RECITATION 6: FAIRNESS METRICS, UNSUPERVISED LEARNING & NAÏVE BAYES

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. A csv file of this dataset could be found in the handout folder.

Neural took the average value of the features (for example, the average value for the first data point is 197.69), and developed the following observation. In general, for all three features in this dataset, a high value indicates better credibility. Hence Neural trained the following decision stump on this dataset: if the average feature value is above the median (198.09), then we determine that the individual will receive the loan (prediction = 1). Otherwise, we decide that the individual will

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

not receive the loan. **For parts (a), (b), (c) below, please round your answer to three decimal places**.

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

Your Answer

Solution 0.400

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

Your Answer

Solution 0.400

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

Your Answer

Solution 0.400

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

Your Answer

Solution 3

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

Your Answer

Solution 1

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

Your Answer

Solution 1

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

Your Answer

Solution 3

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

   ◯ True

   ◯ False

Your Answer

Solution False. In region A, we have number of true negatives = 2 (a), number of true positives = 4 (d), number of false negatives = 1 (c), number of false positives = 3 (b). In region B, we have number of true negatives = 4 (a), number of true positives = 2 (d), number of false negatives = 3 (c), number of false positives = 1 (b). In region A, we get $\frac{b+d}{a+b+c+d} = \frac{7}{10}$, and in region B we get $\frac{3}{10}$

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

   ◯ True

   ◯ False

Your Answer

Solution True. In region A, we have number of true negatives = 2 (a), number of true positives = 4 (d), number of false negatives = 1 (c), number of false positives = 3 (b). In region B, we have number of true negatives = 4 (a), number of true positives = 2 (d), number of false negatives = 3 (c), number of false positives = 1 (b). In region A, we get $\frac{a+d}{a+b+c+d} = \frac{6}{10}$, and in region B we get $\frac{6}{10}$

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

   ◯ True

   ◯ False

Your Answer

Solution False. In region A, we have number of true negatives = 2 (a), number of true positives = 4 (d), number of false negatives = 1 (c), number of false positives = 3 (b). In region B, we have number of true negatives = 4 (a), number of true positives = 2 (d), number of false negatives = 3 (c), number of false positives = 1 (b). In region A, we get $\frac{b}{a+b}/\frac{c}{c+d} = \frac{3}{5}/\frac{1}{5} = 3$, and in region B we get $\frac{1}{5}/\frac{3}{5} = 1/3$

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

   ◯ True

   ◯ False

Your Answer

Solution False. In region A, we have number of true negatives = 2 (a), number of true positives = 4 (d), number of false negatives = 1 (c), number of false positives = 3 (b). In region B, we have number of true negatives = 4 (a), number of true positives = 2 (d), number of false negatives = 3 (c), number of false positives = 1 (b). In region A, we get $\frac{d}{d+b}/\frac{a}{a+c} = \frac{4}{7}/\frac{2}{3} = 6/7$, and in region B we get $\frac{2}{3}/\frac{4}{7} = 7/6$

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

Your Answer

Solution [We should accept any reasonable answer]
Although equality of accuracy between regions A and B is satisfied, other metrics indicate that this model is not fair. In particular, the number of false negatives and false positives between the two regions are quite different.

7. 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?

Your Answer

Solution A Type I error will cause more significant consequences for the bank (issuing a loan to someone who might not be as credible) while a Type II error will cause more significant consequences to the individual, and depending on why they need the loan, a Type II error may have heavier impacts.

2 Naive Bayes

By applying Bayes' rule, we can model the probability distribution $P(Y|X)$ by estimating $P(X|Y)$ and $P(Y)$.

$$P(Y|X) \propto P(Y)P(X|Y)$$

The Naive Bayes assumption greatly simplifies estimation of $P(X|Y)$ - we assume the features X_d are independent given the label. With math:

$$P(X|Y) = \_\_\_\_\_$$

Solution $\prod_{d=1}^{D} P(X_d|Y)$

Different Naive Bayes classifiers are used depending on the type of features.

- Binary Features: Bernoulli Naive Bayes - $X_d \,|\, Y = y \sim \text{Bernoulli}(\theta_{d,y})$
- Discrete Features: Multinomial Naive Bayes - $X_d \,|\, Y = y \sim \text{Multinomial}(\theta_{d,1,y}, \ldots, \theta_{d,K-1,y})$
- Continuous Features: Gaussian Naive Bayes - $X_d \,|\, Y = y \sim \mathcal{N}(\mu_{d,y}, \sigma^2_{d,y})$

We'll walk through the process of learning a Bernoulli Naive Bayes classifier. Consider the dataset below. You are looking to buy a car; the label is 1 if you are interested in the car and 0 if you aren't. There are three features: whether the car is red (your favorite color), whether the car is affordable, and whether the car is fuel-efficient.

Interested?	Red?	Affordable?	Fuel-Efficient?
1	1	1	1
0	0	1	0
0	0	1	1
1	0	0	0
0	0	1	1
0	0	1	1
1	1	1	1
1	1	0	1
0	0	0	0

1. How many parameters do we need to learn?

   Solution 6 for $P(X|Y)$, 1 for $P(Y)$

2. Estimate the parameters via MLE.

Solution

	$Y=1$	$Y=0$
Red?	$\frac{3}{4}$	0
Affordable?	$\frac{1}{2}$	$\frac{4}{5}$
Fuel-Efficient?	$\frac{3}{4}$	$\frac{3}{5}$

3. If I see a car that is red, not affordable, and fuel-efficient, would the classifier predict that I would be interested in it?

   Solution $P(Y=1|\text{red, not affordable, efficient}) \propto \frac{4}{9}\cdot\frac{3}{4}\cdot\frac{2}{4}\cdot\frac{3}{4}=\frac{1}{8}$

   $P(Y=0|\text{red, not affordable, efficient}) \propto \frac{5}{9}\cdot 0\cdot\frac{1}{5}\cdot\frac{3}{5}=0$

4. Is there a problem with this classifier based on your calculations for the previous question? If so, how can we fix it?

   Solution If the car is red, the classifier will always predict I'm interested because $P(\text{not red}|Y=0)=0$. We can use a prior which prevents parameter estimates from being 0, i.e. adding 1 fake count for each feature/label combination.

5. Now we will derive the decision boundary of a 2D Gaussian Naïve Bayes. Show that this decision boundary is quadratic. That is, show that $p(y=1 \mid x_1, x_2) = p(y=0 \mid x_1, x_2)$ can be written as a polynomial function of x_1 and x_2 where the degree of each variable is at most 2. You may fold *unimportant* constants into terms such as C, C', C'', C''' so long as *you are clearly showing each step*.

   Solution Observe that both the LHS and RHS should equal $\frac{1}{2}$ at the decision boundary, so they are both nonzero.

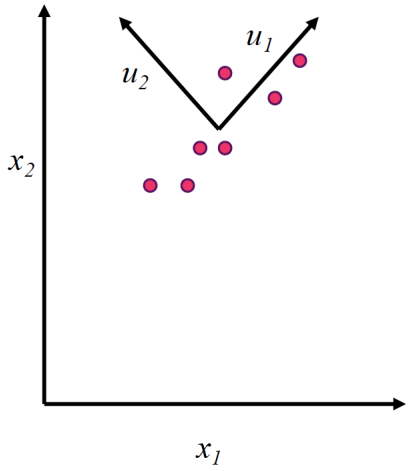
$$
\begin{aligned}
& p(y=1 \mid x_1, x_2) = p(y=0 \mid x_1, x_2) \\
\implies & \frac{p(x_1 \mid y=0)p(x_2 \mid y=0)p(y=0)}{p(x_1, x_2)} = \frac{p(x_1 \mid y=1)p(x_2 \mid y=1)p(y=1)}{p(x_1, x_2)} \\
\implies & 1 = \frac{p(x_1 \mid y=1)p(x_2 \mid y=1)p(y=1)}{p(x_1 \mid y=0)p(x_2 \mid y=0)p(y=0)} \quad (\because \text{nonzero LHS}) \\
\implies & 1 = C\exp\left[\frac{(x_1-\mu_{11})^2}{2\sigma_{11}^2} + \frac{(x_2-\mu_{21})^2}{2\sigma_{21}^2} - \frac{(x_1-\mu_{10})^2}{2\sigma_{10}^2} - \frac{(x_2-\mu_{20})^2}{2\sigma_{20}^2}\right] \\
\implies & 0 = C' + \frac{(x_1-\mu_{11})^2}{2\sigma_{11}^2} + \frac{(x_2-\mu_{21})^2}{2\sigma_{21}^2} - \frac{(x_1-\mu_{10})^2}{2\sigma_{10}^2} - \frac{(x_2-\mu_{20})^2}{2\sigma_{20}^2} \quad (\because \text{nonzero } C)
\end{aligned}
$$

   Since C' is some constant that does not depend on x_1 or x_2, we have shown that the decision boundary is (at most) quadratic x_1 and x_2.

3 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? **Solution**

   There are 4 principal components because the original feature space has dimension 4. Thus, any new basis we construct can only have up to 4 independent components.

2. How much of the variance in the data is preserved by the first two principal components? **Solution** (3.265 + 0.999) / (3.265 + 0.999 + 0.043 + 0.014) = 4.264 / 4.321 = 0.987 * 100 = 99% of the variance.

3. How much of the variance in the data is preserved by the first and third principal components? **Solution** (3.265 + 0.043) / (3.265 + 0.999 + 0.043 + 0.014) = 3.308 / 4.321 = 0.766 * 100 = 76% of the variance.

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?

   Solution The PCA'd dataset is:

$$\begin{bmatrix} 0.52 & -0.36 \\ -1.1 & -1.86 \\ 0.23 & 0.39 \\ -1.9 & 0.41 \\ -4.5 & -0.14 \\ -1.1 & 1.06 \end{bmatrix}$$

   Projected back up to 4 dimensions, we get:

$$\begin{bmatrix} -0.36 & -0.5 & -0.06 & -0.1 \\ 0.16 & -0.77 & 1.33 & 1.51 \\ -0.03 & 0.16 & -0.28 & -0.32 \\ 1.1 & 1.21 & 0.64 & 0.83 \\ 2.36 & 2.09 & 2.02 & 2.5 \\ 0.83 & 1.28 & -0.01 & 0.07 \end{bmatrix}$$

   Reconstruction error is 0.542.

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?

   Solution The new dataset is:

$$\begin{bmatrix} 0.52 & -0.17 \\ -1.1 & -0.08 \\ 0.23 & -0.06 \\ -1.9 & -0.52 \\ -4.5 & -0.03 \\ -1.1 & 0.07 \end{bmatrix}$$

   Projected back up to 4 dimensions, we get:

$$\begin{bmatrix} -0.36 & -0.21 & -0.32 & -0.17 \\ 0.54 & 0.56 & 0.43 & 0.65 \\ -0.15 & -0.1 & -0.13 & -0.09 \\ 0.76 & 1.07 & 0.55 & 1.37 \\ 2.37 & 2.2 & 1.94 & 2.45 \\ 0.62 & 0.52 & 0.52 & 0.54 \end{bmatrix}$$

   Reconstruction error is 5.259.

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?

   Solution Using the first and second principal components:
   Error = $(0.92 - 1.1)^2 + (1.31 - 1.21)^2 + (0.31 - 0.64)^2 + (1.19 - 0.83)^2 = 0.28$

   Using the first and third principal components:
   Error = $(0.92 - 0.76)^2 + (1.31 - 1.07)^2 + (0.31 - 0.55)^2 + (1.19 - 1.37)^2 = 0.17$

   This is because PCA minimizes the mean reconstruction error over all rows, so there may be rows/data points whose reconstruction errors are not minimized (i.e. another choice of projection might yield lower error for those points).

4 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.

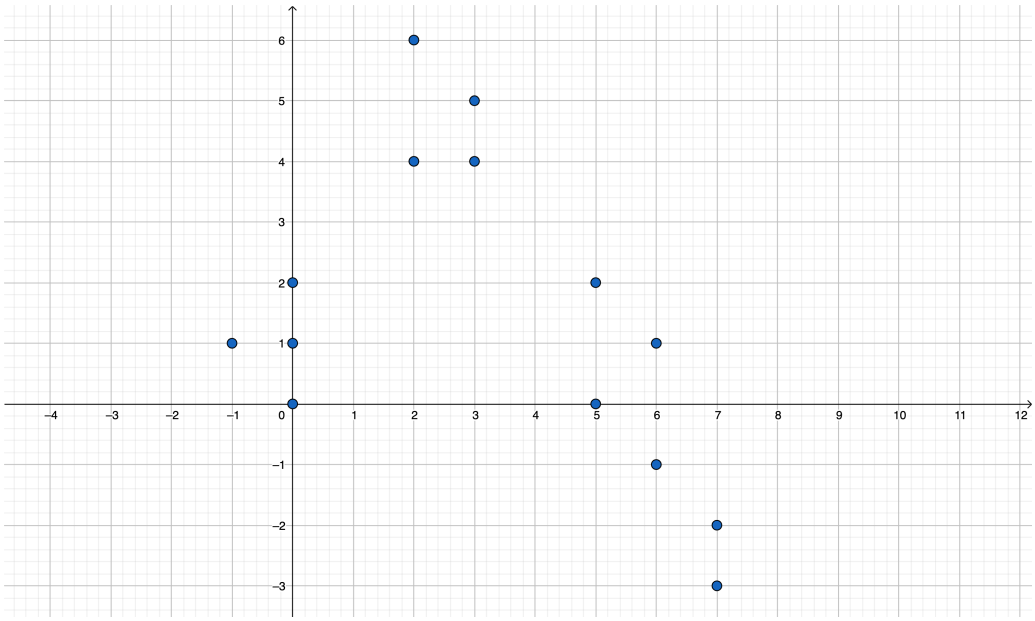
Initialize k cluster centers $\{\boldsymbol{\mu}^{(1)}, ..., \boldsymbol{\mu}^{(k)}\}$ where $\boldsymbol{\mu}^{(i)} \in \mathbb{R}^d$

Repeat until convergence:

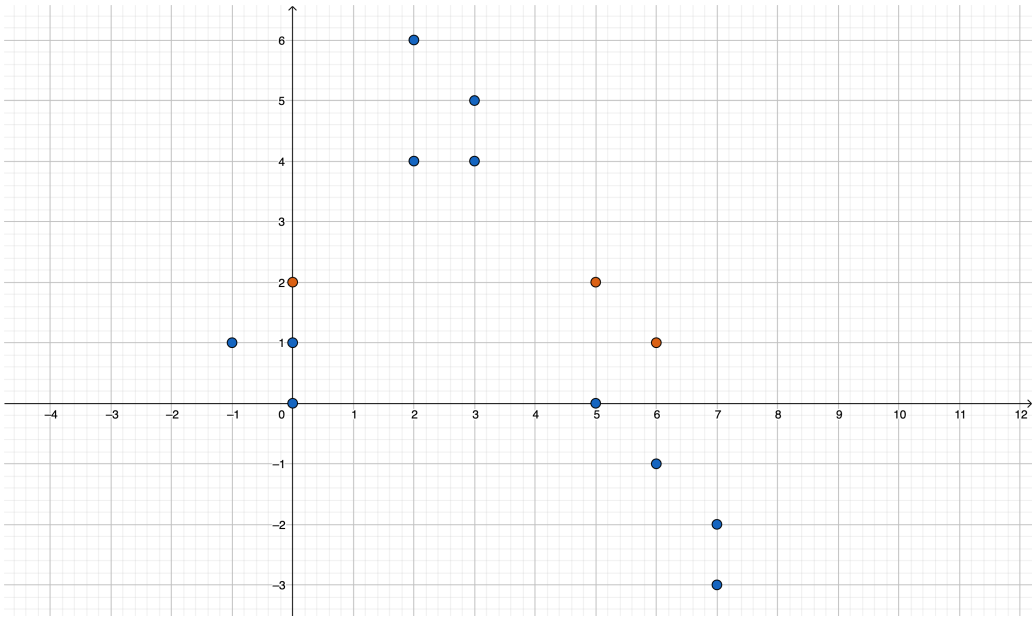
1. Assign each point $\mathbf{x}^{(i)}$ to a cluster $\mathcal{C}^{(j)}$ where $j = \text{argmin}_{1 \leq r \leq k} ||\mathbf{x}^{(i)} - \boldsymbol{\mu}^{(r)}||$
2. Recompute each $\boldsymbol{\mu}^{(i)}$ as the mean of points in $\mathcal{C}^{(i)}$

4.1 Walking through an example

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



Solution Let the cluster centers be initialized to $\mu^{(1)} = (0, 2)$, $\mu^{(2)} = (5, 2)$, $\mu^{(3)} = (6, 1)$ as depicted below in the orange:



Perform one iteration of the K-Means algorithm:

1. What are the cluster assignments? **Solution** $\mathcal{C}^{(1)} = \{(0,0), (-1,1), (0,1), (0,2), (2,4), (2,6)\}$

   $\mathcal{C}^{(2)} = \{(3,4), (3,5), (5,2)\}$

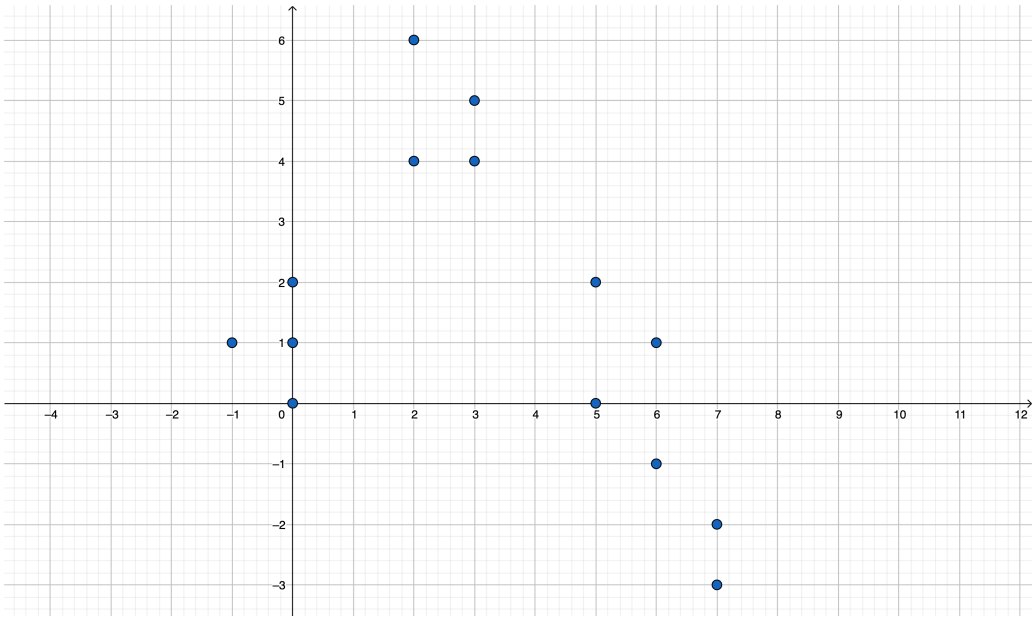
   $\mathcal{C}^{(3)} = \{(5,0), (6,1), (6,-1), (7,-2), (7,-3)\}$

2. What are the recomputed cluster centers? **Solution** $\mu^{(1)} = (0.5, 2.33)$

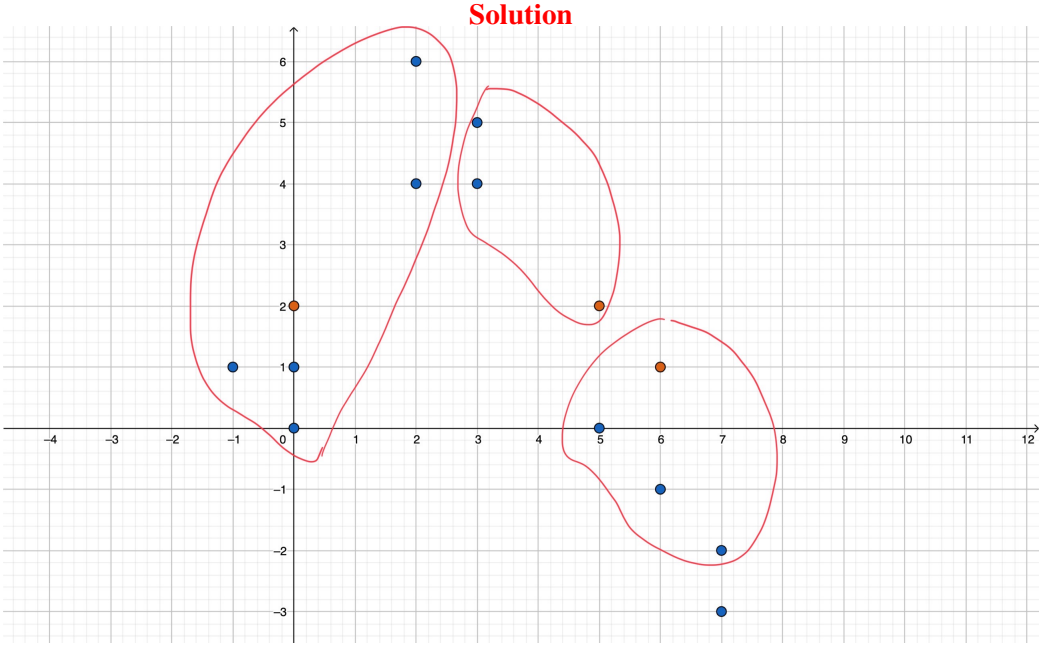
   $\mu^{(2)} = (3.67, 3.67)$

   $\mu^{(3)} = (6.2, -1)$

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

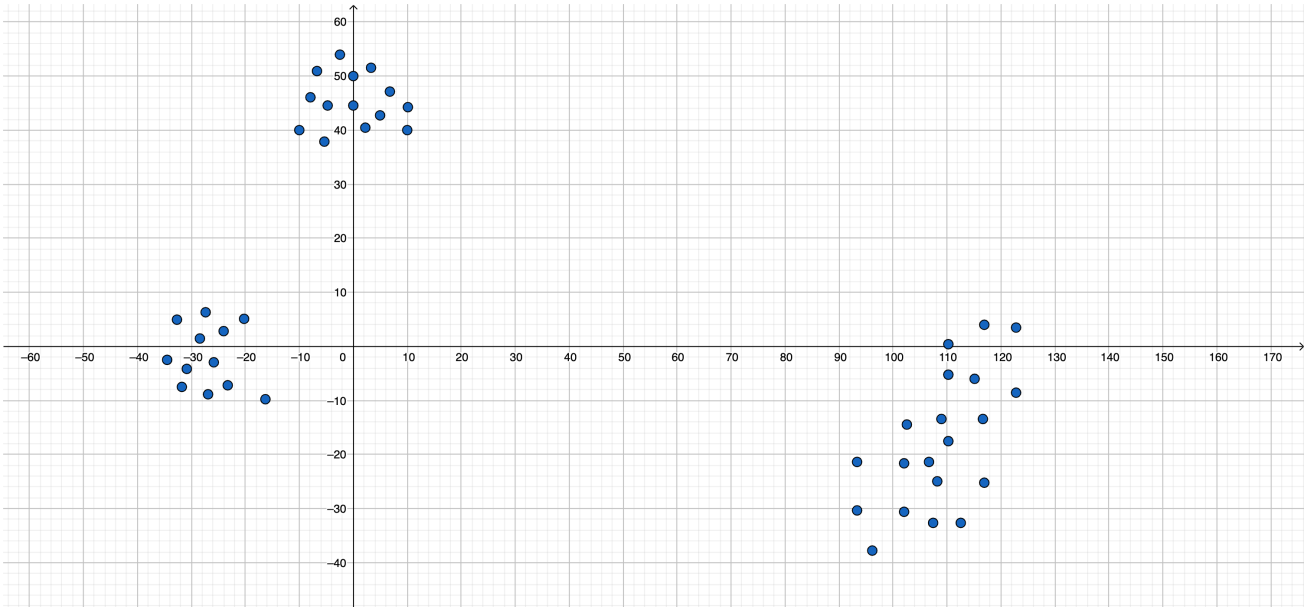


Solution



4.2 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.

   Solution Any three points where each belongs to a different cluster

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.

   Solution For example, one to the upper left corner, the other two at the bottom right corner