RECITATION 6: FAIRNESS METRICS & UNSUPERVISED LEARNING

10-301/10-601 Introduction to Machine Learning (Summer 2024)
http://www.cs.cmu.edu/~hchai2/courses/10601

1 Fairness Metrics

Neural works for the Bank of ML and is given the following dataset from another bank on whether or not to issue a loan to individuals. Each row in this dataset represents one individual's data, which includes their FICO credit score, their savings rate (percentage of their income that goes into their savings), and credit history in months. The data was collected in two different cities, city A and city B, as denoted in the first column. The "Label" column refers to the true label, where "1" refers to loan issued, and "0" refers to no loan issued.

Region	FICO Score	Savings Rate (%)	Credit History (months)	Label	Prediction
A	544.0625	28.0	21	1	0
A	489.0625	33.9	40	0	0
A	433.125	62.3	100	0	1
A	429.0625	56.7	203	1	1
A	417.8125	56.5	5	0	0
A	506.5625	32.7	75	1	1
A	400.625	60.7	216	0	1
A	836.875	10.7	86	1	1
A	471.875	36.2	92	1	1
A	402.8125	62.0	199	0	1
B	809.4285714	5.6	213	1	1
B	480.9375	40.2	72	1	0
B	505.0	31.1	20	0	0
B	438.4375	51.3	122	0	1
B	385.9375	76.2	89	0	0
B	505.625	34.7	39	1	0
B	514.0625	31.0	41	1	0
B	385.9375	76.2	89	0	0
B	446.25	44.5	51	0	0
B	428.75	55.6	215	1	1

1. Neural trains a model on this dataset and gets a prediction for each training data point, also included in the table. **For parts (a), (b), (c) below, please round your answer to three decimal places**.

   (a) Using the model that Neural proposed, what is the training error rate on the entire dataset?

   (b) What is the training error rate for region A?

   (c) What is the training error rate for region B?

   (d) How many false positives were there in region A?

   (e) How many false negatives were there in region A?

   (f) How many false positives were there in region B?

   (g) How many false negatives were there in region B?

2. **True or False**: Using your responses to the previous question, we achieve statistical parity between regions A and B. Justify your answer.

3. **True or False**: We achieve equality of accuracy between regions A and B. Justify your answer.

4. **True or False**: We achieve equality of PPV/NPV between regions A and B. Justify your answer.

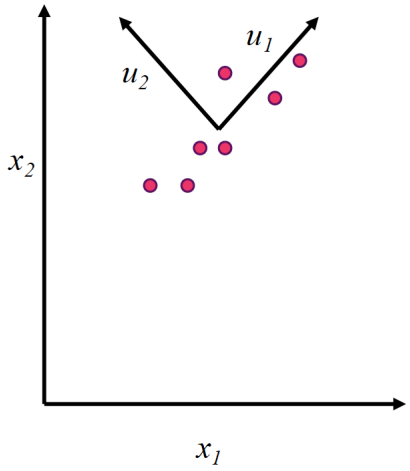
5. Using your responses from the previous questions, comment on the fairness of this model between cities A and B.

6. A Type I error occurs when you erroneously predict a positive label (false positive), and a Type II error is when you erroneously predict a negative label (false negative). Compare and contrast the consequences of making a Type I error and Type II error in this setting. Which would cause more significant consequences?

2 Principal Component Analysis

Principal Component Analysis aims to project data into a lower dimension, while preserving as much as information as possible.

How do we do this? By finding an orthogonal basis (a new coordinate system) of the data, then pruning the "less important" dimensions such that the remaining dimensions minimize the squared error in reconstructing the original data.



In low dimensions, finding the principal components can be done visually as seen above, but in higher dimensions we need to approach the problem mathematically. We find orthogonal unit vectors $\mathbf{v}_1 \dots \mathbf{v}_M$ such that the reconstruction error $\frac{1}{N}\sum_{i=1}^{N}||\mathbf{x}^{(i)}-\hat{\mathbf{x}}^{(i)}||^2$ is minimized, where $\hat{\mathbf{x}}^{(i)} = \sum_{m=1}^{M}(\mathbf{v}_m^T\mathbf{x}^{(i)})\mathbf{v}_m$ are the reconstructed vectors.

If we have M new vectors and d original vectors, with $M = d$, we can reconstruct the original data with 0 error. If $M < d$, it is usually not possible to reconstruct the original data without losing any error. In other words, all the reconstruction error comes from the $M - d$ missing components. This error can be expressed in terms of the covariance matrix of the original data, and is minimized when the principal component vectors $\mathbf{v}_1 \dots \mathbf{v}_M$ are the top M eigenvectors of the covariance matrix (in terms of eigenvalues). The higher the eigenvalues for these eigenvectors are, the more information they store and the lower the reconstruction error.

For the following questions, use this Colab notebook.

Let's assume we've performed PCA on the following dataset:

Row	X1	X2	X3	X4
1	-0.21	-0.61	-0.35	0.08
2	0.15	-0.77	1.26	1.57
3	0.03	0.12	-0.39	-0.25
4	0.92	1.31	0.31	1.19
5	2.51	1.99	1.86	2.57
6	0.91	1.23	-0.01	0.04

And we've obtained the following principal components:

PC1	PC2	PC3	PC4
-0.53	0.23	0.48	-0.66
-0.49	0.7	-0.27	0.44
-0.43	-0.46	0.52	0.57
-0.54	-0.49	-0.65	-0.21

Which correspond to the following eigenvalues:

[3.265, 0.999, 0.043, 0.014]

1. Why are there only 4 principal components?
2. How much of the variance in the data is preserved by the first two principal components?
3. How much of the variance in the data is preserved by the first and third principal components?
4. Perform a dimensionality reduction on the points such that we project them onto the first two principal components. Then, inverse transform it back to four dimensions. What is the reconstruction error for this sample?

5. Perform a dimensionality reduction such that we project the points onto the first and third principal components. Then, inverse transform it back to four dimensions. What is the reconstruction error of this new dataset?

6. Consider the reconstruction error of the fourth row in particular. Is it lower using the first and second principal components or using the first and third? Why might this be the case?

3 K-Means

Clustering is an example of unsupervised machine learning algorithm because it serves to partition **unlabeled** data. There are many different types of clustering algorithms, but the one that is used most frequently and was introduced in class is **K-Means**.

In K-Means, we aim to minimize the objective function:

$$\sum_{i=1}^{n} \min_{j \in \{1,...,k\}} ||\mathbf{x}^{(i)} - \boldsymbol{\mu}_j||^2 \tag{1}$$

Below is the K-Means algorithm:

Let $\mathcal{D} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(n)}\}$ where $\mathbf{x}^{(i)} \in \mathbb{R}^d$ be the set of input examples that each have d features.

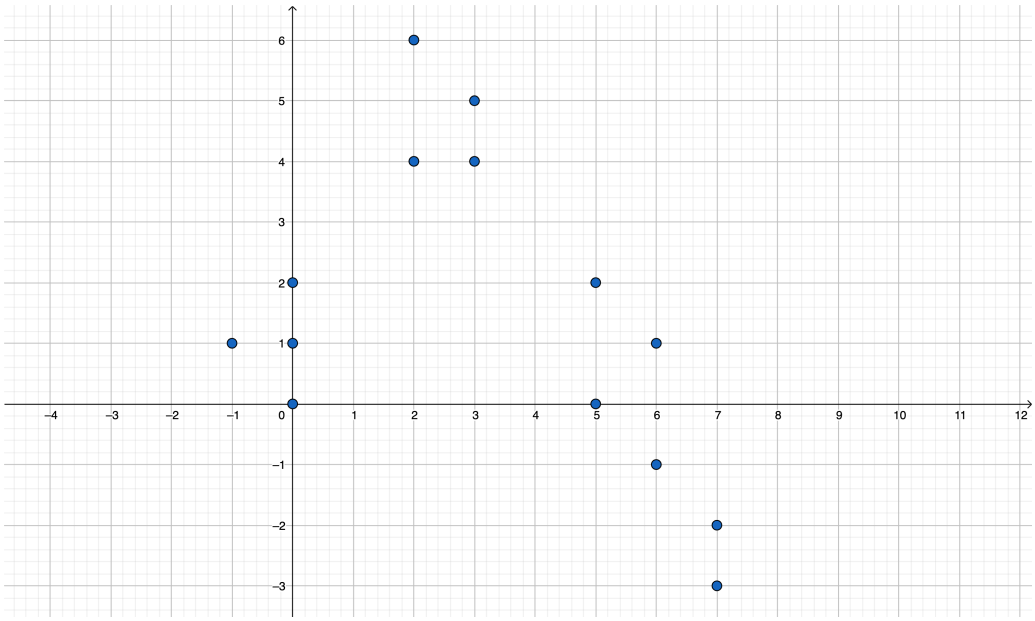
Initialize k cluster centers $\{\boldsymbol{\mu}^{(1)}, ..., \boldsymbol{\mu}^{(k)}\}$ where $\boldsymbol{\mu}^{(i)} \in \mathbb{R}^d$

Repeat until convergence:

1. Assign each point $\mathbf{x}^{(i)}$ to a cluster $\mathcal{C}^{(j)}$ where $j = \text{argmin}_{1 \leq r \leq k} ||\mathbf{x}^{(i)} - \boldsymbol{\mu}^{(r)}||$
2. Recompute each $\boldsymbol{\mu}^{(i)}$ as the mean of points in $\mathcal{C}^{(i)}$

3.1 Walking through an example

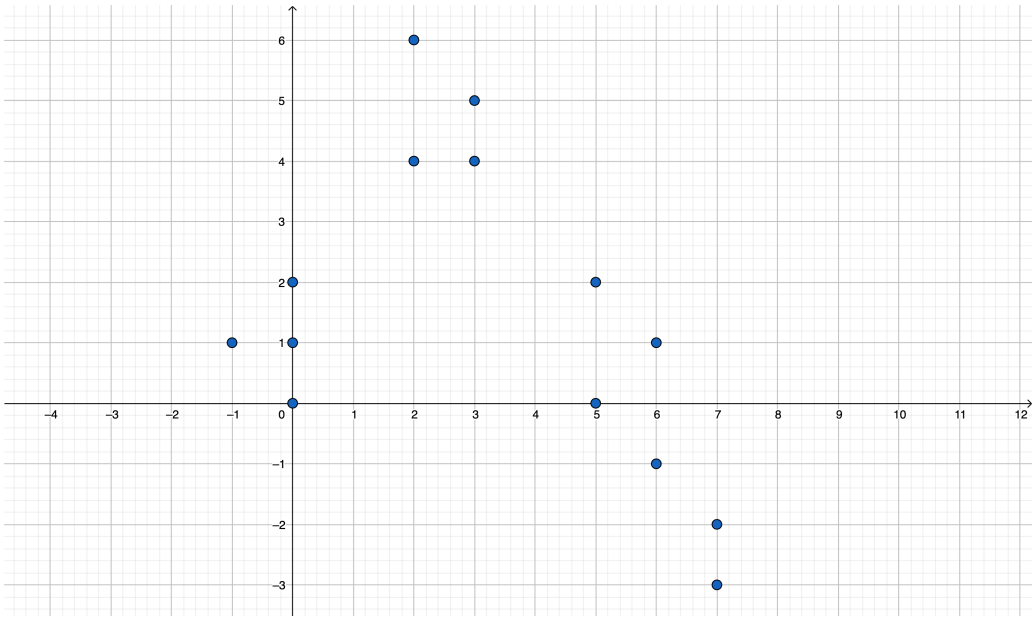
Lets walk through an example of K-Means with $k = 3$ using the following dataset for the first iteration:



Perform one iteration of the K-Means algorithm:

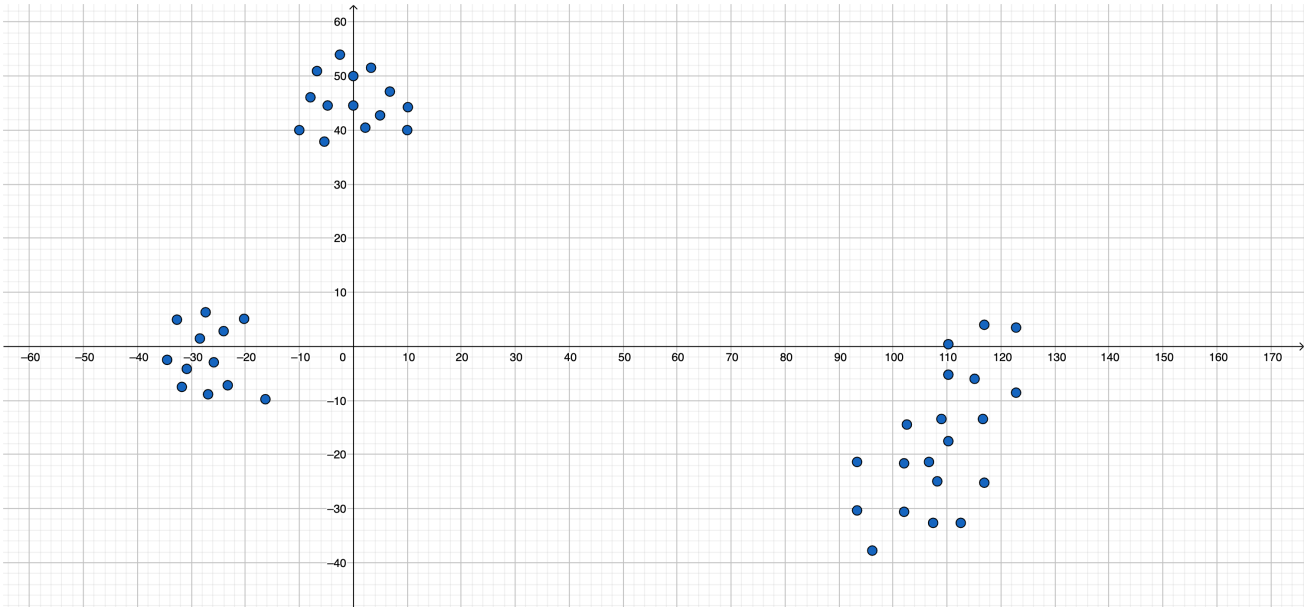
1. What are the cluster assignments?
2. What are the recomputed cluster centers?

3. Draw the cluster assignments after the first iteration on the graph below.



3.2 The importance of initialization

Given the points in the graph below, and assume we will have $k = 3$ cluster centers.



1. Give an example of a set of initialization points such that the K-Means algorithm would converge to a global minimum.

2. Give an example of a set of initialization points such that the K-Means algorithm would converge to a local minimum instead of the global minimum.