Problem 1. (PCA, 4 points) Consider the following dataset $X \in \mathbb{R}^{3\times 2}$ with 3 data entries and two features. Furthermore, its covariance matrix is denoted by Cov:

$$X = \begin{bmatrix} x_1^1 & x_2^1 \\ x_1^2 & x_2^2 \\ x_1^3 & x_2^3 \end{bmatrix} = \begin{bmatrix} 20 & -1 \\ 10 & 2 \\ 15 & 0 \end{bmatrix}, \qquad Cov = \begin{bmatrix} C_{11} & C_{12} \\ C_{12} & C_{22} \end{bmatrix} = \begin{bmatrix} 25 & -7.5 \\ -7.5 & 2.33 \end{bmatrix}.$$

1. Which of the following computes the correlation between feature 1 and feature 2? Circle the correct answer.

   (a) $\frac{\sqrt{C_{12}}}{C_{11}C_{22}}$ (b) $\frac{C_{12}}{\sqrt{C_{11}C_{22}}}$ (c) $-\frac{C_{12}}{C_{11}C_{22}}$

2. The correlation between the two features is: (a) positive (b) negative (c) zero

   Solution:

   $$cor_{x_1,x_2} = \frac{C_{12}}{\sqrt{C_{11}C_{22}}} = \frac{-7.5}{\sqrt{25}\sqrt{2.33}} < 0,$$

   The correlation between feature 1 and feature 2 is negative.

Consider the dataset $\tilde{X}$ that is normalized (by subtracting the mean and dividing by variance for each feature):

$$\tilde{X} = \begin{bmatrix} 1.25 & -1.25 \\ 0 & 1.25 \\ -1.25 & 0 \end{bmatrix}$$

The eigenvalues of $\tilde{X}^\top \tilde{X}$ are $\lambda_1 = 2.3, \lambda_2 = 0.8$ and the corresponding eigenvectors are $v_1 = \begin{bmatrix} 0.7 \\ -0.7 \end{bmatrix}, v_2 = \begin{bmatrix} 0.7 \\ 0.7 \end{bmatrix}$. Now we project the data matrix onto the space spanned by the first principal component, we denote the projected data matrix by Z.

3. The dimension of the projected data matrix $Z \in$ .....

4. Derive the projected data matrix Z. (You may leave your answer as a matrix multiplication).

   Solution: The dimension of the projected data matrix $Z \in \mathbb{R}^{3\times 1}$. Since $\lambda_1 > \lambda_2$ the first principal component is v_1. Thus, the data is projected onto v_1:

   $$Z = \tilde{X}v_1 = \begin{bmatrix} 1.25 & -1.25 \\ 0 & 1.25 \\ -1.25 & 0 \end{bmatrix} \begin{bmatrix} 0.7 \\ -0.7 \end{bmatrix} = \begin{bmatrix} 1.75 \\ -0.875 \\ -0.875 \end{bmatrix}.$$

Problem 2. (k-means, 3 points) You are given a data set $X \in \mathbb{R}^{6\times 2}$ for which you have used k-means clustering with $k = 2$ to cluster your data. Each cluster contains three points and the center of cluster 1, 2, is given by $\mu^1 = (6, 2)$ and $\mu^2 = (1, 6)$, respectively.

1. Consider the sample $(5, ?)$. Which will be the missing entry based on the k-means clusters?
   (a) 0; (b) 2; (c) 4; (d) 6.
2. Consider now a sample $x \in \mathbb{R}^2$ equal to $(1, 2)$. To which cluster would you assign the point?
3. Recompute the center of the cluster with the new point from question 2.

Solution:

1. We first compute the Euclidian distance between the sample and the cluster means μ^1 and μ^2 using only the available components:

$$f^1 = f(x, \mu^1) = |6 - 5| = 1$$

$$f^2 = f(x, \mu^2) = |1 - 5| = 4$$

   The sample is closer to the center of cluster 1. For the second input we simply take the value of $\mu_2^1 = 2$. Thus the inputed value is $(5, 2)$, so the correct answer is (b).

2. To find which cluster x belongs we first calculate the Euclidean distance between x and the cluster means μ^1 and μ^2. The distances are:

$$f^1 = f(x, \mu^1) = \sqrt{(6-1)^2 + (2-2)^2} = \sqrt{25}$$

$$f^2 = f(x, \mu^2) = \sqrt{(1-1)^2 + (6-2)^2} = \sqrt{16}$$

   Thus, x is closer to the center of cluster 2.

3. The center is the average of the points:

$$\mu_1^2 = \frac{1}{|S_2|} \sum_{x \in S_2} x_1$$

$$\mu_2^2 = \frac{1}{|S_2|} \sum_{x \in S_2} x_2$$

   In our case, there were already three points in the

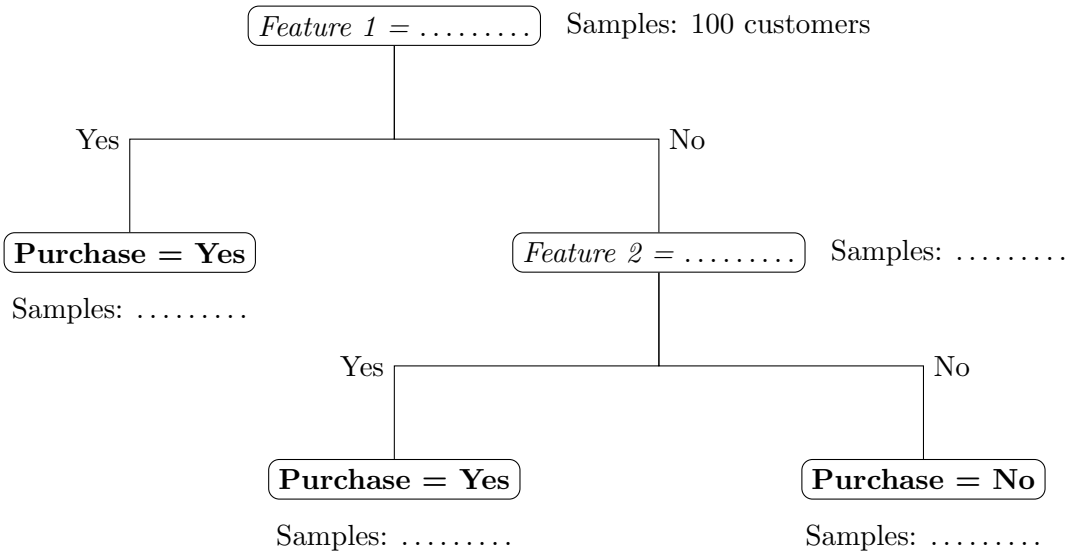
$$\mu_1^2 = \frac{3 * 1 + 1}{4} = 1$$

$$\mu_2^2 = \frac{3 * 6 + 2}{4} = 5$$

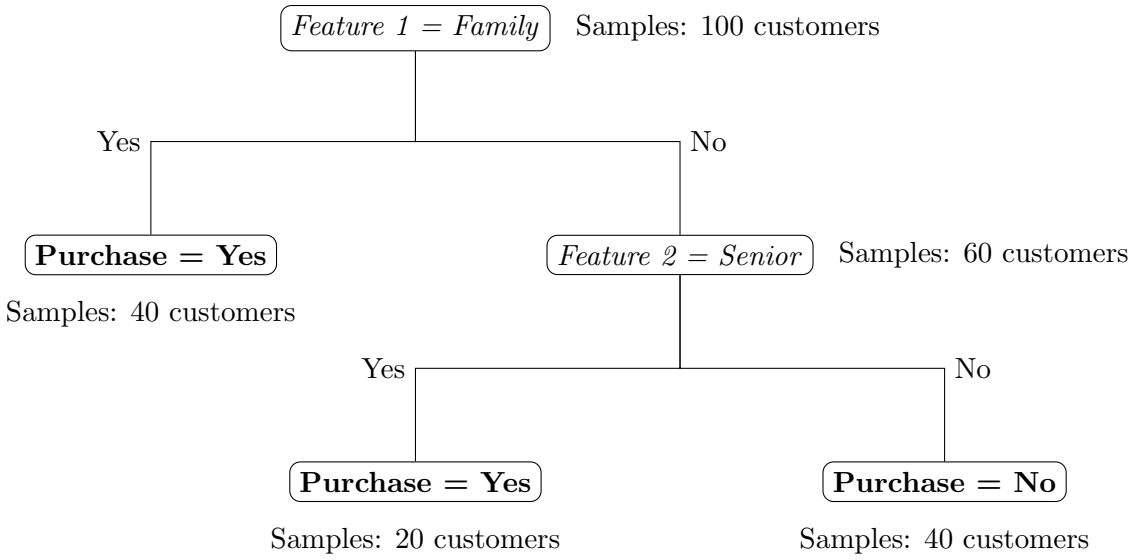
Problem 3. (Decision tree, 3 points)

A company aims to predict the likelihood of customers purchasing a product using a decision tree method, trained on the data of 100 customers, each characterized by two features: 1) Age: Young, Middle-aged, Senior; and 2) Family: Yes, No. It was found that the presence of the "Family" feature had the lowest Gini index. In particular, all 40 customers who had a family purchased the product. For the remaining 60 customers, the "Senior" feature had the lowest Gini index. Specifically, among the 20 senior customers, 15 purchased the product. Among the other 40 customers (i.e., young or middle-aged), only 5 purchased the product. You decided to stop growing your decision tree at depth 2.

1. Fill in the blank parts for features and samples in the decision tree. For example, we fill in the samples for the first node as 100 customers since we are given data for 100 customers.



Solution:



2. What was the Gini index of splitting the "no family" node into a leaf node corresponding to "senior" and "not senior"?

   Solution:

$$\text{Gini index: } = \frac{20}{60}\left(\frac{5}{20}\frac{15}{20} + \frac{15}{20}\frac{5}{20}\right) + \frac{40}{60}\left(\frac{5}{40}\frac{35}{40} + \frac{35}{40}\frac{5}{40}\right) = \frac{13}{48} = 0.4162$$

3. Using the decision tree you constructed, predict whether a customer who is young and does not have a family would purchase the product.

   Solution: Since we split on "Age = non-senior" which results in a leaf node where "Purchase = No", we predict that this customer would not purchase the product.