Problem Set 6: Correcting Bias in Classification

Feedback

Students can provide feedback here.

Background

The datasets we use to train machine learning models can often encode human biases. From a social and ethical standpoint, we want to remove or minimize this bias so that our models are not perpetuating harmful stereotypes or injustices. From a business and legal perspective, we want to produce effective models that adhere to industry standards of fairness.

There are several ways that we can tackle this problem, including pre-processing the data to remove bias before training, in-processing the model to change the way it learns from the data, and post-processing the results to correct for bias. In this assignment, we will be introducing two methods for correcting for bias: a post-processing method that uses alternative classification thresholds for different groups, and an in-processing method to train a logistic regression classifier that maximizes fairness while maintaining a certain level of accuracy. The in-processing method for correcting bias is based on Fairness Constraints: Mechanisms for Fair Classification by Zafar et al. (2017).

Agenda

- Introduction
- Part 1: Defining fairness
- Part 2: Observing a classifier's bias in our dataset
- Part 3: Post-processing with alternative thresholds
- Part 4: In-processing with fairness constraints
- Bonus question: Alternative fairness definitions

In [1]:
```
%%capture
import numpy as np
import pandas as pd
import generate_synthetic_data as generate
import classify_synthetic_data as classify
import matplotlib.pyplot as plt
%matplotlib inline
```

Introduction: Gender bias in hiring

Our dataset represents applicants for a job. We want an algorithm to help making hiring decisions, as this process is rife with human bias. However, studying past data would simply reinforce these biases (as we will see below).

Consider gender in hiring decisions. There is a well-documented gender pay gap: in general, people who identify as women are paid less for the same work than people who identify as men.

Now, imagine that we want to build a classifier that will make hiring decisions. If we train a classifier to make hiring decisions based on prior work experience and income, we would expect the classifier to *learn* a bias against women. Put another way, we can't make a classifier that is "gender-blind" by simply "throwing out" gender, excluding it from our dataset. Information about gender is already correlated with income.

We do not want our algorithm to learn this bias for gender. **Instead, we would like our algorithm to *correct* for this bias.**

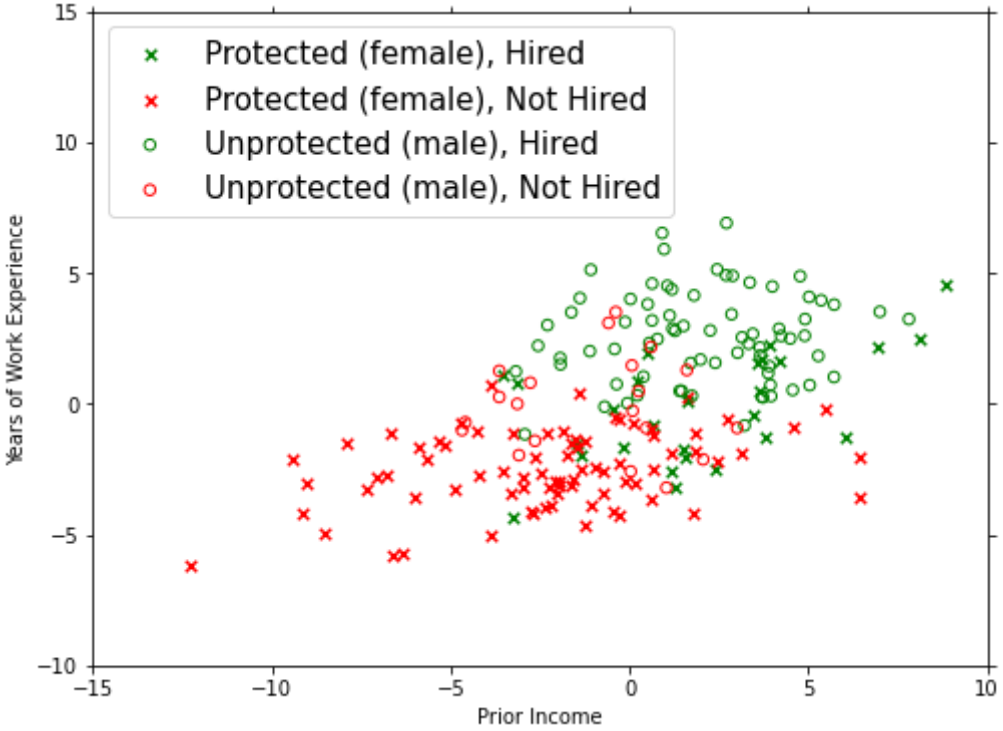
To discuss this issue more deeply, let's introduce some key terms:

- **Sensitive** features: A feature is *sensitive* when it may contribute to bias. For example, gender is a *sensitive* feature.
- **Protected** classes: A class is *protected* when we expect there is bias against it. For example, being female is a *protected* class in this example.

Let's generate a synthetic dataset that contains the prior income, work experience, and gender of 1,000 applicants who were hired and 1,000 applicants who were not. (For a more detailed explanation of how we generate this toy dataset, see the Appendix below.)

We represent each individual on the plot below according to their prior income and work experience (the non-sensitive features). The shape of the marker indicates the applicant's gender (the *sensitive* feature). Females, denoted by `x`s, are the *protected* class in this example. The color indicates whether or not they were hired.

In [2]:
```
X, y, x_sensitive = generate.generate_synthetic_data(plot_data=True)
```



The variables here are `X` (non-sensitive feature: years of prior work experience), `y` (label: hired/not-hired), and `x_sensitive` (male/female).

Here, we have effectively created a biased dataset where the sensitive feature *gender* is strongly related to the applicant's hiring status. A female applicant (`x`) seems less likely than a male applicant (`o`) to be hired. (Again, this is just example data---while gender gap in pay is real, this particular data is synthetic. See the appendix for details.)

Part 1: Defining Fairness

In order to remove bias from our model, we first need to define bias.

In this assignment, we will be using **disparate impact** (also known as the **p%-rule**) to measure the unfairness of a data set with respect to a particular sensitive feature. We can measure *how biased* our dataset is by calculating the p%-rule.

1.1 Calculate the p% rule

If we take $\mathbf{x}$ to represent a (non-sensitive) feature vector, y a 1/-1 class label, and z a 0/1 sensitive attribute, the p%-rule states that the ratio between the fraction of subjects assigned the positive decision outcome ($y = 1$) given that they have a sensitive attribute value and the fraction of subjects assigned the positive outcome given that they do not have that value should be no less than p%.

In the scenario we described in this lab, the non-sensitive feature vector $\mathbf{x}$ represents prior income and years of work experience, the class label y represents the hiring status (1 for hired, -1 for not hired), and the sensitive attribute z represents gender (0 for female, 1 for male). The

p%-rule would tell us the ratio between the fraction of female applicants who were hired and the fraction of male applicants who were hired.

$$p = 100 \left(\frac{Pr(y = 1|z = 0)}{Pr(y = 1|z = 1)} \right)$$

In [3]:
```
sensitive_features_arr  = np.array(x_sensitive['s1'])
```

In [4]:
```
# TODO: Write a function compute_p_rule that takes as inputs a sensitive_feature
# hiring outcomes), and returns the p percent rule. Also have your function prin
# in the protected and unprotected class, and the percent of observations in eac

def calculate_p_rule(sensitive_features_arr, y):

    # TODO: Replace 0 in the next line with a calculation of the # of unprotecte
    i = np.where(y == 1)
    s = []
    for j in i[0]:
        s.append(sensitive_features_arr[j])
    s = np.array(s)
    u = np.unique(s, return_counts = True)
    u0 = np.unique(sensitive_features_arr, return_counts = True)

    number_unprotected = (u[1][1])

    # TODO: Replace 0 in the next line with a calculation of the # of protected
    number_protected = (u[1][0])

    # TODO: Replace 0 in the next line with a calculation of the % of unprotecte
    percent_pos_unprotected = ((number_unprotected)/(len(i[0])))/(u0[1][1])

    # TODO: Replace 0 in the next line  with a calculation of the % of protected
    percent_pos_protected = ((number_protected)/(len(i[0])))/(u0[1][0])

    # TODO: Replace 0 with a calucalation of the p percent rule, using the previ
    p_percent_rule = (percent_pos_protected/percent_pos_unprotected)*100

    print('Number of protected observations: %i' % number_protected)
    print('Number of unprotected observations: %i' % number_unprotected)
    print('Protected in positive class: %i' % round(percent_pos_protected) + '%'
    print('Unprotected in positive class: %i' % round(percent_pos_unprotected) +
    print('P-rule: %i' % round(p_percent_rule) + '%')

    return p_percent_rule
```

Now test your function using the sensitive_features_arr and hiring outcomes (*y*) from our dataset. Verify that your function is correct: it should output p-percent rule of 29.45%. If you get a different number, go back and correct your code!

In [5]:
```
calculate_p_rule(sensitive_features_arr, y)
```

```
Number of protected observations: 255
Number of unprotected observations: 745
Protected in positive class: 0%
```

```
Unprotected in positive class: 0%
P-rule: 29%
```

Out[5]:
```
29.45216169814266
```

QUESTION A: Interpret the output of this method to describe the disparate impact between male and female applicants in the training data.

Answer: Female applicants in this dataset were "hired" at 29% the rate relative to male applicants

QUESTION B: Is there evidence of bias in the training data?

Answer: There is evidence of bias in the training data, as evidence of no bias would result in a p-rule value of 1 or 100%.

Part 2: Machine learning without correcting for bias

Now let's turn back to our dataset. Let's see what happens when we *don't* correct for bias and train a machine learning model on our data.

First, we pre-process the data. We add an intercept column and split the data into train and test sets.

In [6]:
```
intercept = np.ones(X.shape[0]).reshape(X.shape[0], 1)
X = np.concatenate((intercept, X), axis = 1)
pd.DataFrame(X, columns=["intercept", "income", "experience"]).head() # to help
```

Out[6]:

	intercept	income	experience
0	1.0	-1.738503	-1.977414
1	1.0	-4.908148	-3.267064
2	1.0	3.006536	-0.925507
3	1.0	0.779730	2.473575
4	1.0	1.193537	-2.595904

In [7]:
```
# split the data into training and test using a 70/30 split
train_fold_size = 0.7
X_train, y_train, x_sensitive_train, X_test, y_test,  x_sensitive_test =\
    classify.split_into_train_test(X, y, x_sensitive, train_fold_size)
```

To set our baseline, we'll train a standard logistic regression classifier on our data and see how it performs. (See appendix for more on how this classifier training works).

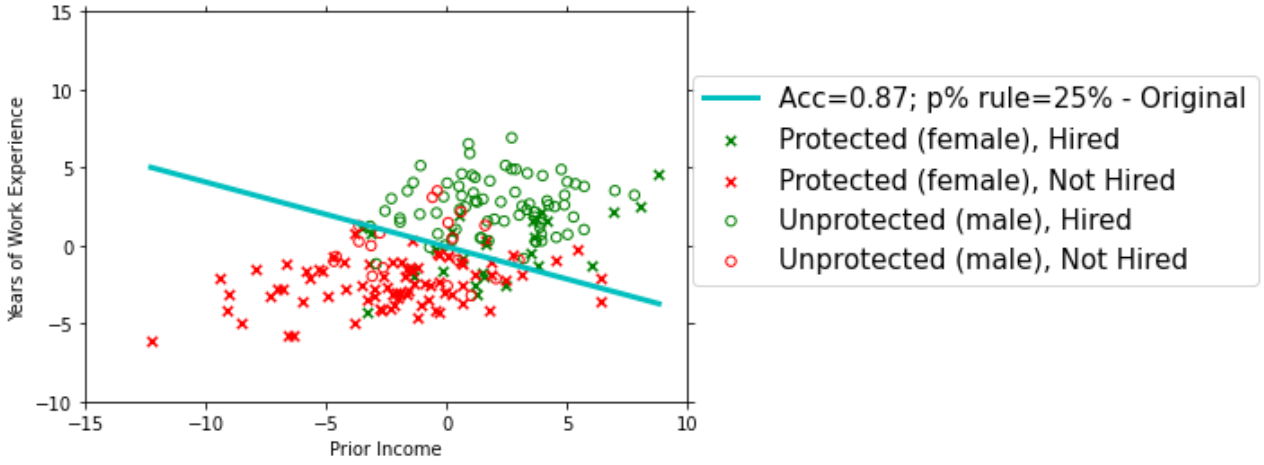
In [8]:
```
theta, p_rule, score, distances_from_decision_boundary = classify.train_test_cla
    X_train, y_train,  x_sensitive_train,
    X_test, y_test, x_sensitive_test,
    ['s1'],  # our list of sensitive features.
```

```
    apply_fairness_constraints=0 # We are NOT applying any fairness constraints
)
```

```
Accuracy: 0.87
Protected/non-protected in positive class: 86% / 22%
P-rule achieved: 25%
Covariance between sensitive feature and decision from distance boundary : 1.181
```

In [9]:

```
classify.plot_boundaries(X, y, np.array(x_sensitive['s1']), theta, p_rule[0], sc
```



The blue line represents the decision boundary for our standard logistic regression classifier. The model predicts points above this line as "hired" and points below as "not hired."

QUESTION A: Is our classifier accurate at predicting class labels (as determined by its accuracy rate)?

Answer: Our classifier is 87% accurate at predicting class labels.

QUESTION B: Is our classifier fair (as determined by its p% rule)? Explain. Hint: female applicants would be labeled "hired" by this classifier at what rate relative to male applicants?

Answer: Our classifier is not fair, as female applicants are labeled "hired" by our classifier at 25% the rate relative to male applicants. P-rule should be 100% to indicate fairness.

QUESTION C: How did the classifier learn bias for gender, even though gender was not included?

Answer: Our classifier learned bias for gender even though gender was not included from our income variable because there is a gender pay gap. This is an example of the removing sensitive features error.

Part 3: Post-processing with alternative thresholds

One option for correcting for this bias is postprocessing model outputs with a method called *thresholding*. This method makes no changes to the classifier itself, but instead processes classifier outputs differently for observations in the protected class and observations in the unprotected class. In the context of this problem set, we can think of thresholds as alternative

restrictions on distance from the decision boundary: observations without the sensitive characteristic (males) must be above the decision boundary in order to be hired, but observations with the sensitive characteristic (females) can be up to a distance of *k* below the decision boundary.

3.1 Applying one alternative threshold

First, use the distances_from_decision_boundary obtained earlier from the train_test_classifier function to test out a threshold of 0 for males and a threshold of -1 for females. Identify which observations in the test set will be hired according to these thresholds, and use the compute_p_rule function you wrote earlier to calculate the p% rule.

In [10]:
```
sensitive_feature_arr = np.array(x_sensitive_test['s1'])
```

In [11]:
```
sensitive_feature_arr = np.array(x_sensitive_test['s1'])
```

In [12]:
```
# TODO: Post-process the outputed distances_from_decision_boundary to apply a th
# protected observations (sensitive_feature = 1) and a threshold of -1 or over f
# (sensitive_feature = 0). Calculate the p% rule using these post-processed outp
df1 = np.concatenate((sensitive_feature_arr.reshape(-1, 1), distances_from_decis
df2 = pd.DataFrame(df1, columns=["isMale", "distance"])
df2.loc[(df2['isMale'] == 1) & (df2['distance'] > 0), 'isHired'] = 1
df2.loc[(df2['isMale'] == 1) & (df2['distance'] <= 0), 'isHired'] = -1
df2.loc[(df2['isMale'] == 0) & (df2['distance'] > -1), 'isHired'] = 1
df2.loc[(df2['isMale'] == 0) & (df2['distance'] <= -1), 'isHired'] = -1
calculate_p_rule(df2['isMale'],df2['isHired'])
```

```
Number of protected observations: 102
Number of unprotected observations: 229
Protected in positive class: 0%
Unprotected in positive class: 0%
P-rule: 35%
```

Out[12]:
```
35.234308805318385
```

QUESTION A: Is this decision method more or less fair than the naive decision method with equal thresholding implemented earlier? Why?

Answer: This decision method is slightly more fair than the naive decision method with equal thresholding implemented earlier, because we are allowing more values to be labelled as "hired" for our protected class in our sensitive feature (female). Since this is where the bias occurs, our p-rule becomes more fair, but it is clear that our dataset is still biased.

3.2 Varying the threshold

Now let's try out a set of different thresholds. Assume that we will leave the threshold for the unprotected class (males) at 0. Try out a set of thresholds for the protected class (females) between -8 and 0. Try out at least 20 thresholds in this range, and plot the results in a scatterplot, with the threshold on the x-axis and the p% rule on the y-axis. Calculate the

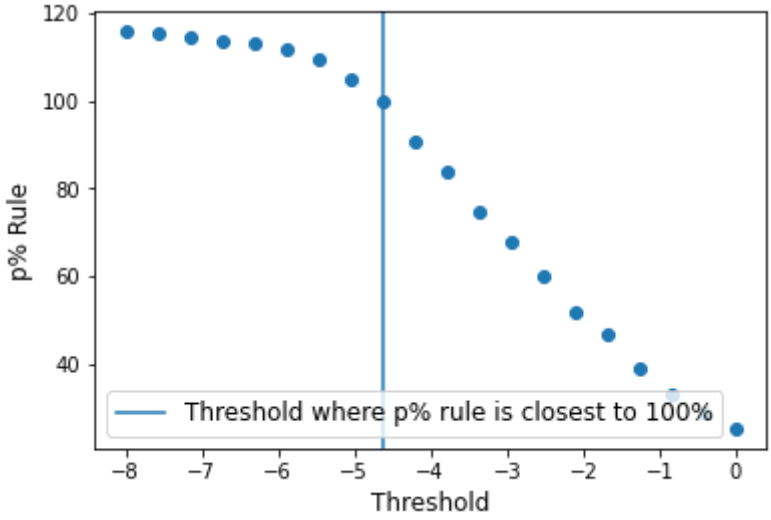
threshold at which the p% rule is closest to 100%, and add a vertical line to the scatterplot showing this threshold. Remember to make sure your plot is readable and well-labeled

In [13]:
```
# TODO: Post-process the outputed distances_from_decision_boundary to apply a th
# observations and a threshold of k for unprotected observations (sensitive_feat
# p% rule. Experiment with at least 20 different values of k between -8 and 0.

space = np.linspace(0, -8, num=20, endpoint=True)

pvals = []

for x in space:
    df3 = pd.DataFrame(df1, columns=["isMale", "distance"])
    df3.loc[(df3['isMale'] == 1) & (df3['distance'] > 0), 'isHired'] = 1
    df3.loc[(df3['isMale'] == 1) & (df3['distance'] <= 0), 'isHired'] = -1
    df3.loc[(df3['isMale'] == 0) & (df3['distance'] > x), 'isHired'] = 1
    df3.loc[(df3['isMale'] == 0) & (df3['distance'] <= x), 'isHired'] = -1
    pvals.append(calculate_p_rule(df3['isMale'],df3['isHired']))
dist = []
for i in pvals:
    dist.append(abs(100-i))
min_index = dist.index(min(dist))
plt.scatter(space, pvals)
plt.axvline(x=space[min_index], label = "Threshold where p% rule is closest to 1
plt.xlabel('Threshold', fontsize='large')
plt.ylabel('p% Rule', fontsize='large')
plt.legend(loc='best', fontsize='large')
plt.show()
```

```
Number of protected observations: 73
Number of unprotected observations: 229
Protected in positive class: 0%
Unprotected in positive class: 0%
P-rule: 25%
Number of protected observations: 84
Number of unprotected observations: 229
Protected in positive class: 0%
Unprotected in positive class: 0%
P-rule: 29%
Number of protected observations: 95
Number of unprotected observations: 229
Protected in positive class: 0%
Unprotected in positive class: 0%
P-rule: 33%
Number of protected observations: 113
Number of unprotected observations: 229
Protected in positive class: 0%
Unprotected in positive class: 0%
P-rule: 39%
Number of protected observations: 135
Number of unprotected observations: 229
Protected in positive class: 0%
Unprotected in positive class: 0%
P-rule: 47%
Number of protected observations: 150
Number of unprotected observations: 229
Protected in positive class: 0%
Unprotected in positive class: 0%
P-rule: 52%
Number of protected observations: 173
Number of unprotected observations: 229
```

```
Protected in positive class: 0%
Unprotected in positive class: 0%
P-rule: 60%
Number of protected observations: 196
Number of unprotected observations: 229
Protected in positive class: 0%
Unprotected in positive class: 0%
P-rule: 68%
Number of protected observations: 216
Number of unprotected observations: 229
Protected in positive class: 0%
Unprotected in positive class: 0%
P-rule: 75%
Number of protected observations: 243
Number of unprotected observations: 229
Protected in positive class: 0%
Unprotected in positive class: 0%
P-rule: 84%
Number of protected observations: 263
Number of unprotected observations: 229
Protected in positive class: 0%
Unprotected in positive class: 0%
P-rule: 91%
Number of protected observations: 289
Number of unprotected observations: 229
Protected in positive class: 0%
Unprotected in positive class: 0%
P-rule: 100%
Number of protected observations: 304
Number of unprotected observations: 229
Protected in positive class: 0%
Unprotected in positive class: 0%
P-rule: 105%
Number of protected observations: 317
Number of unprotected observations: 229
Protected in positive class: 0%
Unprotected in positive class: 0%
P-rule: 110%
Number of protected observations: 323
Number of unprotected observations: 229
Protected in positive class: 0%
Unprotected in positive class: 0%
P-rule: 112%
Number of protected observations: 327
Number of unprotected observations: 229
Protected in positive class: 0%
Unprotected in positive class: 0%
P-rule: 113%
Number of protected observations: 329
Number of unprotected observations: 229
Protected in positive class: 0%
Unprotected in positive class: 0%
P-rule: 114%
Number of protected observations: 331
Number of unprotected observations: 229
Protected in positive class: 0%
Unprotected in positive class: 0%
P-rule: 114%
Number of protected observations: 334
Number of unprotected observations: 229
Protected in positive class: 0%
Unprotected in positive class: 0%
P-rule: 115%
Number of protected observations: 335
Number of unprotected observations: 229
```

```
Protected in positive class: 0%
Unprotected in positive class: 0%
P-rule: 116%
```



Part 4: In-processing with fairness constraints

An alternative method for correcting for bias is applying fairness constraints in training the machine learning algorithm.

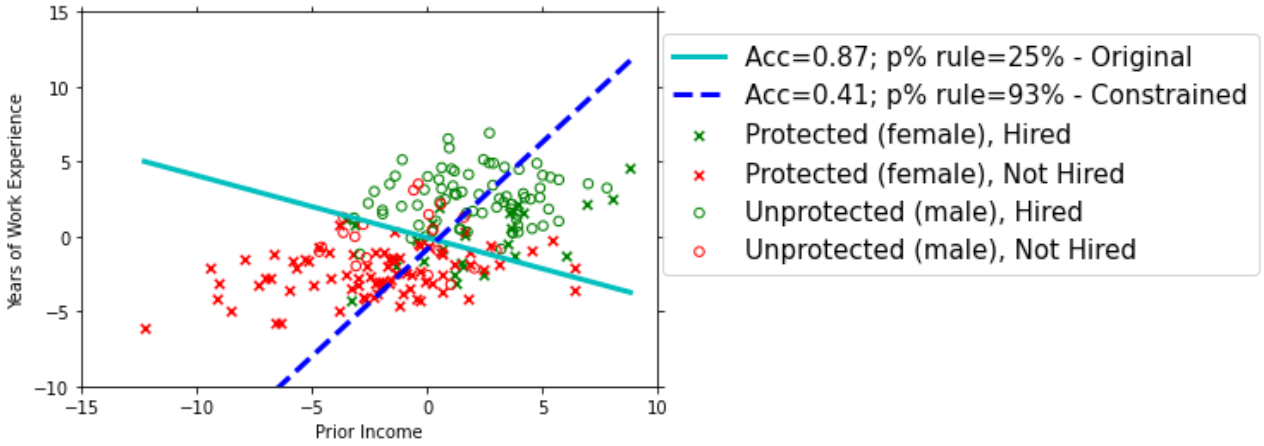
You can see the appendix for more information on how this process works (see Appendix: Applying fairness constraints. But it's best illustrated with an example.

Below, you're going to pick a **hyperparameter gamma** (γ). I've set a value of 1.5. See what happens.

In [14]:

```
theta1, p_rule1, score1, distances_from_decision_boundary = classify.train_test_
    X_train, y_train, x_sensitive_train,
    X_test, y_test, x_sensitive_test,
    ['s1'],  # our sensitive feature
    apply_accuracy_constraint=1, # applying our fairness constraint.
    gamma=1.5
    )

classify.plot_boundaries(X, y,
                         sensitive_features_arr,
                         theta,
                         p_rule[0],
                         score,
                         theta1,
                         p_rule1[0],
                         score1)
```

```
Accuracy: 0.41
Protected/non-protected in positive class: 60% / 56%
P-rule achieved: 93%
Covariance between sensitive feature and decision from distance boundary : 0.002
```



The solid line represents the decision boundary from our naive classifier---the one that scored poorly on the p% rule. That's the same line we saw above.

The *dashed* line represents our new decision boundary: the one from our **constrained** model, to which we applied our gamma hyperparameter.

After applying our bias correction method, the decision boundary *rotates* to produce a more fair distribution of class labels with respect to the sensitive feature.

At a gamma of 1.5, this method achieves a 93% p-rule, *but* a less accurate classifier. That's our tradeoff: we're effectively saying the original class labels are biased, and therefore wrong. So we're consciously making this tradeoff.

4.1 Fairness-accuracy trade-off

Using this method, we can specify how much we are willing to let the accuracy change by choosing an appropriate γ parameter for the model. The larger γ is, the more loss we are willing to incur in our corrected model compared to the baseline model. Again, refer to Appendix: Applying fairness constraints to get a better understanding of how this works.

Let's visually explore the *fairness-accuracy trade-off* by varying the value of gamma. Test out at least 20 values of gamma between 0 and 5. For each value, train a constrained model and record the accuracy and p-rule. Then, produce three scatterplots: The first should show gamma on the x-axis and accuracy on the y-axis. The second should show gamma on the x-axis and the p-rule on the y-axis. The third should show the p-rule on the x-axis and accuracy on the y-axis. Remember to make sure the plots are easy to read and well-labeled.

In [15]:
```
# TODO: experiment with gamma values between 0 and 3. Create plots of gamma vs.
# and p-rule vs. accuracy.

gammas = np.linspace(0, 5, num=20, endpoint = True)
accuracies = []
pvals = []

for i in gammas:
    theta1, p_rule1, score1, distances_from_decision_boundary = classify.train_t
        X_train, y_train, x_sensitive_train,
        X_test, y_test, x_sensitive_test,
```

```
        ['s1'],  # our sensitive feature
        apply_accuracy_constraint=1, # applying our fairness constraint.
        gamma=i
        )
    accuracies.append(score1)
    pvals.append(p_rule1)
```

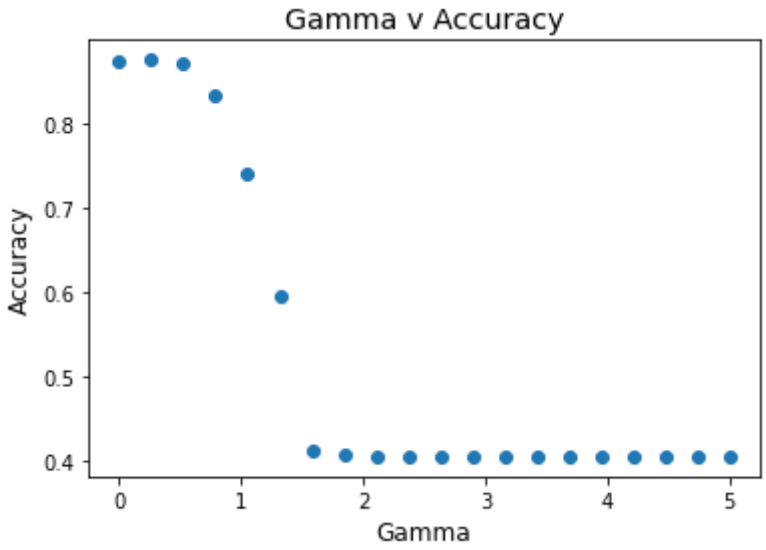
```
Accuracy: 0.87
Protected/non-protected in positive class: 86% / 22%
P-rule achieved: 25%
Covariance between sensitive feature and decision from distance boundary : 1.181

Accuracy: 0.88
Protected/non-protected in positive class: 86% / 21%
P-rule achieved: 25%
Covariance between sensitive feature and decision from distance boundary : 0.482

Accuracy: 0.87
Protected/non-protected in positive class: 86% / 22%
P-rule achieved: 26%
Covariance between sensitive feature and decision from distance boundary : 0.290

Accuracy: 0.83
Protected/non-protected in positive class: 83% / 30%
P-rule achieved: 36%
Covariance between sensitive feature and decision from distance boundary : 0.165

Accuracy: 0.74
Protected/non-protected in positive class: 67% / 38%
P-rule achieved: 57%
Covariance between sensitive feature and decision from distance boundary : 0.069

Accuracy: 0.59
Protected/non-protected in positive class: 82% / 73%
P-rule achieved: 89%
Covariance between sensitive feature and decision from distance boundary : 0.000

Accuracy: 0.41
Protected/non-protected in positive class: 58% / 54%
P-rule achieved: 93%
Covariance between sensitive feature and decision from distance boundary : 0.002

Accuracy: 0.41
Protected/non-protected in positive class: 57% / 54%
P-rule achieved: 94%
Covariance between sensitive feature and decision from distance boundary : 0.003

Accuracy: 0.41
Protected/non-protected in positive class: 56% / 54%
P-rule achieved: 96%
Covariance between sensitive feature and decision from distance boundary : 0.003

Accuracy: 0.41
Protected/non-protected in positive class: 56% / 54%
P-rule achieved: 96%
Covariance between sensitive feature and decision from distance boundary : 0.003

Accuracy: 0.41
Protected/non-protected in positive class: 56% / 54%
P-rule achieved: 96%
Covariance between sensitive feature and decision from distance boundary : 0.003

Accuracy: 0.41
Protected/non-protected in positive class: 56% / 54%
```

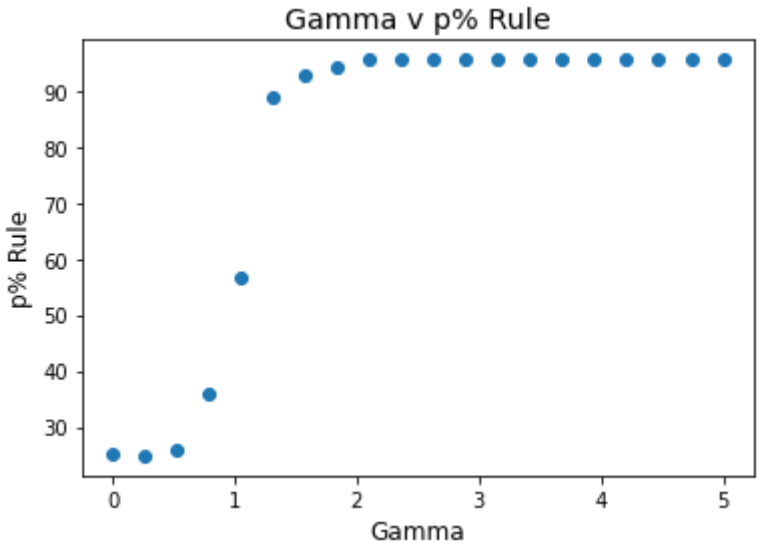
```
P-rule achieved: 96%
Covariance between sensitive feature and decision from distance boundary : 0.003

Accuracy: 0.41
Protected/non-protected in positive class: 56% / 54%
P-rule achieved: 96%
Covariance between sensitive feature and decision from distance boundary : 0.003

Accuracy: 0.41
Protected/non-protected in positive class: 56% / 54%
P-rule achieved: 96%
Covariance between sensitive feature and decision from distance boundary : 0.003

Accuracy: 0.41
Protected/non-protected in positive class: 56% / 54%
P-rule achieved: 96%
Covariance between sensitive feature and decision from distance boundary : 0.003

Accuracy: 0.41
Protected/non-protected in positive class: 56% / 54%
P-rule achieved: 96%
Covariance between sensitive feature and decision from distance boundary : 0.003

Accuracy: 0.41
Protected/non-protected in positive class: 56% / 54%
P-rule achieved: 96%
Covariance between sensitive feature and decision from distance boundary : 0.003

Accuracy: 0.41
Protected/non-protected in positive class: 56% / 54%
P-rule achieved: 96%
Covariance between sensitive feature and decision from distance boundary : 0.003

Accuracy: 0.41
Protected/non-protected in positive class: 56% / 54%
P-rule achieved: 96%
Covariance between sensitive feature and decision from distance boundary : 0.003

Accuracy: 0.41
Protected/non-protected in positive class: 56% / 54%
P-rule achieved: 96%
Covariance between sensitive feature and decision from distance boundary : 0.003
```

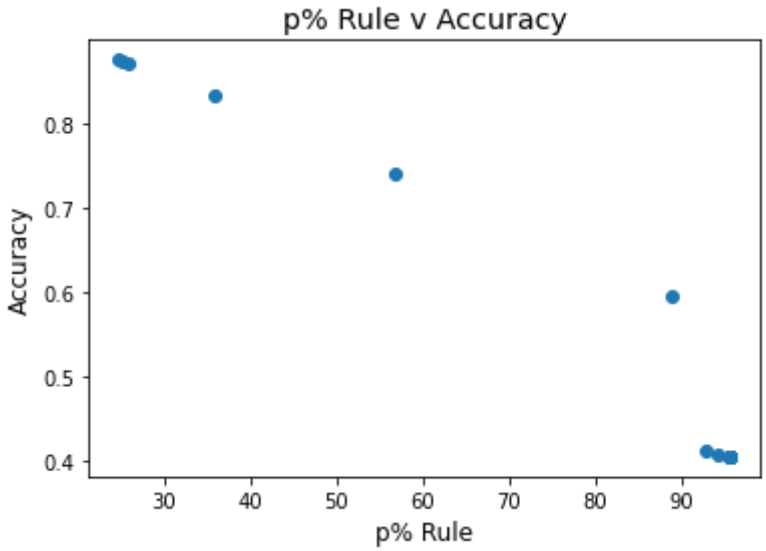
In [16]:

```
#Then, produce three scatterplots: The first should show gamma on the x-axis and
#The second should show gamma on the x-axis and the p-rule on the y-axis.
#The third should show the p-rule on the x-axis and accuracy on the y-axis.
#Remember to make sure the plots are easy to read and well-labeled.
plt.scatter(gammas, accuracies)
plt.xlabel('Gamma', fontsize='large')
plt.ylabel('Accuracy', fontsize='large')
plt.title('Gamma v Accuracy', fontsize='x-large')
plt.show()

plt.scatter(gammas, pvals)
plt.xlabel('Gamma', fontsize='large')
plt.ylabel('p% Rule', fontsize='large')
plt.title('Gamma v p% Rule', fontsize='x-large')
plt.show()

plt.scatter(pvals, accuracies)
plt.xlabel('p% Rule', fontsize='large')
plt.ylabel('Accuracy', fontsize='large')
```

```
plt.title('p% Rule v Accuracy', fontsize='x-large')
plt.show()
```







The following questions have no right answer. Instead, they are meant to make you think about the nuances of navigating correcting for bias in practice.

QUESTION A: What is the "right" value for γ? Why? Can you think of an empirical way to justify your choice for the "right" gamma value?

Answer: We want our classifier to be as accurate as possible while also being as fair as possible. One way to do this is to find the gamma where the tradeoff of achieving a higher p-value would be less than the percentage of accuracy lost. See below for the application of this idea where we find that the "right" gamma is 1.32. This is the point at which our p-rule is 88.9% and our accuracy is .595.

In [17]:
```
pdiff = []
adiff = []
for i in range(0,19):
    adiff.append((abs(accuracies[i]-accuracies[i+1]))/accuracies[i])
    pdiff.append((abs(pvals[i][0]-pvals[i+1][0]))/pvals[i][0])

for i in range(0,19):
    if adiff[i] >= pdiff[i]:
        print(i)
        break

print(gammas[5])
print(pvals[5][0])
print(accuracies[5])
```

```
5
1.3157894736842104
88.90182116938244
0.595
```

QUESTION B: What kinds of discussions or decision making processes could help us agree on γ?

Answer: We should discuss different methods for determining gamma. Each discussion will likely be different for each problem. For example, with this problem, the goal of why we are predicting hiring labels should be discussed. In this case, maximizing p-value may be more important than accuracy if our goal is to correct the bias in the future. However, if the goal is to determine hiring labels if no difference in hiring methods is made, then maximizing accuracy may be most effective.

QUESTION C: In the case of hiring, whom should these discussions involve? How about other cases? Remember that different stakeholders have different levels of understanding about machine learning, and different levels of understanding about social issues such as gender and racial bias.

Answer: These discussions should involve the designers as well as those who are the subject of the design. In this case, females are of particular interest to include in the discussions as they are the group the subject of the bias. In general, the protected groups should always be involved in these discussions.

QUESTION D: Is there some point at which we should stop trying to correct for bias in our data? If so, how do we determine that point?

Answer: There does have to be a point when we stop correcting for bias, as we have to accept

that we will never be able to fully fix the problem of bias through technical measures. Each determination of when we reach that point will depend on the unique problem and should ultimatley be decided in the discussion of the decision making process to decide between the ultimate tradeoff between accuracy and p-rule value for each problem.

Bonus Question: Alternative Ways of Defining Fairness

In this assignment, we have used disparate impact -- as quantified by the p% rule -- to measure bias in our dataset and predictions. However, disparate impact is not the only way to quantify fairness. For this question, we'll look at another option for defining fairness: *disparate mistreatment*, introduced by Zafar et al. (2017b). Disparate mistreatment focuses on the misclassification rates between groups -- that is, it compares the *true positive rate* for the protected class to the *true positive rate* for the unprotected class. In the context of our dataset, disparate mistreatment compares the rate at which qualified females are hired in comaprison to the rate at which qualified men are hired.

Let $\hat{y}$ be the prediction of the classifier for an observation. Then disparate mistreatment is defined as:

$$m = 100 \left(\frac{Pr(\hat{y} = 1 | y = 1, z = 1)}{Pr(\hat{y} = 1 | y = 1, z = 0)} \right)$$

Write a function calculate_disparate_mistreatment that takes in a sensitive_features_arr and *y* (the true hiring labels), along with *distances_from_decision_boundary* as outputed by the *train_test_classifier* method, and returns *m*, the measure of disparate mistreatment. Also have your function print out the true positive rate for each of the classes.

In [18]:
```
# Code up disparate mistreatment
def calculate_disparate_mistreatment(sensitive_features_arr, y, distances_from_d

    # TODO: Replace 0 with the true positive rate in the unprotected class
    tpr_unprotected = 0

    # TODO: Replace 0 with the the true positive rate in the protected class
    tpr_protected = 0

    # TODO: Replace 0 with your calculation of disparate mistreatment
    disparate_mistreatment = 0

    print('True positive rate for unprotected class: %i' % tpr_unprotected + '%'
    print('True positive rate for protected class: %i' % tpr_protected + '%')
    print('Disparate mistreatment: %i' % disparate_mistreatment + '%')

    return disparate_mistreatment
```

In [19]:
```
# Unconstrained classifier
theta, p_rule, score, distances_from_decision_boundary = classify.train_test_cla
    X_train, y_train,  x_sensitive_train,
    X_test, y_test, x_sensitive_test,
    ['s1'],  # our list of sensitive features.
```

```
    apply_fairness_constraints=0 # We are NOT applying any fairness constraints
)
```

```
Accuracy: 0.87
Protected/non-protected in positive class: 86% / 22%
P-rule achieved: 25%
Covariance between sensitive feature and decision from distance boundary : 1.181
```

In [20]:

```
# Test disparate mistreatment on the predictions of the unconstrained classifier
sensitive_features_arr = np.array(x_sensitive_test['s1'])
calculate_disparate_mistreatment(sensitive_features_arr, y_test, distances_from_
```

```
True positive rate for unprotected class: 0%
True positive rate for protected class: 0%
Disparate mistreatment: 0%
```

Out[20]: 0

In [21]:

```
# Constrained classifier
theta1, p_rule1, score1, distances_from_decision_boundary = classify.train_test_
    X_train, y_train, x_sensitive_train,
    X_test, y_test, x_sensitive_test,
    ['s1'],  # our sensitive feature
    apply_accuracy_constraint=1, # applying our fairness constraint.
    gamma=1.5
)
```

```
Accuracy: 0.41
Protected/non-protected in positive class: 60% / 56%
P-rule achieved: 93%
Covariance between sensitive feature and decision from distance boundary : 0.002
```

In [22]:

```
# Test disparate mistreatment on the predictions of the constrained classifier
sensitive_features_arr = np.array(x_sensitive_test['s1'])
calculate_disparate_mistreatment(sensitive_features_arr, y_test, distances_from_
```

```
True positive rate for unprotected class: 0%
True positive rate for protected class: 0%
Disparate mistreatment: 0%
```

Out[22]: 0

QUESTION A: Did using a classifier that was constrained to reduce disparate impact improve disparate mistreatment?

Answer: Replace with your answer

QUESTION B: In general, what is the relationship between disparate impact and disparate mistreatment? Why? *Hint*: A useful resource for this question is Barocas et al. (2019), Chapter 2. Barocas et al. use different terms for the fairness metrics we study here: disparate impact is referred to as *independence* and disparate mistreatment is referred to as *separation*.

Answer: Replace with your answer

QUESTION C: Tell a story where reducing disparate impact may unintentionally increase disparate mistreatment. This story can be in the context of this problem set, but it doesn't have to be.

Answer: Replace with your answer

QUESTION D: How might we decide which fairness metric is employed to test for bias in a classifier? Alternatively, how might we balance multiple fairness criteria? What kinds of decision mamking processes or tools could be employed, and who should be involved in these discussions?

Answer: Replace with your answer

Conclusion

Even if we exclude sensitive attributes from our training procedure, models trained on biased data with strong correlations between sensitive and non-sensitive attributes can still replicate or even emphasize that bias in their predictions.

We can correct for that bias, but doing so will *always* incur tradeoffs in accuracy. In a way, our bias correction is a claim that the data are "misclassified" in the first place. In this lab's example, we think there are too few women labeled as "hired." So we're finding a compromise between correcting that bias and fitting the data we already have.

Based on a paper by Zafar et. al, we demonstrated a bias correction method we can control by setting the hyperparameter γ. But remember: different industries often have different standards for how fair is "fair enough" or how accurate is "accurate enough" for a model to be considered acceptable for use.

Appendix

Generating the toy dataset

To generate the non-sensitive features, we assign each class a bivariate Gaussian distribution and draw the 2 non-sensitive features for each data point from the appropriate distribution. To generate a biased sensitive feature correlated with the non-sensitive features, we rotate the non-sensitive features; the closer the rotation angle is to 0, the more correlated our sensitive feature will be to the non-sensitive features. We use these rotated coordinates to generate a Bernoulli distribution from which to draw the binary sensitive feature. We set the parameters of this distribution such that an individual in the dataset is more likely to be in the unprotected sensitive group (sensitive feature = 1) if the rotated non-sensitive features are also more likely to belong to the positive class (class label = 1).

Applying fairness constraints

To do so, we first define the *decision boundary covariance* as a measure of the covariance between the sensitive feature $\mathbf{z}$ and the classifier's decision $d_{\boldsymbol{\theta}}(\mathbf{x})$. For a logistic regression classifier, the classifier's decision is based on the dot product of the parameters $\boldsymbol{\theta}$ that the model learns and the feature values $\mathbf{x}$.

$$Cov(\mathbf{z}, d_{\boldsymbol{\theta}}(\mathbf{x})) = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_i - \bar{\mathbf{z}})\boldsymbol{\theta}^T\mathbf{x}_i$$

Now, the whole idea here is that we *don't* want the sensitive feature to influence our classifier's decision. Put another way, we want to demonstrate that the sensitive feature and the classifier's decision boundary don't covary---or, at least that the decision boundary covariance is close to 0. Now, we could adjust the classifier to minimize the decision boundary covariance, but this might lead to a significant loss in accuracy which could make our model unusable in practice.

Instead, we want to make our decisions as fairly as possible, with some limit on how low the accuracy can be to produce a functional classifier. So we can set up a mathematical way to frame the trade-off between accuracy and fairness: we can minimize the decision boundary covariance (maximize fairness) given some upper bound on the model's loss (an accuracy constraint).

If $Loss(\boldsymbol{\theta}^*)$ represents the loss of our baseline classifier without any bias correction, we can set the parameter γ to specify how much additional loss we are willing to add to the baseline loss. So our bias correction method can be written:

$$\text{find } \boldsymbol{\theta} \text{ that minimizes } \quad \left|Cov(\mathbf{z}, d_{\boldsymbol{\theta}}(\mathbf{x}))\right|$$

$$\text{subject to } \quad Loss(\boldsymbol{\theta}) \leq (1+\gamma)Loss(\boldsymbol{\theta}^*)$$

We refer to this constraint as the hyperparameter gamma (γ).

In [ ]: