# EE708, Assignment 4 Submission

**Dhruv Gupta**, Roll no: 240354

# Problem 1

Question

A dataset contains 200 samples classified into two classes:
120 positive and 80 negatives.

a Compute the Gini index before splitting.

b If a split results in subsets:
Left: (50 positive, 10 negative) Right: (70 positive, 70 negative)
Compute the weighted Gini index and determine whether the split improves purity

Answer

Formula for Gini Index is:

$$\text{Gini-Index} = 1 - \sum_{i=1}^{K} p(i)^2$$

Where $p(i)$ represents the probability of the object being classified into a particular class out of $K$ classes.
We have, $p(\text{positive}) = 120/200 = 0.6$ and $p(\text{negative}) = 80/200 = 0.4$ Hence, our Gini-Index before splitting can be written as:

$$\text{GI}_{initial} = 1 - \left(0.6^2 + 0.4^2\right) = 0.48$$

After the split happens:

$$p(\text{positive and left}) = 50/60 = 0.833 \qquad p(\text{positive and right}) = 70/140 = 0.5$$

$$p(\text{negative and left}) = 10/60 = 0.167 \qquad p(\text{negative and right}) = 70/140 = 0.5$$

$$\text{GI}_{left} = 1 - \left(0.833^2 + 0.167^2\right) = 0.278 \qquad \text{GI}_{right} = 1 - \left(0.5^2 + 0.5^2\right) = 0.5$$

Weighted sum of Gini-Indices after the split can be written as:

$$\text{GI}_{final} = p(\text{left}) \cdot \text{GI}_{left} + p(\text{right}) \cdot \text{GI}_{right}$$

$$\implies \text{GI}_{final} = \frac{60}{200} \cdot 0.278 + \frac{140}{200} \cdot 0.5 = 0.43$$

So, we get: the final Gini-Index after splitting as 0.43, which is less than the initial GI, hence, **the split improves purity.**

# Problem 2

Question

Consider the given dataset with two independent variables ($x_1$, $x_2$) and one dependent variable ($y$):

a Use the sum of squared errors (SSE) to determine the best splitting point for $x_1$.

b Construct the first split of a regression tree using SSE as the impurity measure.

Answer

**(a) Best Splitting Point for $x_1$ Using SSE**

Calculate Total SSE Without Splitting: The mean of $y$ is:

$$\bar{y} = \frac{10 + 12 + 15 + 18 + 21 + 25 + 28 + 30}{8} = 19.875$$

The total SSE is:

$$SSE = \sum (y_i - 19.875)^2 = 484.875$$

Evaluate Possible Splits for $x_1$ and Compute SSE:
For each split, we divide the data into left ($x_1 \leq k$) and right ($x_1 > k$) groups. The SSE for each group is calculated using the formula:

$$SSE = \sum (y_i - \bar{y})^2$$

After evaluating all possible splits, the split at $x_1 = 4.5$ results in the lowest SSE, making it the best choice.

**(b) First Split Using SSE as the Impurity Measure**

We can consider the first split from either $x_1$ or $x_2$. On calculation, we see that the SSE is same for first split at either $x_1 = 4.5$ or $x_2 = 10$. At $x_1 = 4.5$:

$$\bar{y}_L = \frac{10 + 12 + 15 + 18}{4} = 13.75 \qquad \bar{y}_R = \frac{21 + 25 + 28 + 30}{4} = 26$$

$$SSE_R = (21 - 26)^2 + (25 - 26)^2 + (28 - 26)^2 + (30 - 26)^2 = 34$$

$$SSE_L = (10 - 13.75)^2 + (12 - 13.75)^2 + (15 - 13.75)^2 + (18 - 13.75)^2 = 21.5$$

Total SSE After Split:

$$SSE_{\text{split}} = 21.5 + 34 = 55.5$$

Thus, the first split is made at $x_1 = 4.5$ using SSE as the impurity measure.

# Problem 3

## Question

Consider a 2-dimensional feature space with a dataset of $N = 10$ points. VQ system maps these points into $K = 3$ clusters using a codebook. Given the following initial cluster centroids:

$$C_1 = (2, 3) \qquad C_2 = (5, 8) \qquad C_3 = (9, 4)$$

Assign the following data points to their closest centroid using squared Euclidean distance:

$$(1, 2) \qquad (3, 4) \qquad (6, 7) \qquad (8, 3), \qquad (5, 5)$$

a Compute the new centroids after one iteration of vector quantization.
b Show whether the distortion decreases after this iteration.

## Answer

**Computing distances:** Given Centroids: **Cluster Assignment** Given the initial centroids:

$$C_1 = (2, 3), \quad C_2 = (5, 8), \quad C_3 = (9, 4)$$

And the five data points:

$$P_1 = (1, 2), \quad P_2 = (3, 4), \quad P_3 = (6, 7), \quad P_4 = (8, 3), \quad P_5 = (5, 5)$$

We compute the squared Euclidean distance (distortion) for each data point to the centroids using the following formula:

$$d^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2$$

**Assigning clusters:** For the first point, $P_1$, we have the distances:

$$d^2(P_1, C_1) = 2 \qquad d^2(P_1, C_1) = 52 \qquad d^2(P_1, C_1) = 68$$

Hence,$P_1$ belongs in cluster 1.
Similarly, following this process of assigning clusters, we get the following results:

$$(1, 2) \to C_1, \quad (3, 4) \to C_1,$$
$$(6, 7) \to C_2, \quad (8, 3) \to C_3, \quad (5, 5) \to C_2$$

**Centroid Updates:** New centroids are calculated as the mean of points in each cluster:

$$C_1' = \frac{(1, 2) + (3, 4)}{2} = (2, 3) \qquad C_2' = \frac{(6, 7) + (5, 5)}{2} = (5.5, 6) \qquad C_3' = (8, 3)$$

**This completes our first iteration of vector quantisation.**

Answer continued

**Distortion Comparison:** The total distortion is calculated as:

$$D = \sum_{\text{all points}} d^2(\text{point, its centroid})$$

Initial distortion (with centroids $C_i$): $D_{\text{initial}} = 2 + 2 + 2 + 2 + 9 = 17$
Final distortion (with centroids $C_i'$): $D_{\text{final}} = 2 + 2 + 1.25 + 1.25 + 0 = 6.5$

Clearly, the **distortion has decreased significantly** after this iteration of VQ.

# Problem 4

Question

$$E_Z[\ln p(X, Z|\mu, \Sigma, \pi)] = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})(\ln \pi_k + \ln \mathcal{N}(x_n|\mu_k, \Sigma_k))$$

Show that if we maximize the first equation with respect to $\Sigma_k$ and $\pi_k$ while keeping the responsibilities $\gamma(z_{nk})$ fixed, we obtain the closed-form solutions given by the following equations:

$$\Sigma_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T \qquad \pi_k = \frac{N_k}{N}$$

Answer

We perform the process of "Expectation Maximization"; we have to maximize the log-likelihood function with respect to $\Sigma_k$ and $\pi_k$, hence we will partially differentiate it with respect to these parameters and set the equation equal to zero.

**Gaussian Distribution:** $\mathcal{N}(x_n|\mu_k, \Sigma_k)$ is a Gaussian distribution function given by:

$$\mathcal{N}(x|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_k)^T\Sigma_k^{-1}(x - \mu_k)\right)$$

Logarithm of $\mathcal{N}(x_n|\mu_k, \Sigma_k)$ would be:

$$ln\ \mathcal{N}(x|\mu_k, \Sigma_k) = -\frac{1}{2}ln|\Sigma_k| - \left(\frac{1}{2}(x - \mu_k)^T\Sigma_k^{-1}(x - \mu_k)\right) + \text{constant}$$

We will use this result in further steps.

Answer continued

**Maximization of expected log likelihood function:** We substitute the calculated value of $ln\ \mathcal{N}(x|\mu_k, \Sigma_k)$ into the equation:

$$E_Z[\ln p(X, Z|\mu, \Sigma, \pi)] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \left(\ln \pi_k + \ln \mathcal{N}(x_n|\mu_k, \Sigma_k)\right)$$

Note: We are keeping the responsibility $\gamma(z_{nk})$ constant here for all further calculations as mentioned in the question.

**Maximize with respect to $\Sigma_k$:** We differentiate the expected log likelihood with respect to $\Sigma_k$ and set it equal to zero.

$$\frac{\partial}{\partial \Sigma_k} \sum_{n=1}^{N} \gamma(z_{nk}) \left( -\frac{1}{2} \ln |\Sigma_k| - \frac{1}{2} (x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k) \right) = 0$$

Solving this results in the closed-form solution:

$$\Sigma_k = \frac{1}{\psi} \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T, \qquad \text{where } \psi = \sum_{n=1}^{N} \gamma(z_{nk})$$

**Maximize with respect to $\pi_k$:** Here, $\pi_k$ represents the mixing coefficients, so there is a constraint that $\sum_{k=1}^{K} \pi_k = 1$, this means we can use the method of Lagrange Multipliers and solve this.

We differentiate the Log-Likelihood wrt $\pi_k$ and get:

$$\pi_k = \frac{\psi}{N}, \qquad \text{where } \psi = \sum_{n=1}^{N} \gamma(z_{nk})$$

**$\psi$ represents the effective number of points assigned to a specific cluster**

# Problem 5

Question

Consider a density model given by a mixture distribution

$$p(x) = \sum_{k=1}^{K} \pi_k p(x|k)$$

and suppose that we partition the vector $x$ into two parts so that $x = (x_a, x_b)$. Show that the conditional density $p(x_b|x_a)$ is itself a mixture distribution and find expressions for the mixing coefficients and component densities.

**Answer**

Using Bayes' theorem:

$$p(x_b|x_a) = \frac{p(x_a, x_b)}{p(x_a)}$$

Now, $p(x) = p(x_a, x_b)$

$$p(x_b|x_a) = \frac{\sum_{i=1}^{K} \pi_i p(x_a, x_b|i)}{\sum_{j=1}^{K} \pi_j p(x_a|j)}$$

Now, $p(x_a, x_b|i) = p(x_a|i) \cdot p(x_b|x_a, i)$

$$p(x_b|x_a) = \frac{\sum_{i=1}^{K} \pi_i p(x_a|i) \cdot p(x_b|x_a, i)}{\sum_{j=1}^{K} \pi_j p(x_a|j)}$$

Define a new mixing component, $\gamma_i(x_a) = \frac{\pi_i p(x_a|i)}{\sum_{j=1}^{K} \pi_j p(x_a|j)}$

$$p(x_b|x_a) = \sum_{i=1}^{K} \gamma_i(x_a) \cdot p(x_b|x_a, i)$$

This proves that $p(x_b|x_a)$ is also in itself a mixture distribution.

# Problem 6

**Question**

Consider a mixture of Gaussian distributions given by

$$p(x|\Theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

where:
$K$: Number of Gaussian components
$\pi_k$: mixing coefficients such that $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k > 0$
$\mathcal{N}(x|\mu_k, \Sigma_k)$: Gaussian density with mean $\mu_k$ and covariance $\Sigma_k$
$\Theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}$: represents the parameters of the model.

a Write down the complete log-likelihood function for a dataset $\{x_1, x_2 ... x_N\}$ assuming that the data points are drawn independently from the mixture model.

b Derive the Maximum Likelihood Estimation (MLE) update rules for $\pi_k, \mu_k$ and $\Sigma_k$, assuming that the component that generated each data point is known.

Answer

To write the complete log-likelihood function, we will first find the likelihood function, $L(\Theta)$. Probability of $x$ being modeled by $\Theta$ is the product of probabilities of each individual component $x_n$ being modeled by $\Theta$

$$p(x|\Theta) = \prod_{n=1}^{N}\sum_{k=1}^{K} p(x_n|\Theta) \implies L(\Theta) = \prod_{n=1}^{N}\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)$$

So our log-likelihood function becomes:

$$lnL(\Theta) = \sum_{n=1}^{N} ln\left(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)\right)$$

**MLE Update Rules:**
Since we're assuming we know which component each datapoint belongs to, we can introduce a latent variable, $z_{nk}$, where $z_{nk} = 1$ if the point $x_n$ was generated by component $k$, otherwise zero. Hence, our log likelihood can be written as:

$$lnL(\Theta) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \cdot ln\left(\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)\right)$$

$$\implies lnL(\Theta) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \cdot [ln\pi_k + ln\mathcal{N}(x_n|\mu_k, \Sigma_k)]$$

After performing Expectation Maximization on the log-liklihood function with latent variable introduced, we get the following update rules:

$$\pi_k = \frac{\sum_{n=1}^{N} z_{nk}}{N} \qquad \mu_k = \frac{\sum_{n=1}^{N} z_{nk} \cdot x_n}{\sum_{n=1}^{N} z_{nk}}$$

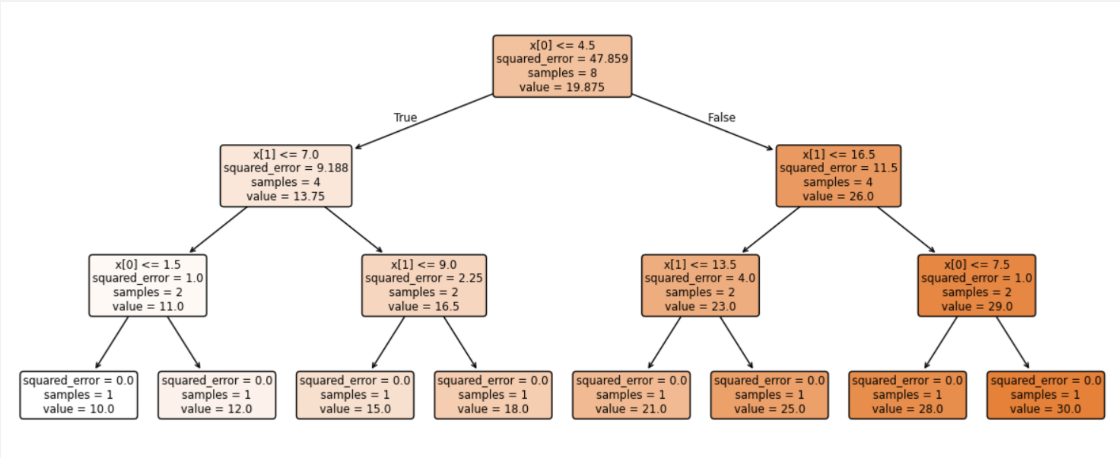
$$\Sigma_k = \frac{\sum_{n=1}^{N} z_{nk} \cdot (x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{n=1}^{N} z_{nk}}$$

# Problem 7

Question

Write a code to obtain a fully grown regression tree for the data given in Q2 and visualize the regression tree.

## Output




# Problem 8

## Question

Binary classification tree:

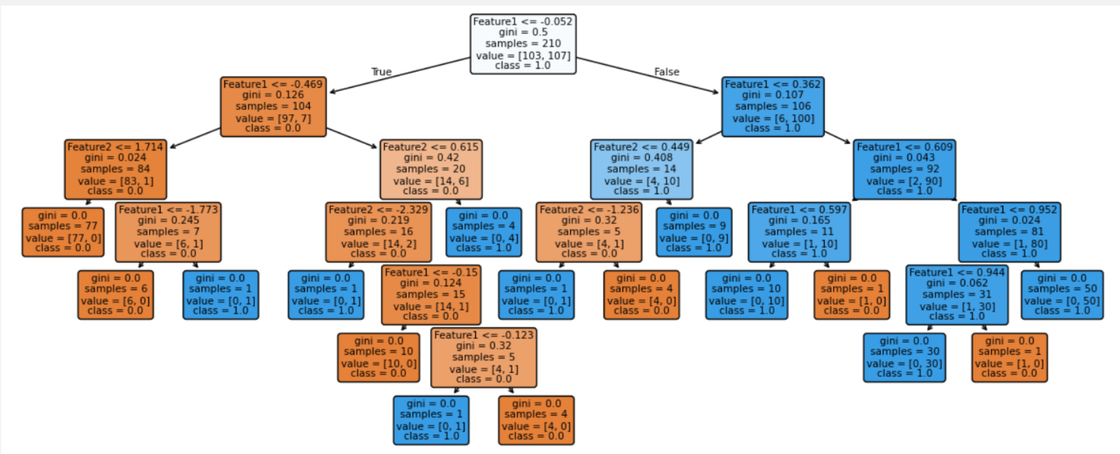
a Train a fully grown binary classification tree based on Gini impurity using the dataset A4-train.csv and visualize it.

b Compute the Sum of Squared Errors (SSE) on the test dataset (A4-test.csv) at each depth and plot the variation of SSE with depth.

c Determine the optimal pruning depth by selecting the depth where SSE change is minimal.

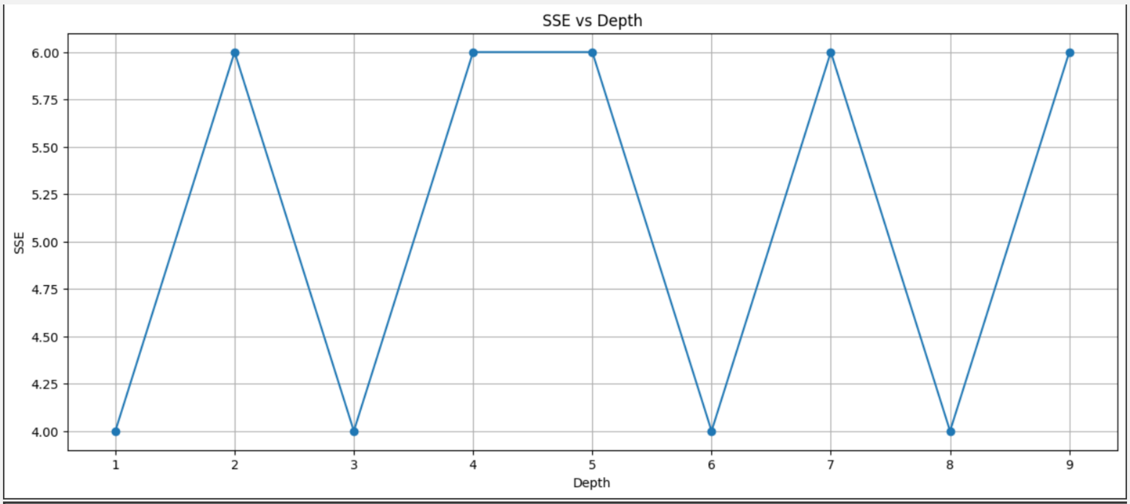
d Visualize the pruned tree.

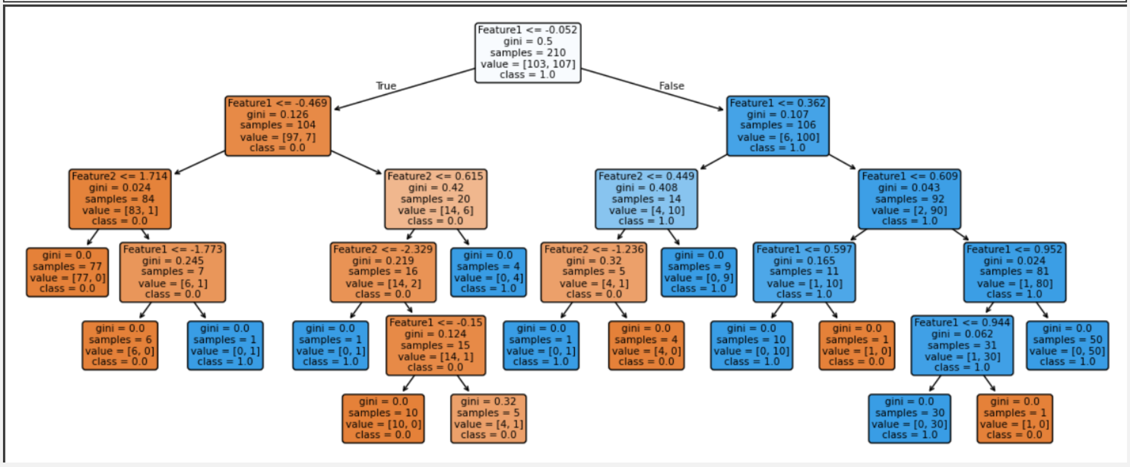
## Output




Initial depth: 6

## Ouput continued




Optimal depth: 5