Assignment #4

CS181-S21

Due: 7:59pm EST, March 19, 2021

Homework 4: SVM, Clustering, and Ethics

Introduction

This homework assignment will have you work with SVMs, clustering, and engage with the ethics lecture. We encourage you to read Chapters 5 and 6 of the course textbook.

Please submit the **writeup PDF to the Gradescope assignment 'HW4'**. Remember to assign pages for each question.

Please submit your **LaTeX file and code files to the Gradescope assignment 'HW4 - Supplemental'**.

Problem 1 (Fitting an SVM by hand, 10pts)

For this problem you will solve an SVM by hand, relying on principled rules and SVM properties. For making plots, however, you are allowed to use a computer or other graphical tools.

Consider a dataset with the following 7 data points each with $x \in \mathbb{R}$ and $y \in \{-1, +1\}$:

$$\{(x_i, y_i)\}_{i=1}^{7} = \{(-3, +1), (-2, +1), (-1, -1), (0, +1), (1, -1), (2, +1), (3, +1)\}$$

Consider mapping these points to 2 dimensions using the feature vector $\boldsymbol{\phi}(x) = (x, -\frac{8}{3}x^2 + \frac{2}{3}x^4)$. The hard margin classifier training problem is:

$$\begin{aligned} &\min_{\mathbf{w}, w_0} \frac{1}{2}\|\mathbf{w}\|_2^2 \\ \text{s.t.} \quad &y_i(\mathbf{w}^\top \boldsymbol{\phi}(x_i) + w_0) \geq 1, \ \forall i \in \{1, \dots, n\} \end{aligned}$$

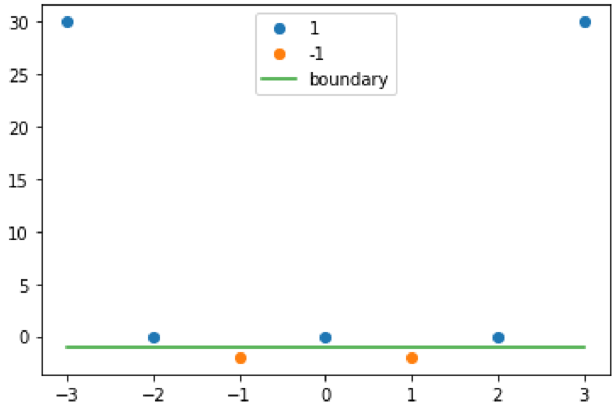
Make sure to follow the logical structure of the questions below when composing your answers, and to justify each step.

1. Plot the transformed training data in $\mathbb{R}^2$ and draw the optimal decision boundary of the max margin classifier. You can determine this by inspection (i.e. by hand, without actually doing any calculations).
2. What is the value of the margin achieved by the optimal decision boundary found in Part 1?
3. Identify a unit vector that is orthogonal to the decision boundary.
4. Considering the discriminant $h(\boldsymbol{\phi}(x); \mathbf{w}, w_0) = \mathbf{w}^\top \boldsymbol{\phi}(x) + w_0$, give an expression for *all possible* $(\mathbf{w}, w_0)$ that define the optimal decision boundary from 1.1. Justify your answer.

   Hint: The boundary is where the discriminant is equal to 0. Use what you know from 1.1 and 1.3 to solve for $\mathbf{w}$ in terms of w_0. (If you solve this problem in this way, then w_0 corresponds to your free parameter to describe the set of all possible $(\mathbf{w}, w_0)$.)
5. Consider now the training problem for this dataset. Using your answers so far, what particular solution to $\mathbf{w}$ will be optimal for the optimization problem?
6. What is the corresponding optimal value of w_0 for the $\mathbf{w}$ found in Part 5 (use your result from Part 4 as guidance)? Substitute in these optimal values and write out the discriminant function $h(\boldsymbol{\phi}(x); \mathbf{w}, w_0)$ in terms of the variable x .
7. Which points could possibly be support vectors of the classifier? Confirm that your solution in Part 6 makes the constraints above tight—that is, met with equality—for these candidate points.

Solution

1. The figure is shown below.



2. Margin is 1.
3. (0,1)
4. We know $\mathbf{w}$ is in the direction of (0,1), so we write $\mathbf{w}^T = k(0,1)$. Now we have

$$0\phi_1(x) + k\phi_2(x) + w_0 = 0,$$

or

$$k = \frac{-w0}{\phi_2(x)}.$$

So

$$\mathbf{w}^T = (0, \frac{-w0}{\phi_2(x)})$$

We also know

$$\phi_2(x) = -1,$$

So

$$\mathbf{w}^T = (0, w0)$$

5. Let

$$w^T\phi(x) + w_0 = 1$$

for (-2,1) and (0,1), and let

$$w^T\phi(x) + w_0 = -1$$

for (-1,1) and (1,-1), I get

$$w^T = (0,1)$$

6. The corresponding w_0 is 1.
7. Support vectors: (-2,1) and (0,1), where the discriminant is 1, (-1,1) and (1,-1), where the discriminant is -1. Because the discriminant is 1 or -1 for these four points, it confirms that my solutions make the constraints above tight.

Problem 2 (K-Means and HAC, 20pts)

For this problem you will implement K-Means and HAC from scratch to cluster image data. You may use `numpy` but no third-party ML implementations (eg. `scikit-learn`).

We've provided you with a subset of the MNIST dataset, a collection of handwritten digits used as a benchmark for image recognition (learn more at http://yann.lecun.com/exdb/mnist/). MNIST is widely used in supervised learning, and modern algorithms do very well.

You have been given representations of MNIST images, each of which is a 784×1 greyscale handwritten digit from 0-9. Your job is to implement K-means and HAC on MNIST, and to test whether these relatively simple algorithms can cluster similar-looking images together.

The code in `T4_P2.py` loads the images into your environment into two arrays – `large_dataset`, a 5000x784 array, will be used for K-means, while `small_dataset`, a 300x784 array, will be used for HAC. In your code, you should use the ℓ_2 norm (i.e. Euclidean distance) as your distance metric.

Important: Remember to include all of your plots in your PDF submission!

Checking your algorithms: Instead of an Autograder file, we have provided a similar dataset, `P2_Autograder_Data`, and some visualizations, `HAC_visual` and `KMeans_visual`, for how K-means and HAC perform on this data. Run your K-means (with $K = 10$ and `np.random.seed(2)`) and HAC on this second dataset to confirm your answers against the provided visualizations. Do **not** submit the outputs generated from `P2_Autograder_Data`. Load this data with `data = np.load('P2_Autograder_Data.npy')`.

1. Starting at a random initialization and $K = 10$, plot the K-means objective function (the residual sum of squares) as a function of iterations and verify that it never increases.

2. Run K-means for 5 different restarts for different values of $K = 5, 10, 20$. Make a plot of the final K-means objective value after your algorithm converges (y-axis) v. the values of K (x-axis), with each data point having an error bar. To compute these error bars, you will use the 5 final objective values from the restarts for each K to calculate a standard deviation for each K.

   How does the final value of the objective function and its standard deviation change with K? (Note: Our code takes 10 minutes to run for this part.)

3. For $K = 10$ and for 3 random restarts, show the mean image (aka the centroid) for each cluster. To render an image, use the pyplot `imshow` function. There should be 30 total images. Include all of these images as part of a single plot; your plot must fit on one page.

4. Repeat Part 3, but before running K-means, standardize or center the data such that each pixel has mean 0 and variance 1 (for any pixels with zero variance, simply divide by 1). For $K = 10$ and 3 random restarts, show the mean image (centroid) for each cluster. Again, present the 30 total images in a single plot. Compare to Part 3: How do the centroids visually differ? Why?

5. Implement HAC for min, max, and centroid-based linkages. Fit these models to the `small_dataset`. For each of these 3 linkage criteria, find the mean image for each cluster when using 10 clusters. Display these images (30 total) on a single plot. How do these centroids compare to those found with K-means? **Important Note:** For this part ONLY, you may use `scipy`'s `cdist` function to calculate Euclidean distances between every pair of points in two arrays.

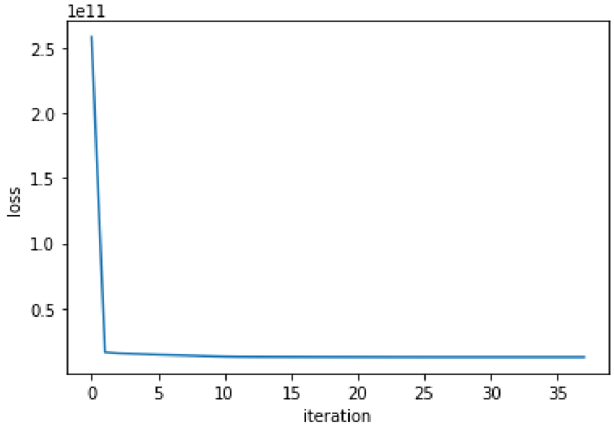
6. For each of the 3 HAC linkages (max/min/centroid), plot ”Distance between most recently merged clusters” (y-axis) v. ”Total number of merges completed” (x-axis). Does this plot suggest that there are any natural cut points?

7. For each of the max and min HAC linkages, make a plot of ”Number of images in cluster” (y-axis) v. ”Cluster index” (x-axis) reflecting the assignments during the phase of the algorithm when there were $K = 10$ clusters. Intuitively, what do these plots tell you about the difference between the clusters produced by the max and min linkage criteria?

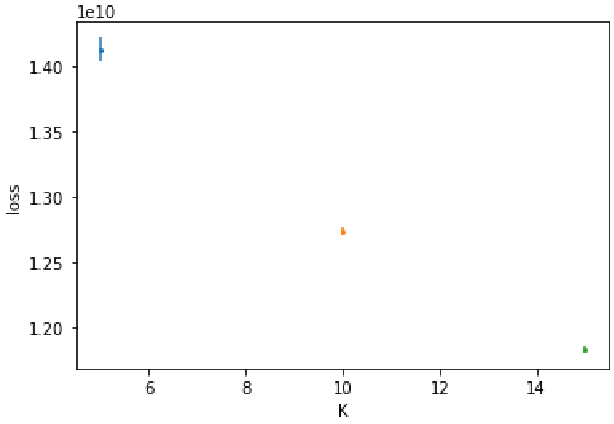
8. For your K-means with $K = 10$ model and HAC min/max/centroid models using 10 clusters on the `small_dataset` images, use the `seaborn` module's `heatmap` function to plot a confusion matrix of clusters v. actual digits. This is 4 matrices, one per method, each method vs. true labeling. The cell at the ith row, jth column of your confusion matrix is the number of times that an image with the true label of j appears in cluster i. How well do the different approaches match the digits? Is this matching a reasonable evaluation metric for the clustering? Explain why or why not.

Solution

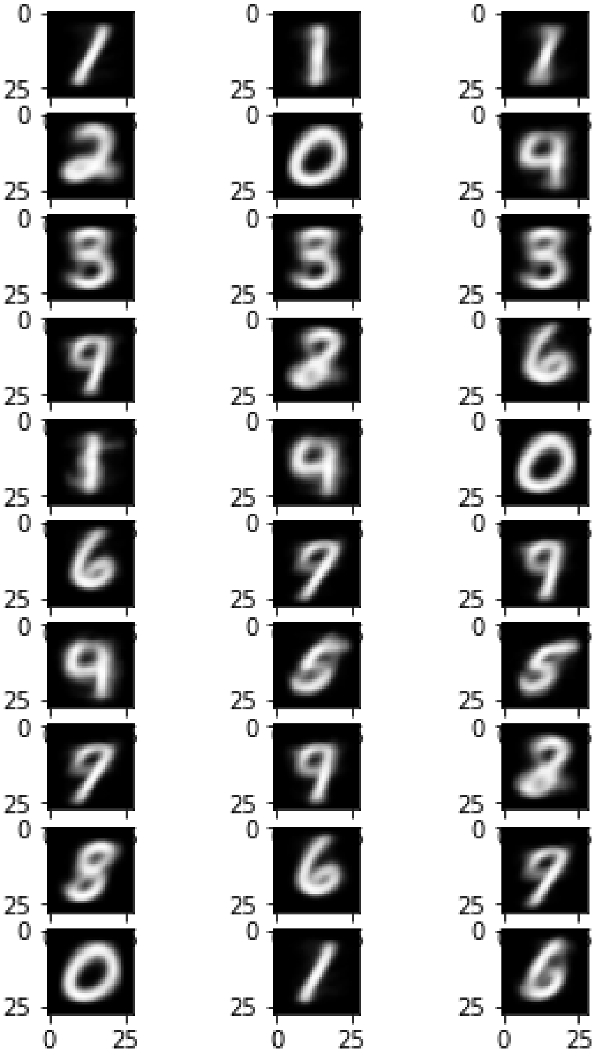
1. The objective function is plotted below. We see it never increases.



2. The figure is shown here. Here the errorbar is quite small, and sometimes even smaller than the size of the markers. As K increases, the final value of the objective function decreases, and the errorbar also decreases.

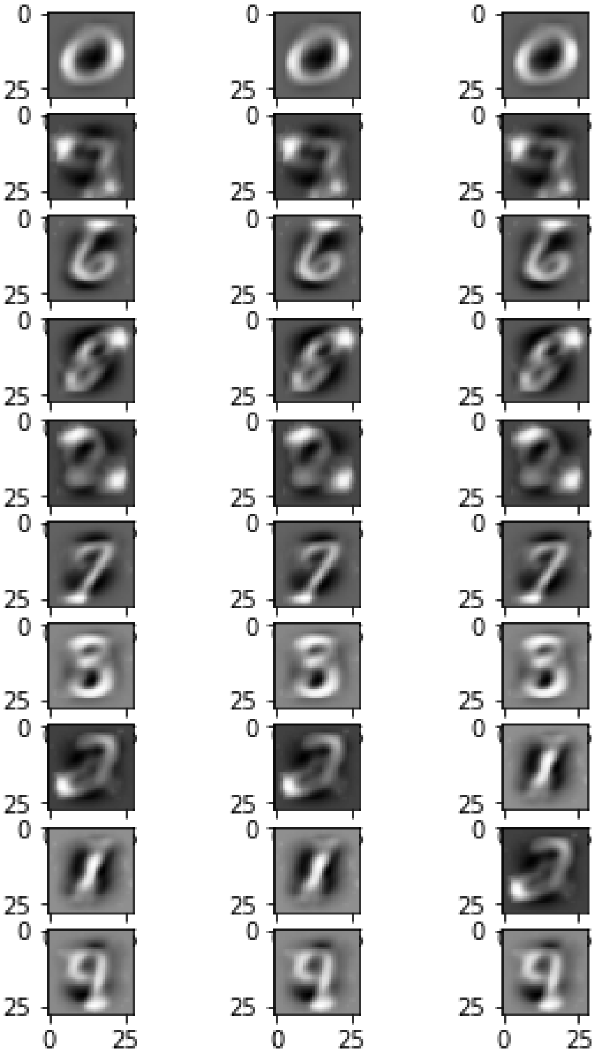


3. The result is shown below.

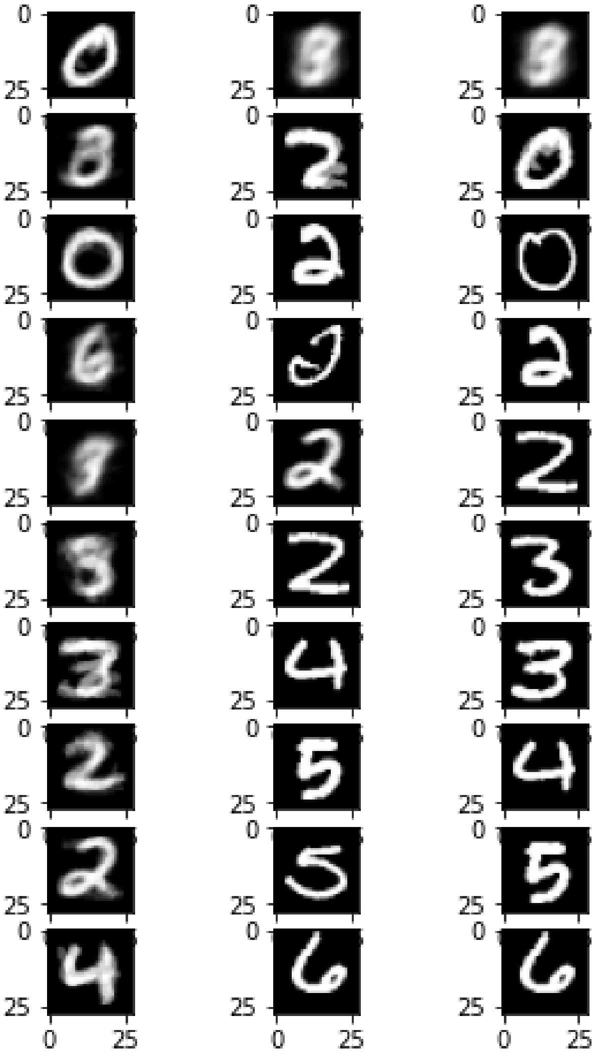


4. The figure is shown here.

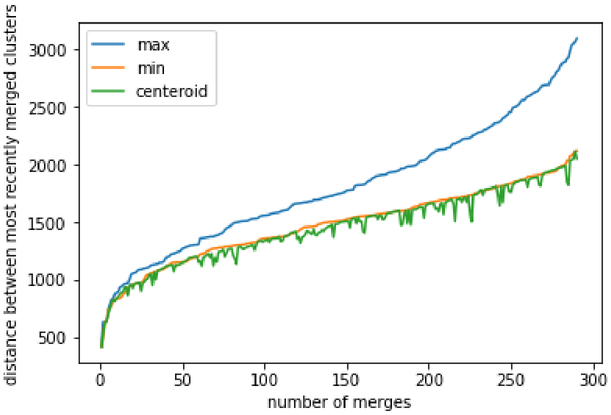
   Compared to the result in Part 3, the result here become more blurry, or more 'grey'. That is because when we are doing the standardization by pixel, we are actually twisting the original patterns for each figure, because the scaled result for a certain figure depends on other figures. And what we are seeing now the 'anomaly', or the normalized deviation from the mean value, so it is no longer clear '0-9' digits. What's more, We should generally see an improved result for standardization, because it prevents the dominant influence of certain pixels(axis). For this case, the result seems to be more robust, we get more similar results for multiple runs, with standardization, but the improvement is not quite significant.



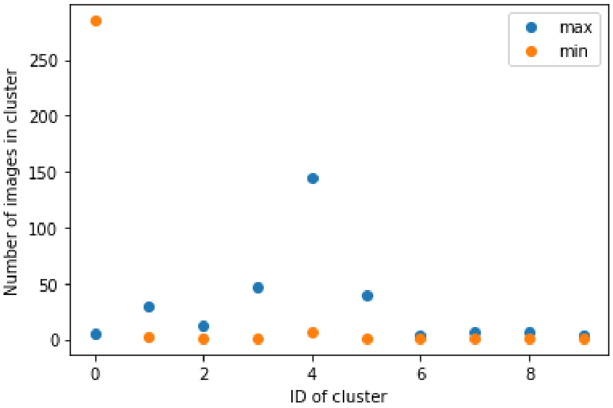
5. The result is shown below. The first column here is 'max' linkage, second is 'min', third is 'centroid' linkage. The result become worse compared to the result in part3. We see for 'max' linkage, there are a few clusters which are blurry, for 'min' and 'centroid' linkage, we see a single cluster which looks like a blend of a lot of datapoints, while other clusters have very clear edge, which implies these clusters contains very few points. For all three linkage methods, we can see multiple centroids of the same numbers, like multiple 2's 3's and 5's. That means the data are not correctly classified as desired.



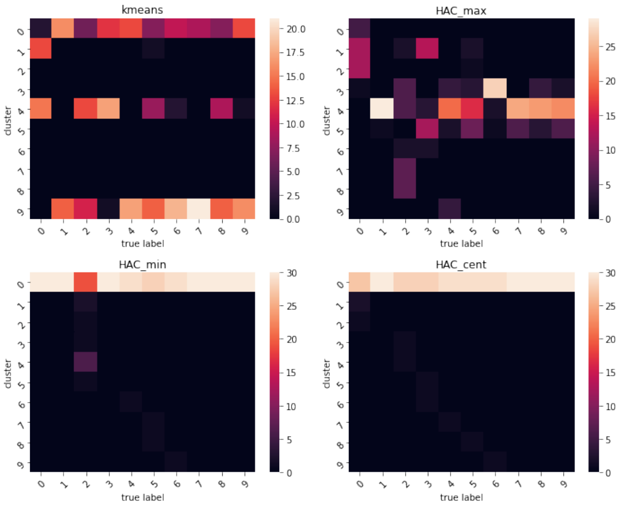
6. As the figure shown below, we can see certain turning points around 10-50 merges, but that does not seem like meaningful cutting points, for a cutting point we should expect the distance increase rapidly instead of increasing slower than before. And if those are cut points, we will get more than 250 clusters. Also from the centroids we see in part5, we can infer the methods does not catch the characters of the digits quite well, so in conclusion there is not a natural cut point.



7. The result below tell me the 'max' linkage method is a better one compared to the 'min' linkage, for this specific question. Because for 'min' linkage, most of the data are put into a single large cluster, which is meaningless, and the rest of clusters are almost empty. For the 'max' method, the points are more evenly distributed into clusters, although the result is still far from perfect.



8. None of the approches match the digits well enough. Here we do not have a label with name for the classified result, but we can still evaluate the result by looking at whether each true value is assigned into a single predicted clusters. The short answer is no for all the methods. For K means, we see each digit is put into more than two clusters, and most of the points go to clusters 0,4,and 9. For HAC_max, we see the result is relatively well compared to other two linkage methods, because the points are relatively evenly distributed into multiple clusters. However, we still see most digits go to more than one clusters, and cluster 4 contains significantly larger amount of data than other clusters. For HAC_min and HAC_centroid, the result is really bad, we see almost all data goes into cluster 0, for both of the methods, which means different digits are not really distinguished by the method, they are still entangled together.


kmeans
cluster
0 1 2 3 4 5 6 7 8 9
true label
0 1 2 3 4 5 6 7 8 9
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
HAC_max
cluster
0 1 2 3 4 5 6 7 8 9
true label
0 1 2 3 4 5 6 7 8 9
25
20
15
10
5
0
HAC_min
cluster
0 1 2 3 4 5 6 7 8 9
true label
0 1 2 3 4 5 6 7 8 9
30
25
20
15
10
5
0
HAC_cent
cluster
0 1 2 3 4 5 6 7 8 9
true label
0 1 2 3 4 5 6 7 8 9
30
25
20
15
10
5
0

Problem 3 (Ethics Assignment, 15pts)

Read the article "Amazon Doesn't Consider the Race of Its Customers. Should It?". Please write no more than 1 concise paragraph each in response to the below reflection questions. We do not expect you to do any outside research, though we encourage you to connect to lecture materials and the reading for the module where relevant.

1. Some people think that Amazon's process for determining which neighborhoods would receive same-day delivery was wrongfully discriminatory, but others disagree. Based on our definitions and discussions from lecture, do you believe that Amazon's same-day delivery process was wrongfully discriminatory? Explain your reasoning.

2. Basing decisions about how to treat others on social group membership often strikes us as being wrongfully discriminatory. For example, most people would say that refusing to hire someone because they are a woman is wrongful discrimination, at least under normal circumstances.

   However, there are some cases in which some people argue that social group membership *should* be taken into consideration when deciding how to treat others. The title of the article poses the question: Do you think that should Amazon consider the race of its customers in its same-day delivery processes? If so, then how?

3. There are many different technical definitions of fairness in machine learning. In this problem, we'll introduce you to the intuition behind two common definitions and invite you to reflect on their limitations.

   Say that Amazon decides to develop a new algorithm to decide its same-day delivery coverage areas. Given your machine learning expertise, Amazon hires you to help them.

   Assume for simplification that Amazon's same-day delivery coverage algorithm f takes as input x, features about an individual Amazon user's account, and outputs binary class label y for whether or not same-day delivery will be offered to user x. User x also has (often unobserved) sensitive attributes a. In this example, we will assume a is a binary label so that $a = 1$ if the user is not white, 0 otherwise.

   One technical notion of algorithmic fairness is called "group fairness"[a]. An algorithm satisfies group fairness with respect to protected attribute a if it assigns the same proportion of positive labels to the group of white and the group of non-white Amazon users. In other words, if 50% of white users have access to same-day shipping, then 50% of non-white users should have access to same-day shipping too.

   What are some limitations or potential issues that may arise with enforcing this definition of fairness in practice? Are there ways that a classifier that satisfies group fairness may still result in discriminatory outcomes?

4. Another technical notion of algorithmic fairness is called "individual fairness"[b]. An algorithm satisfies individual fairness if for all pairs of users x_1 and x_2 that are similar *without taking into consideration their race* a, the algorithm will assign similar probabilities for the attaining the positive label (roughly, $x_1 \sim x_2 \Rightarrow p(x_1) \sim p(x_2)$)[c]. In other words, if two individuals have almost-identical user profiles, then they should both be eligible for same-day shipping, even if one is white and the other is non-white.

   What are some limitations or potential issues that may arise with enforcing this definition of fairness in practice? Are there ways that a classifier that satisfies individual fairness may still result in discriminatory outcomes?

[a] Group fairness is also sometimes referred to as "independence" or "demographic parity". https://fairmlbook.org/classification.html

[b] https://arxiv.org/pdf/1104.3913.pdf

[c] This is an intuitive description of individual fairness (likewise with group fairness above) rather than a precise formalization.

Solution

1. Yes, it is wrongfully discriminatory. Because although Amazon claim they do not consider the race of customers, the result turns out that the area where the residents are almost black people tend to be excluded by the same-day delivery service in a number of cities. Although they claim they are not doing the discrimination on purpose, it turns out the black group are treated different from other group of people, even though their profiles are the same to other customers, which does not seem right. What's more, according to the article, black communities might need that service more urgent than other communities, due to the locations of the traditional markets, but that is not considered by the algorithm.

2. Yes I think it should. Because as the class discussed, sometimes a too neutral representation of the real world will magnify the unfairness and cause the wrongful discrimination in the data analysis process. I think this is one of the cases. When only consider the distribution of the prime members, it might seems lead to the current decision. The the fact that there are less prime members in black communities might be related to some historical reason. The algorithm can simply give more weight to the black/minority community in the algorithm for deciding the same-day delivery areas. It can also try to keep the ratio of the black customers having access to the same-day delivery service similar to the ratio of the black customers or black populations in the city.

3. This definition only aim to treat people in different groups in the same way, but does not consider the difference between groups. People from different group with same profiles may be treated wrongfully differently. For example, if white people tend to not use same-day shipping, but black people tend to use same-day shipping, having group fairness might be unfair to black customers because more black customers need the service. Another example is that if black people live in places where the traditional service is not good enough, in that case, more black customers should be provided that one-day delivery service.

4. This definition only treat people as individuals and assume individuals with same profiles have same backgrounds. But in fact, different group of people have different backgrounds, even though their final profiles are similar. The story for the same-day delivery decision of Amazon is a good example for the discrimination when individual fairness is fulfilled. Amazon claim they do not consider the race of the customers, and only consider the customer profiles, but the actual result is that black zip-codes tend to be excluded by the same-day delivery. That is due to historical reasons, black groups tend to be more difficult to meet the criterion of the customer profiles of Amazon's same day delivery. It is unfair to take the historical reasons as well as the requirement of the same-day delivery of groceries of black communities into account.

Problem 4 (Bonus Ethics Assignment, 0pts)

Estimated total time for completion: 45 minutes.

In our lecture from class, we discussed philosophical and legal frameworks to examine algorithmic discrimination. But these aren't the only frameworks! A growing body of work in the humanities and social sciences, particularly in feminist studies, critical race studies, sociology, anthropology, and the history of science has emerged to study the social consequences of machine learning.

In this bonus problem, you will be introduced to one framework inspired by scholarship in science and technology studies. To complete the below questions, first watch the 28-minute 2019 NeurIPS talk "The Values of Machine Learning" by Ria Kalluri. Please write no more than 1 paragraph in response to each of the below reflection questions:

1. In their talk, Ria discussed opportunities for shifting power to each of four possible stakeholders: an input source, data source, expert, and decision-maker. Choose one of these stakeholders, and discuss one possible way we as machine learning practitioners and model-builders could shift power to them.
2. What do you think will it take to achieve a world where AI can shift power towards historically marginalized communities? What obstacles or barriers stand in between our current world and your own AI dreams?

Solution

1.

Name

Boer Zhang

Collaborators and Resources

Whom did you work with, and did you use any resources beyond cs181-textbook and your notes? I worked with Hanwen Zhang. I also refer to this website when working on HAC https://github.com/Darkprogrammerpb/DeepLearningProjects/blob/master/Project40/agglomerative_hierarchial_clustering/Hierarchial%20Agglomerative%20clustering.ipynb.

Calibration

Approximately how long did this homework take you to complete (in hours)? About 24 hours.