Question 1: Regression (7pts)

Consider the following table that represents a dataset of houses based on their size (in square meters) and their price (in thousands of dollars):

#	Size (sq.m)	Price (in $1000s)
1	50	200
2	70	280
3	100	400
4	120	470
5	130	500

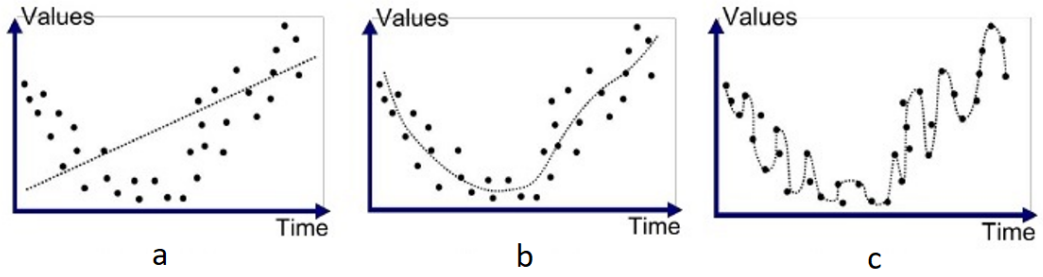
A) Describe the linear regression model and the primary objective when applying it to a dataset. (2pts)

B) Based on the linear regression equation $y = mx + b$ where $m = 4$ and $b = 50$, calculate the predicted prices for each house size in the dataset. (1pt)

C) Given the ground truth prices and the predicted prices you've calculated, compute the Mean Squared Error (MSE). (1pt)

D) Plot the ground truth prices alongside the predicted prices on a scatter plot. The x-axis should represent the house size, and the y-axis should represent the price. (1pt)

E) Refer to the figure below, which illustrates three regression models fitted to a dataset. Identify and label the models as (a), (b), and (c). Determine which one appears to be underfitted, which one is overfitted, and which one is a good fit. (2pt)



Question 2: K-Means Clustering on a 2D Plane (6pts)

Consider the following table that represents a dataset of fish based on their average length and width:

#	Length (cm)	Width (cm)
1	10	5
2	9	4.5
3	20	10
4	21	10.5
5	15	7
6	14	6.5

A) Explain the K-Means clustering algorithm. Briefly outline the steps required to perform clustering on the given dataset. (2pts)

B) For $k = 2$, explain the k-means++ method of initializing centroids. Based on this method, calculate the initial centroids for the provided dataset. (2pts)

C) Once you have determined the initial centroids using k-means++ and the clusters have been formed, a new fish with a length of 12 cm and width of 6 cm is introduced. Explain how you would determine to which cluster this new fish belongs. (1pt)

D) Observe the diagram below which illustrates a common method used to determine the optimal number of clusters in k-means clustering. Describe the method and explain how it?

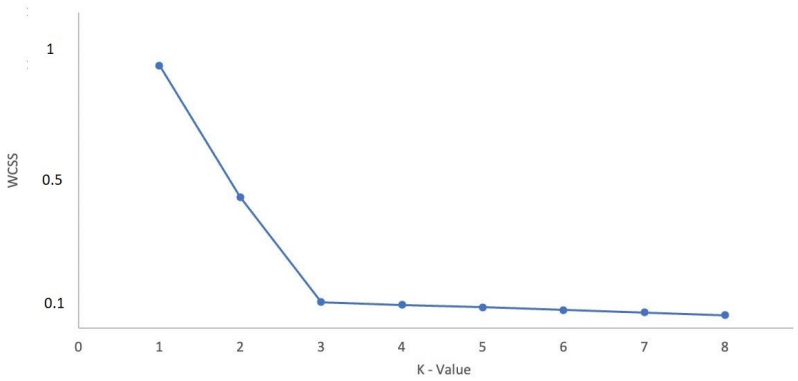


Figure 1: Sum of Squared Errors based on K

Based on the diagram, what is the best value for k and how many fish clusters will we have? (Hint: The answer is $k = 3$). (1pt)

A) Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. In its simplest form, with one independent variable, it is called simple linear regression. The relationship is expressed as:

$$y = mx + b$$

Where:

- y is the dependent variable.
- x is the independent variable.
- m is the slope.
- b is the y-intercept.

The primary objective of linear regression is to find the values of m and b such that the difference between observed values and values predicted by the model is minimized, typically using the method of least squares.

B) Using the equation $y = 4x + 50$:

- For size 50: $y = 4(50) + 50 = 250$
- For size 70: $y = 4(70) + 50 = 330$
- For size 100: $y = 4(100) + 50 = 450$
- For size 120: $y = 4(120) + 50 = 530$
- For size 130: $y = 4(130) + 50 = 570$

C) Mean Squared Error (MSE) is given by:

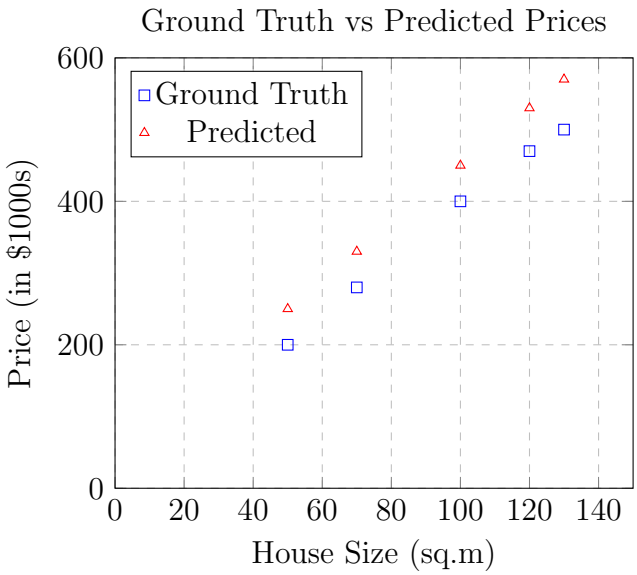
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Substituting the given values:

$$MSE = \frac{1}{5}[(200-250)^2+(280-330)^2+(400-450)^2+(470-530)^2+(500-570)^2]$$

$$MSE = 3200$$

D) Scatter plot of ground truth prices vs. predicted prices:


Ground Truth vs Predicted Prices
Ground Truth
Predicted
Price (in $1000s)
House Size (sq.m)
0
200
400
600
0
20
40
60
80
100
120
140

Answers

A) Kmeans Algorithm

- Create k points for starting centroids (e.g., randomly).
- While any point has changed cluster assignment:
  - For every point in our dataset:
    - For every centroid:
      - Calculate the distance between the centroid and point.
      - Assign the point to the cluster with the lowest distance.
    - For every cluster:
      - Calculate the mean of the points in that cluster.
      - Assign the centroid to the mean.

B) Kmeans++

Pseudo Algorithm [1]

1. Choose a point from the data (i.e., Forgy).
2. For each data point x not chosen yet, compute $D(x)$, the distance between x and the nearest center that has already been chosen.
3. Choose one new data point at random with a probability proportional to $D(x)^2$.
4. Repeat Steps 2 and 3 until k centers have been chosen.

Reference: https://en.wikipedia.org/wiki/K-means
Given data:

1. (10, 5)
2. (9, 4.5)
3. (20, 10)
4. (21, 10.5)
5. (15, 7)
6. (14, 6.5)

Step 1: Randomly select the first centroid. We choose **(10, 5)**.
Step 2: Compute the squared Euclidean distance from each data point to the nearest centroid.

- Point 2: 1.25

- Point 3: 125
- Point 4: 136.25
- Point 5: 29
- Point 6: 17.25

Step 3: Choose the next centroid with a probability proportional to the distance squared from the nearest centroid. We choose **(21, 10.5)**.

Our initialized centroids for k=2 are:

1. (10, 5)
2. (21, 10.5)

C)

To determine the cluster of the new fish:

1. Calculate the distance between the fish's data point (12, 6) and each of the centroids.
2. Assign the fish to the cluster of the nearest centroid based on the distance calculation.

D)

The described method is called the "Elbow Method". It involves running the k-means clustering on the dataset for a range of values of k, and then for each value of k compute the sum of squared distances from each point to its assigned center. When these overall dispersions are plotted against k values, the "elbow" of the curve represents an optimal value for k. The elbow point suggests that adding more clusters doesn't provide a significantly better fit to the data. Based on the hint provided, the best value for k is 3.