first implement the postprocessing functions in Part I as discussed in Section 2.2. We will run the unit tests to test the correctness of the functions for Part I. For Parts II and III, we will evaluate your report. Please note that if your methods for Part I are not implemented correctly, you will automatically lose all points for Part II. You will have to provide the *market submission* model if you want to be considered for the extra credit as part of the competition described in Section 6.

1 Setup

In 2016, the independent non-profit news organization ProPublica released a report evaluating Northpointe's COMPAS system, which is an algorithm widely used across the country for considerations in pretrial detention and sentence determination.

- Propublica Story

- Northpointe's Response

COMPAS evaluates criminals on over 100 factors and outputs a risk score which indicates how likely it is that someone will recidivate (go on to commit another crime in the future). These scores are then taken into consideration by judges when assigning sentences, determining bail/parole eligibility, etc. Critically, race is not one of the factors used in by COMPAS.

ProPublica reviewed the output of COMPAS on a dataset of over 7000 individuals from Broward County, Florida. They found that the algorithm correctly predicted recidivism at similar rates for both white (59%) and black defendants (63%). However, when the algorithm was incorrect it tended to skew very differently for each of these groups. White defendants who re-offended within two years were mistakenly labelled low-risk almost twice as often as their black counterparts. Additionally, black defendants who did not recidivate were rated as high-risk at twice the rate of comparable white defendants.

Northpointe submitted a rebuttal of ProPublica's evaluation, claiming that PP misrepresented certain statistical values. They assert that their model is entirely fair across racial lines when base rate recidivism levels are taken into account.

On the heels of these reports your company has decided to release a new system as a potential candidate for replacing COMPAS. 3 machine learning models have been designed by your development team and trained on the Broward County data. These include:

1. A linear support vector machine based regression model (SVM)
2. A feed forward neural network (NN)
3. A naive Bayes classifier (NBC)

NOTE: All of the above models are setup to produce a real-valued prediction between 0 and 1 that indicates the probability of the input subject to commit a crime in future (recidivism).

In addition, your research team has been scanning machine learning research papers and have determined 5 potential post-processing methods that enforce various constraints in attempts to reflect different measures of fairness. These include:

1. Maximum profit / maximum accuracy
2. Single threshold
3. Predictive parity
4. Demographic parity
5. Equal Opportunity

As the lead development team of your company it falls upon your group to implement these 5 methods on each model. You must then determine which model/fairness combination to put forward as your finished product. Since this is an important social issue, you are certain that rival companies and socially inclined NGOs will also be releasing similar systems. You need to be sure to sufficiently justify all algorithmic/ethical considerations about your model so that it can withstand the scrutiny of the public eye and ultimately be chosen as the replacement for COMPAS.

1.1 Your Role

You will approach this assignment as **volunteers of a humanitarian NGO (non-governmental organization)**. Your NGO is a non-profit, and thus you do not receive a paycheck and are not reliant on this work for money to survive. Your group's mission revolves around advocating for the fair treatment of individuals within the U.S. criminal justice system. To this end, societal considerations and positive reform are your main concerns. However, it will be difficult to get various local and state governments to adopt your software if you completely disregard financial concerns.

2 Coding Tasks

2.1 Models and Data

As listed above, you are given 3 working machine learning models: a linear support vector regressor (`Compas_SVM.py`), a feed-forward neural network (`Compas_NN.py`), and a naive Bayes classifier (`Compas_Naive_Bayes.py`). Each of these models has a single function which loads the data, builds and runs the model, and outputs a set of predictions alongside the rest of the useful data needed for evaluation. The data consists of 4 files: `Compas_train_data.npy`, `Compas_train_labels.npy`, `Compas_test_data.npy`, and `Compas_test_labels.npy`, all of which are contained within `Broward_Data.zip`. All files provided to you should be extracted/downloaded into the same directory for them to run properly.

Running one of the model files will load the data, classify the data, separate all relevant predictions and labels into groups based on race, and call the `report_results` function from `Report_Results.py`. This function calls each of the 5 post-processing methods (which you must complete) and prints out some useful metrics that can be used to determine if your functions are working correctly. The `report_results` function will be used by the graders on these models in order to determine the accuracy of your final solution.

Additionally, you are provided with `utils.py`, a collection of functions used to gather various metrics from the classified data. You are welcome to write your own functions, but everything you need for completing the postprocessing methods is provided within this file.

2.2 Postprocessing

The implementation section requires you to finish methods for 5 different postprocessing techniques, which can be found in `Postprocessing.py`. The function headers have all been written for you, but your team must complete the function bodies. Each function takes `categorical_results` as an input. This is a dictionary - the keys are the 5 different racial groups, and the entry for each key is a list of `(prediction, label)` tuples for all people within that racial group.

Some functions also take ϵ as an additional parameter, which acts as a tolerance for enforcing some of the fairness constraints. The specific ϵ values you need to satisfy are included alongside each specific function.

> NOTE: The listed ϵ values must be consistently upheld for the SVM and Naive Bayes Classifier. The neural network may occasionally fail or provide trivial solutions using these values.

While COMPAS gives a recidivist rating of 1-10, with additional labels of low, medium, and high risk, the models you have been given for this assignment all output a list of real-valued predictions between [0, 1]. The postprocessing methods you must implement will determine threshold values for these predictions; anything above the value will be a 1, and anything below the value will be a 0. This will give a final result of a list of binary decisions. Some functions will require you to find different threshold values for each racial group.

These functions must return the two things: the classified data which has been thresholded to meet the requirements of the function, and the thresholds themselves. The output data should be in the same form as the input data. The thresholds should also be a dictionary with racial groups as the keys, but the entries will be the scalar threshold value for that racial group.

Definitions of necessary terms and a few basic pros/cons of each method are provided in the supplementary handout - `primer.pdf`. You have to implement the following functions:

- **Maximum Profit/Maximum Accuracy/Minimal Loss** [**10 Points**]: The maximizing method produces the best possible results from the given classifier. "Best" is context-specific; it can refer to metrics such as the highest accuracy, the maximum profit the minimal loss. This function requires finding thresholds for each racial group to maximize your chosen secondary optimization criteria (either cost or accuracy, see Section 2.3).

- **Single Threshold** [**10 Points**]: Arguably the simplest of the thresholding methods, single threshold holds all groups to the same decision standard by using one threshold value for all decisions.

- **Predictive Parity** [**10 Points**]: A classifier that enforces predictive parity will have the same positive predictive value (PPV) for each group. According to Northpointe, the COMPAS system satisfies predictive parity across racial groups when base rates of recidivism are accounted for. Your method must enforce predictive parity within an ϵ tolerance of ±1%.
- **Demographic Parity** [**10 Points**]: Demographic parity means that each group has an equal proportion of members classified in each way. Since our system is binary, this means that each group must have the same percentage of people who are classified as recidivists. Your function must be able to enforce demographic parity within an ϵ tolerance of ±2%.
- **Equal Opportunity** [**10 Points**]: Equal opportunity refers to all groups having the same opportunity to achieve the "privileged" outcome. This terminology can be somewhat misleading, as in our case it refers to being labelled as a recidivist. Enforcing equal opportunity means that all groups will have the same true positive rate (TPR). Your functions must be able to enforce equal opportunity within an ϵ tolerance of ±1%.

To be clear, you do not need to implement each of these methods 3 times. If you correctly implement the functions in `Postprocessing.py` they should work for all 3 of the provided models.

2.3 Secondary Optimization

Many of these constraints can be satisfied using many different threshold values. For example, a single-threshold system can be achieved by choosing any number between 0 and 1 as a threshold. Of course, your group wants the very best results, and those can only be achieved by performing a secondary optimization.

Two metrics can be used for this secondary optimization: accuracy and financial cost. For each one of the 5 post-processing methods you must return a set of thresholds that both satisfy the necessary constraint and also return the best possible overall value for one of these two metrics. You are welcome to choose either one, but you must use the same secondary optimization metric for all 5 functions. You must list the metric you have chosen in the space provided at the top of `Postprocessing.py`.

3 Evaluation

3.1 Report Essentials

After you have implemented the 5 postprocessing methods, you must choose a single model and postprocessing method as your group's market submission. More details about this aspect of the project are provided in Section 5.

As explained in the supplementary handout, there is no definitive answer to what constitutes fairness in machine learning. Therefore, you must provide a detailed report justifying the model you choose to submit. The report must be no longer than 2 pages. There is no minimum length, as long as you have fulfilled the criteria listed below. The beginning of this report MUST include your model choice, algorithm choice, secondary optimization criteria (cost or accuracy), the overall cost of your system's choices to society, and your overall accuracy.

REPORT 1.

The following questions are necessary for the body of report:

1. What is the motivation for creating a new model to replace COMPAS? What problem are you trying to address? [**8 Points**]
2. Who are the stakeholders in this situation? [**8 Points**]
3. What biases might exist in this situation? Are there biases present in the data? Are there biases present in the algorithms? [**8 Points**]

4. What is the impact of your proposed solution? [**8 Points**]

5. Why do you believe that your proposed solution a better choice than the alternatives? Are there any metrics (TPR, FPR, PPV, etc?) where your model shows significant disparity across racial lines? How do you justify this? [**8 Points**]

Additionally, **5 Points** of the report will come from professionalism. You must reference at least one outside source in support of your argument. This can be either one that we've provided, or one you've found on your own. Do not just rely on the pros/cons provided in the fairness primer.

Finally, the remaining **5 Points** of the report will come from ethical consistency. Does your argument make sense, and do you show that you have a strong understanding of the case you're presenting? No stance is inherently wrong, but you must be able to justify your position without contradicting yourself.

3.2 Report Extra Credit

Answering these questions constitutes the bare minimum that you should provide. **5 Points** extra credit will be provided to the group that provides the best justification, se we strongly suggest that you include additional information and references to papers that support your choice. We have provided an incomplete list of extra topics to potentially include. You are also welcome to provide any other information or arguments that you consider relevant.

- How do you justify valuing one metric over the other as constituting "fairness"?
- What assumptions are made in the way we have presented the assignment? Are certain answers presupposed by the way we have phrased the questions?
- In what ways do these simplifications not accurately reflect the real world?
- How do uncertainty and risk tolerance factor into your decision?
- To what extent should base rates of criminality / recidivism among different groups be factored into your decision?
- The tools we provide can split the predictions into different protected categories, such as by age or gender. What disparities arise in these groups? How do these disparities compare to those shown when the predictions are split by race?

4 Financials

A system such as this has far-reaching social implications for the real world. One such consequence to be considered is the monetary cost of the decisions provided by your model. Each type of classification carries a unique cost or benefit to society as a whole. These values are detailed below:

True Positives: -$60,076 per person per year.

True positives reflect people who were correctly predicted to recidivate. The cost for this type of prediction is the cost of keeping one inmate housed in jail for one year in New York State. This value includes the cost of living and upkeep for an inmate and the fractional salary and benefits of the corrections officers needed to oversee the inmates.

Source: price of prisons

True Negatives: +$23,088 per person per year.

True negatives reflect people who were correctly predicted not to re-offend. This is actually a positive value, and is determined by the annual pre-tax earnings of a full-time worker making minimum wage (2019) in NYS. While this value is somewhat simplified and doesn't necessarily represent all of the social value created by a functioning member of society, it is often used as the baseline in criminology research. *Source*: minimum wage

False Positives: -$110,076 per person per year.

False positives represent people who were incorrectly predicted to recidivate. While this prediction may not directly lead to a jail sentence, there are many U.S. districts that rely very heavily on systems such as COMPAS. A false positive could thus potentially represent a wrongfully imprisoned person, and this value reflects the cost of housing one person in jail plus the compensation cost for wrongful imprisonment. Many states have laws and precedents that set an expected value to be collected by a lawsuit brought forth from wrongful imprisonment. NYS doesn't actually have a limit on this amount, but the most common value from other states is $50,000 per year.

Source: compensation

False Negatives: -$202,330 per person per year.

False negatives represent people who were incorrectly predicted to not recidivate and then went on to commit another crime. Our value represents the average annual cost of a crime on society, a topic which is understandably highly contentious in legal literature. Different crimes obviously have different associated values, but since this system is being deployed in the real world it is a reasonable assumption that we won't know what type of crime was committed beforehand.

A baseline for this value was obtained from the source listed below, but additional factors were included to reflect the nature of this assignment. We adjusted their calculations for inflation, and slightly skewed them upwards to represent the fact that we are calculating crime cost per year. Non-violent crimes are significantly more likely to happen, meaning that there is a much higher chance of a single offender committing multiple crimes per year before getting arrested.

Source: cost of crime

These values will be applied to the results of your model to determine the total cost to society. This metric will be considered in the market evaluation, described in Section 5.

5 Market Evaluation Competition

Since fairness is a contested subject within machine learning, we have no authority to proclaim that one of these solutions is objectively better than the others. Oftentimes in society the moral choice at any given moment is determined by the popular voice, for better or worse. While it is up to you to justify why your solution is better than the alternatives, the decisions of the whole class will determine what constitutes the "best" fairness metric in the context of a market game termed Vox Populi.

The basic idea is that a method will only qualify as "fair" if it obtains a certain percentage of approval from the total class. Methods falling below this threshold percentage will be considered "unfair" and subsequently disregarded.

In the first round of the competition the popularity of each postprocessing algorithm will be calculated as a percentage of the total submissions. Any function used by 20% or more submissions will be considered a de facto standard for fairness. The remaining techniques will be deemed unfair by our mock society, and any submission that uses them will be not be considered for the next round of the competition.

The second round will look at the secondary optimization metrics. All fair models will be checked to see whether accuracy or profit was considered to be the more important secondary optimization metric overall. This chosen metric will then be used to determine the best performing model of the class.

For this competition you are allowed to freely change the model you decided to use. This includes adjusting any hyper-parameters, seeds, kernels, or network structures. If you really want to you can build your own model, as long as it matches the criteria listed below. You are also free to change the data metrics used by your model, which is determined by the metric list provided to the preprocess function. However, you must use the same secondary optimization method that you declared in your report. This model will be tested on the standardized dataset that has been provided to you, and we must be able to recreate your results from the submission for your group to qualify for the competition. Additionally, all tolerances (described in part 2) must still be met for your submission to qualify.

Your model should be a .py file and must adhere to the following naming procedure: `marketsubmission_<classsection>_<groupnumber>.py`, with no brackets. So if you are a CSE474 student in group #22, you would submit `marketsubmission_474_22.py`.

When placed in the same directory as the data, your model file should act similarly to the models provided to you. It must call the preprocessing function with a metric list, load the data, build the model, make predictions, group those predictions by race, apply the postprocessing method of your choice, and print all relevant associated metrics and thresholds. You only need to print the results of your chosen postprocessing method, and thus you should not call `report_results` in your submitted model. You can use pieces of that code if you find it convenient. Any model that does not print its final results or whose results do not match those presented in the report will not be considered for the market competition.

Keep in mind that not all groups in the class have been assigned the same role. Some of you will be approaching this problem from very different perspectives, which will certainly affect which models and algorithms you choose. Open discussion and debate about what qualifies as fairness is encouraged on Piazza. However, conspiring as a class in an attempt to game the system (such as by all agreeing to use the same postprocessing algorithm) will not be tolerated and will result in the cancellation of any applicable extra credit points for this project.

Winning the market evaluation will result in **15 Points** extra credit for first place, **10 Points** extra credit for second place, and **5 Points** extra credit for third place. In the event of a tie for any position the winners will be determined by the strength of the argument provided in their report.

6 Grading Overview

The high-level grading scheme is described below.

6.1 Implementation

The 5 post-processing techniques described in Section 2 are worth 10pts each.

- Your solutions must solve the optimization problems presented; hard-coding thresholds will result in no points.
- All tolerances must be observed where specified.
- Your methods will be applied as the post-processing steps of a standard model and data set, and there are definitive optimal numerical answers that should be provided by each function.
- Each method must perform secondary optimization using either accuracy or financial cost, and you must specify which metric you have chosen at the beginning of your report. Not choosing a metric or switching metrics between functions will result in a loss of points!

6.2 Evaluation

Each of the 5 main questions in Section 3 are worth 8pts.

- Report MUST include your model choice, postprocessing algorithm choice, secondary optimization criteria (cost or accuracy), the total cost of your system's choices to society, and your overall accuracy. Reports without these criteria will not be graded.
- Answers should be well-thought out and reflect the results of a serious exploration of the problem and its various aspects.
- Single word or single sentence answers will not be accepted.
- All answers must be compounded into a coherent report justifying the choices made by your group.

- Satisfactory reports must not only be factually accurate, but also grammatically correct and persuasive.
- Additional 10pts come from professionalism and consistency – refer to Section 3 for more details.

6.3 Extra Credit

- The group with the best report will be awarded an extra 5pts.
- The winners of the market evaluation competition will be awarded an extra 15pts, 10pts, or 5pts for first, second, and third place respectively.

7 Libraries

- numpy
- scikit-learn
- matplotlib
- tensorflow 1.15.0
- keras 2.2.4