PROXIMITY/DISTANCE MEASURES-PART 3

Dr. Umarani Jayaraman
Assistant Professor

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY,
DESIGN AND MANUFACTURING,
KANCHEEPURAM

Topic:

- Mahalanobis distance

Quadratic form distance

$$d_Q(x, y) = \sqrt{(x - y)^T A (x - y)}$$

- For example $A_{ij}=1-c_{ij}/c_{max}$ for color histograms
- c_{ij} is bin-to-bin distance and c_{max} the maximum distance
- Note
  - If A is an identity matrix, then Euclidean
  - If A is a diagonal matric, then weighted Euclidean
  - If A is a inverse of covariance matrix, then Mahalanobis distance
  - **Is it a Metric? Yes , if A is positive definite**

Mahalanobis distance

- We know the quadratic form distance

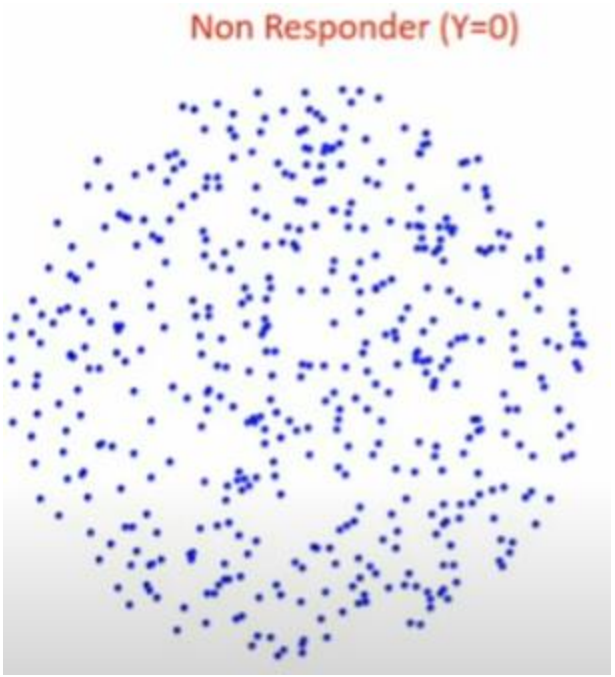
$$d_Q(x, y) = \sqrt{(x - y)^T A (x - y)}$$

- Replace A in quadratic form distance by inverse of covariance matrix $\sum$ to get **Mahalanobis distance**

$$d_M(x, y) = \sqrt{(x - y)^T \Sigma^{-1} (x - y)}$$

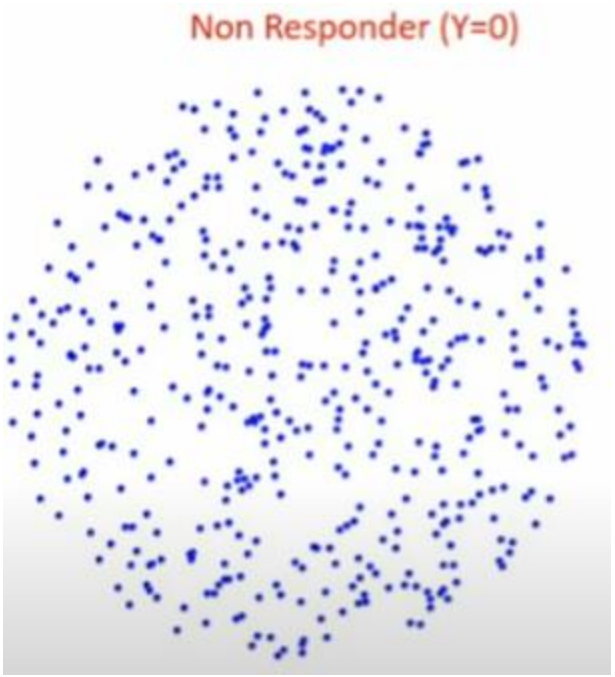
Issue of Euclidean distance

- Classification Problem



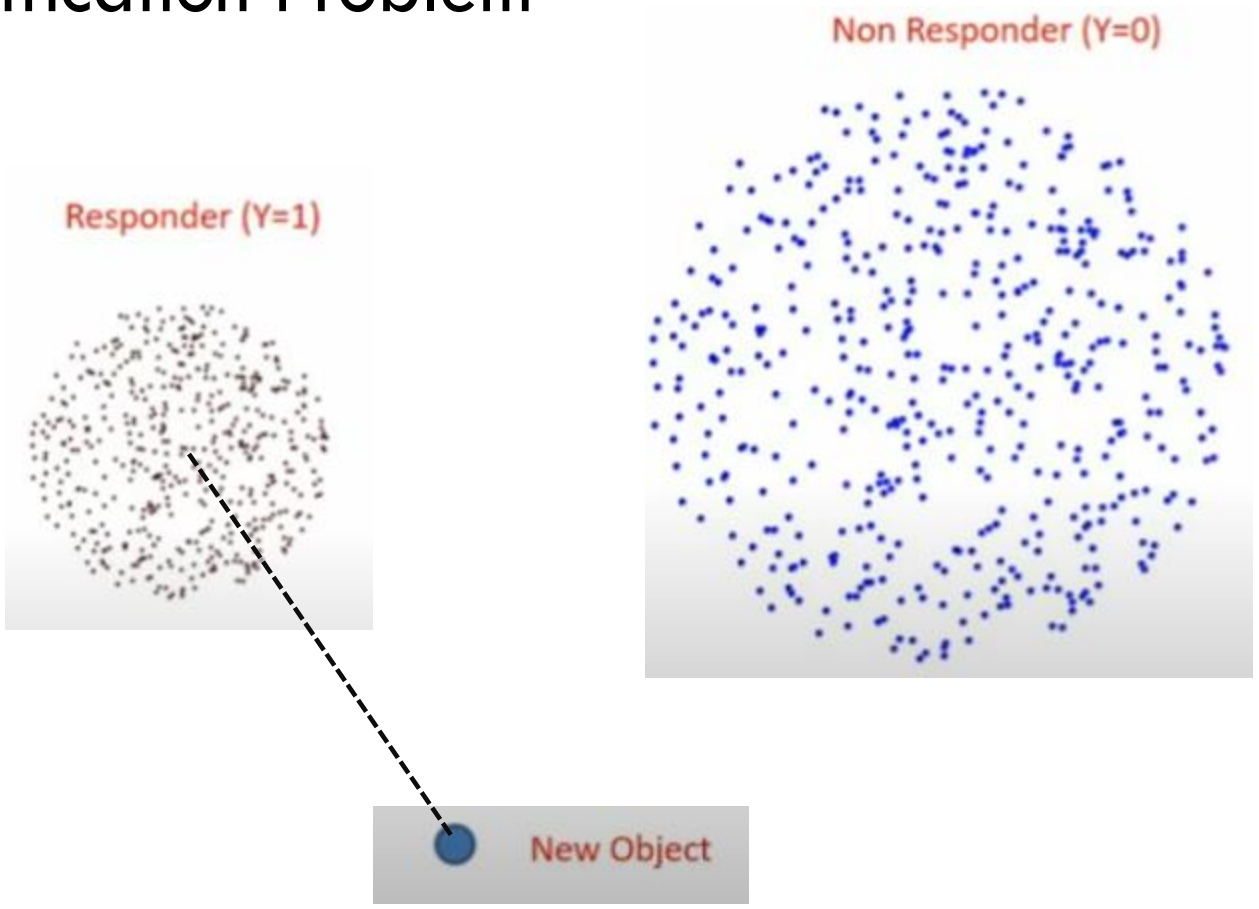
Issue of Euclidean distance

- Classification Problem



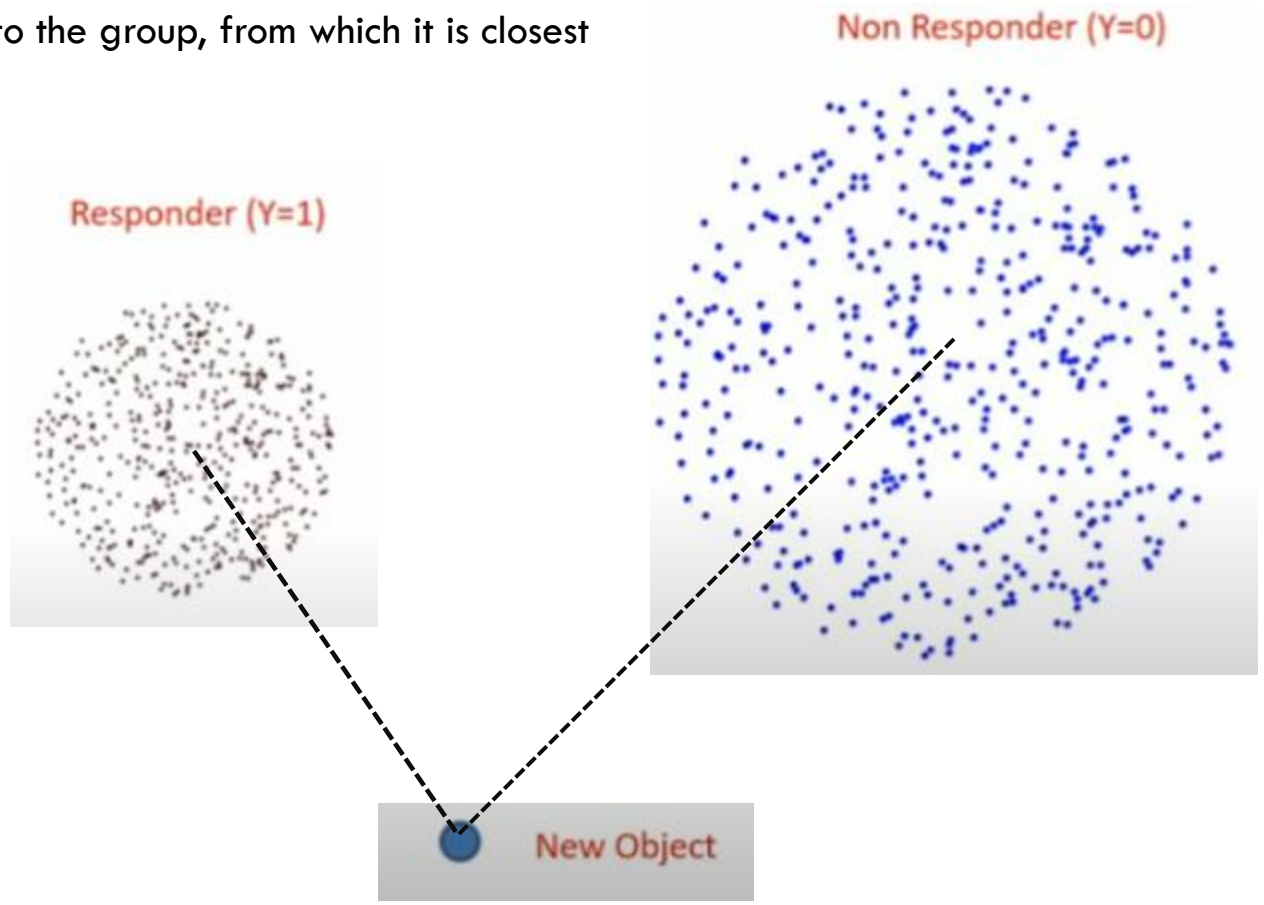
Issue of Euclidean distance

- Classification Problem



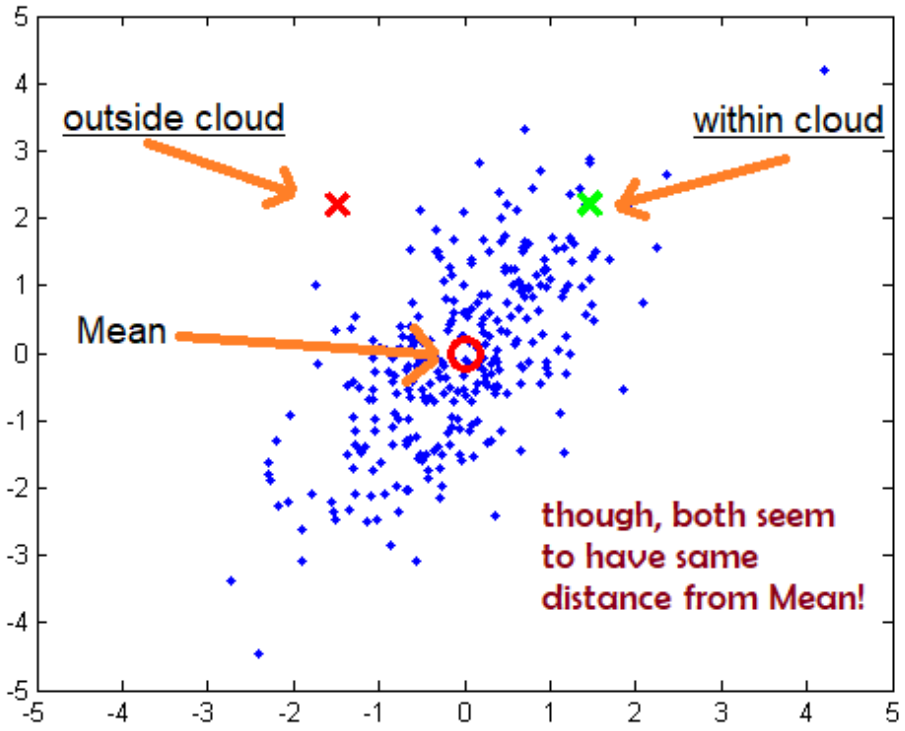
Issue of Euclidean distance

- Calculate distance of the new object from mean of different populations
- Assign it to the group, from which it is closest

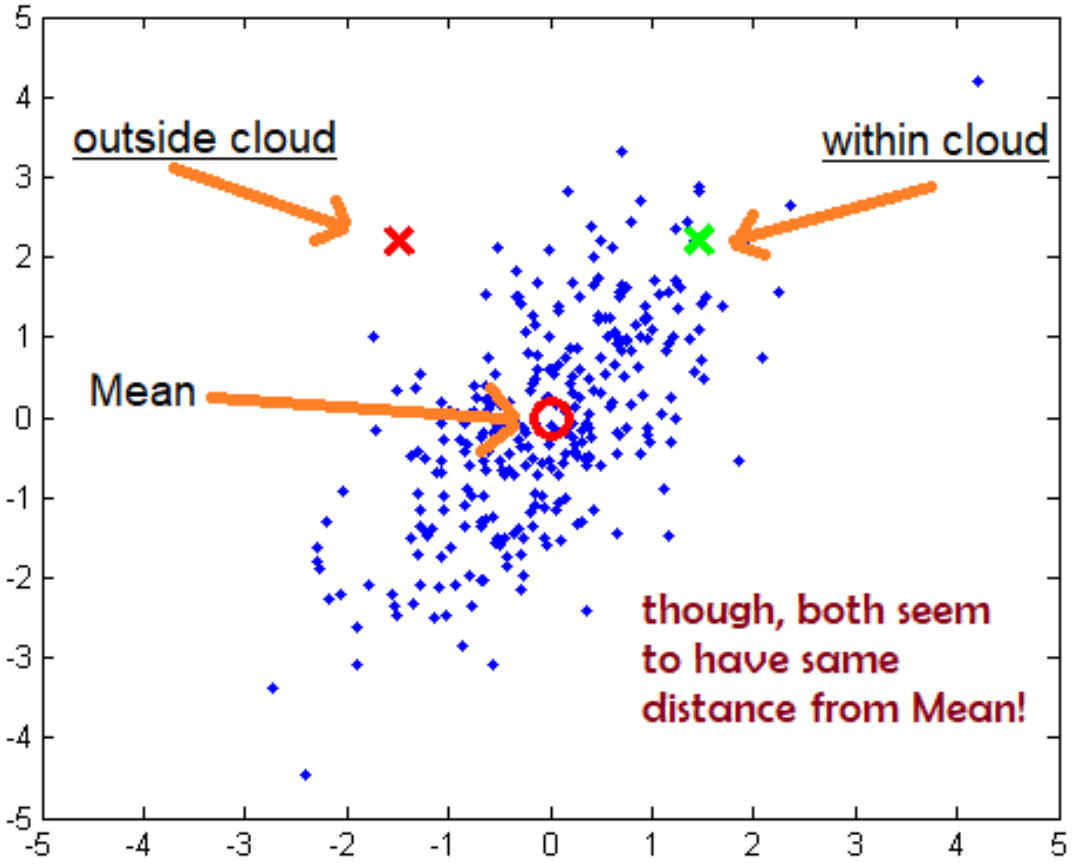


Issue of Euclidean distance

- It fails to capture the points which is outside the distribution as outliers

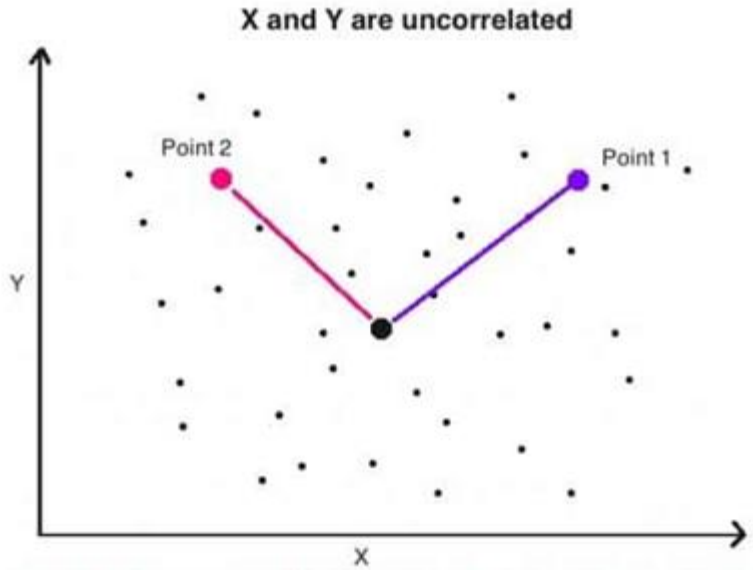


Intuition of MD: Finding Outliers



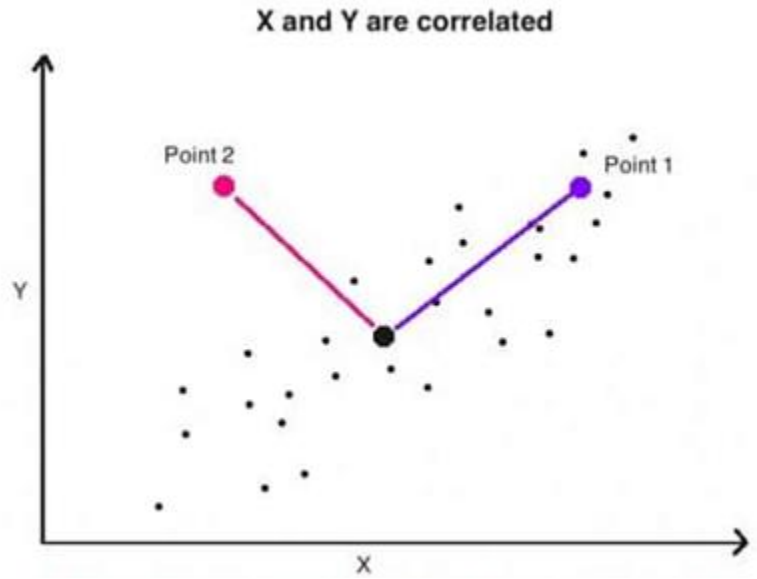
Intuition of MD

- When X an Y are uncorrelated, the Euclidean distance from the centroid can be useful to infer if a point is member of the distribution
- If the distance is lesser, then it is more likely a member of the distribution

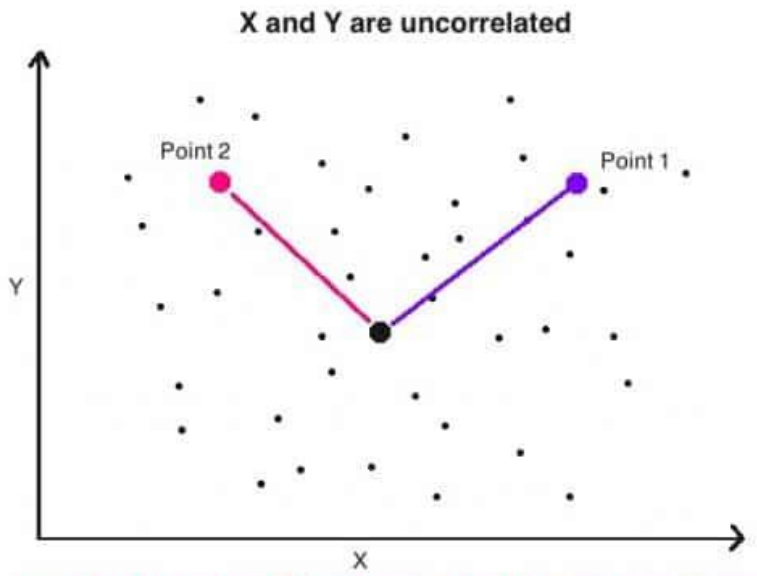


Intuition of MD

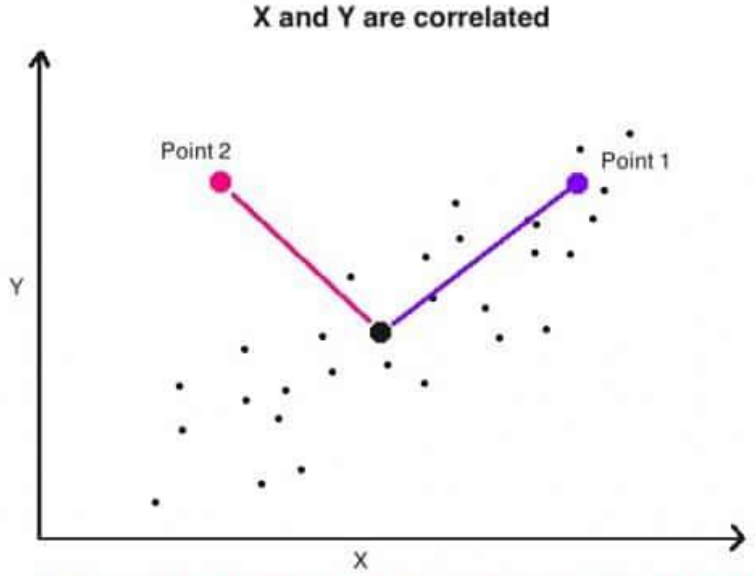
- Both point 1 and point 2 have the same Euclidean distance from centroid.
- But only point 1 is a member of the distribution
- To detect point 2 as outlier, dist (point 2, centroid) should be much higher than dist (point 1,centoid)
- Mahalanobis distance can be used here instead.



Intuition of MD



When X and Y are uncorrelated, the Euclidean distance from the Centroid can be useful to infer if a point is member of the distribution. The farther it is, the less likely it is a member.



Both Point 1 and Point 2 have the same Euclidean distance from centroid. But only Point 1 is a member of the distribution. To detect Point 2 as outlier, dist(Point 2, centroid) should be much higher than dist(Point 1, Centroid). Mahalanobis distance can be used here instead.

Mahalanobis distance

- Mahalanobis distance does the following
  - It transforms the variables into uncorrelated variables
  - It makes their variance equal to 1 (unit variance)
  - Then it calculates the simple Euclidean distance

Mahalanobis distance

- Normalized Euclidean
  - $x_i' = x_i - \mu_i / \sigma_i$ (i is the i^{th} component of the feature vector)
- Mahalanobis distance

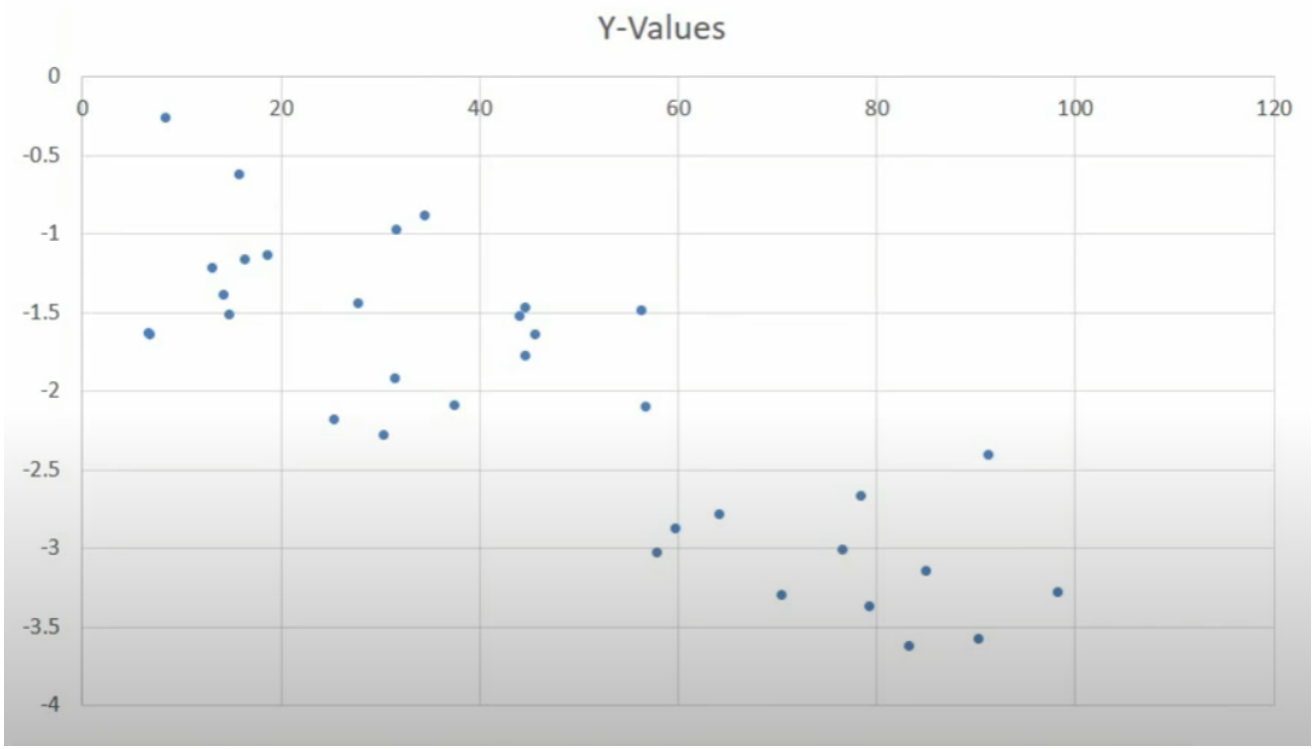
$$D^2 = (\mathbf{x} - \mathbf{m})^T \mathbf{C}^{-1} (\mathbf{x} - \mathbf{m})$$

Distance from mean

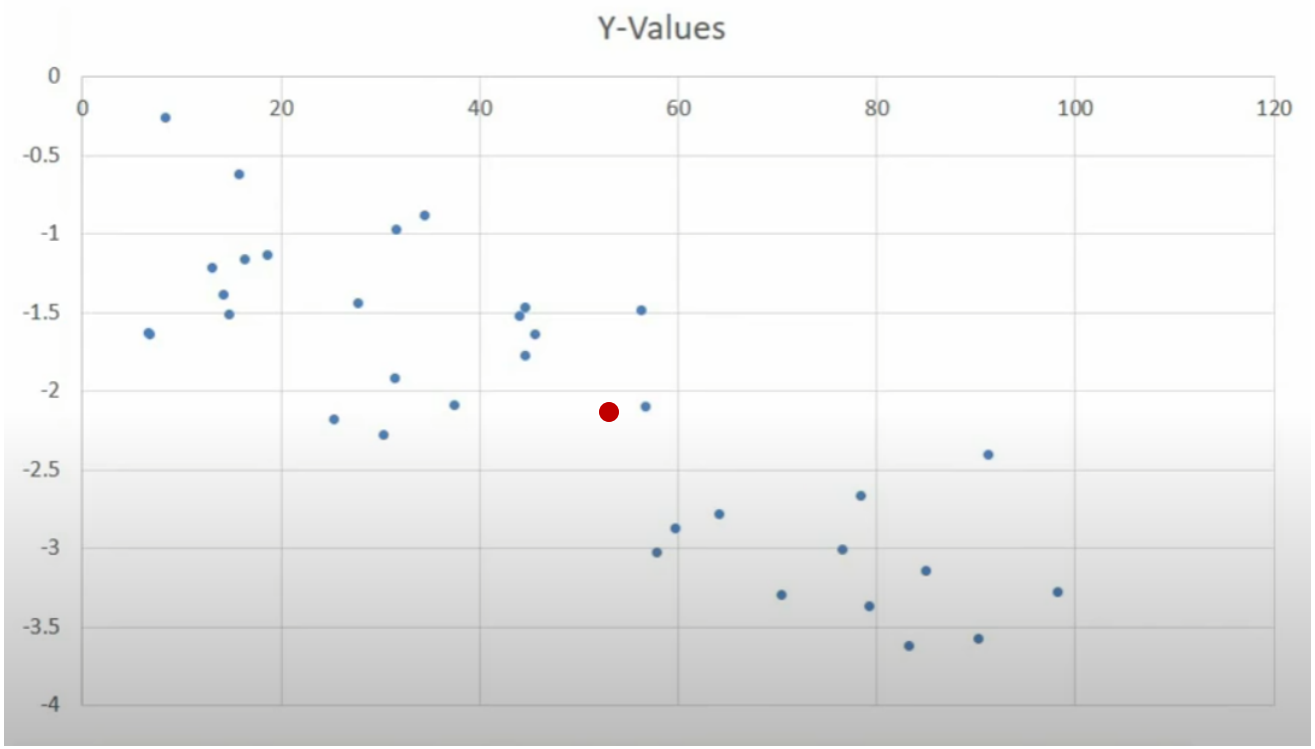
Divide by covariance matrix

- It actually does the normalization in multi variate before computing the Euclidean distance
- If the covariance is higher then we divide it by high value (covariance matrix)
- If the covariance is lesser then we divide it by low value (covariance matrix)

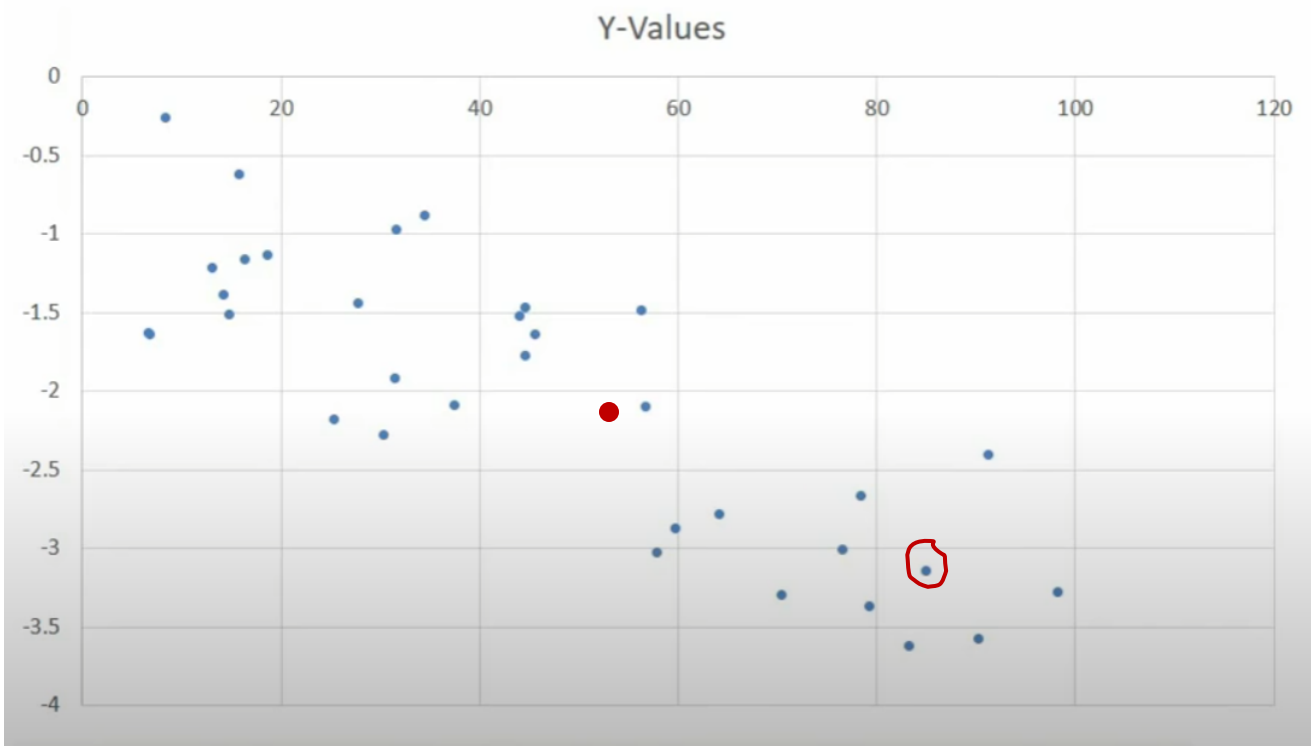
Graphic representation



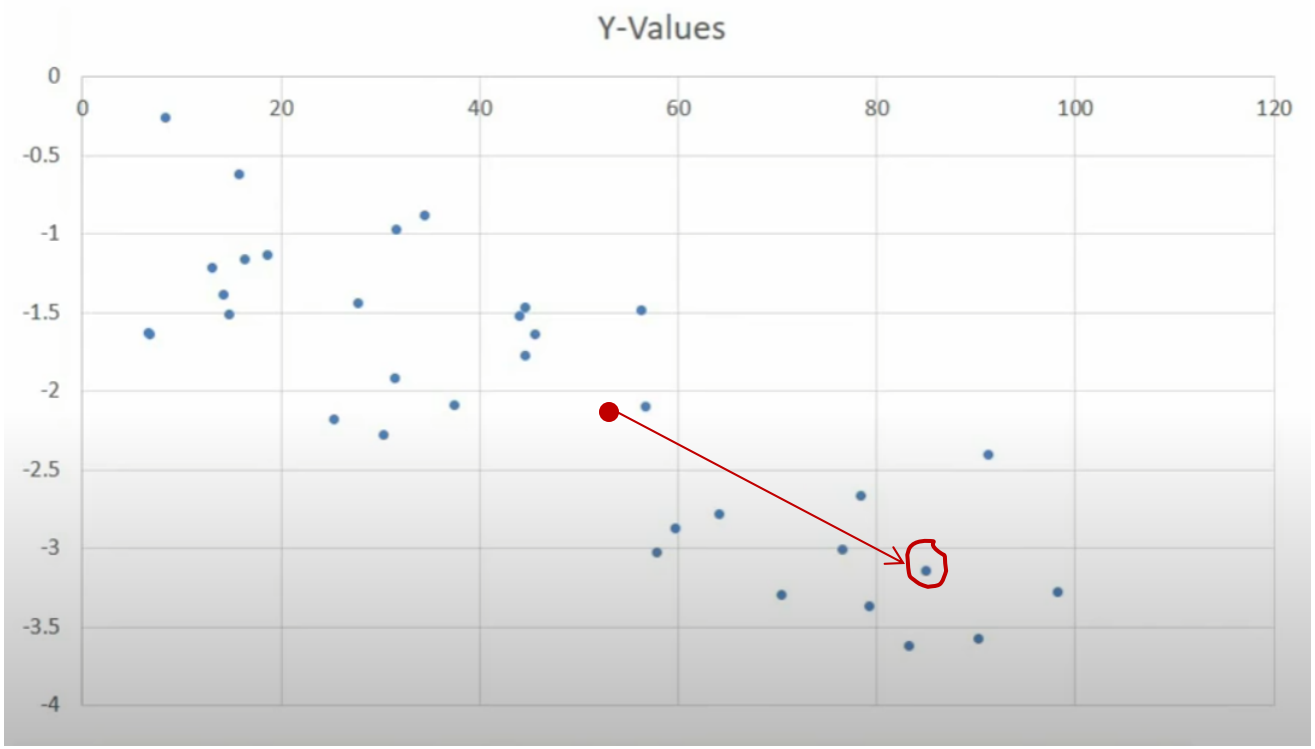
Graphic representation



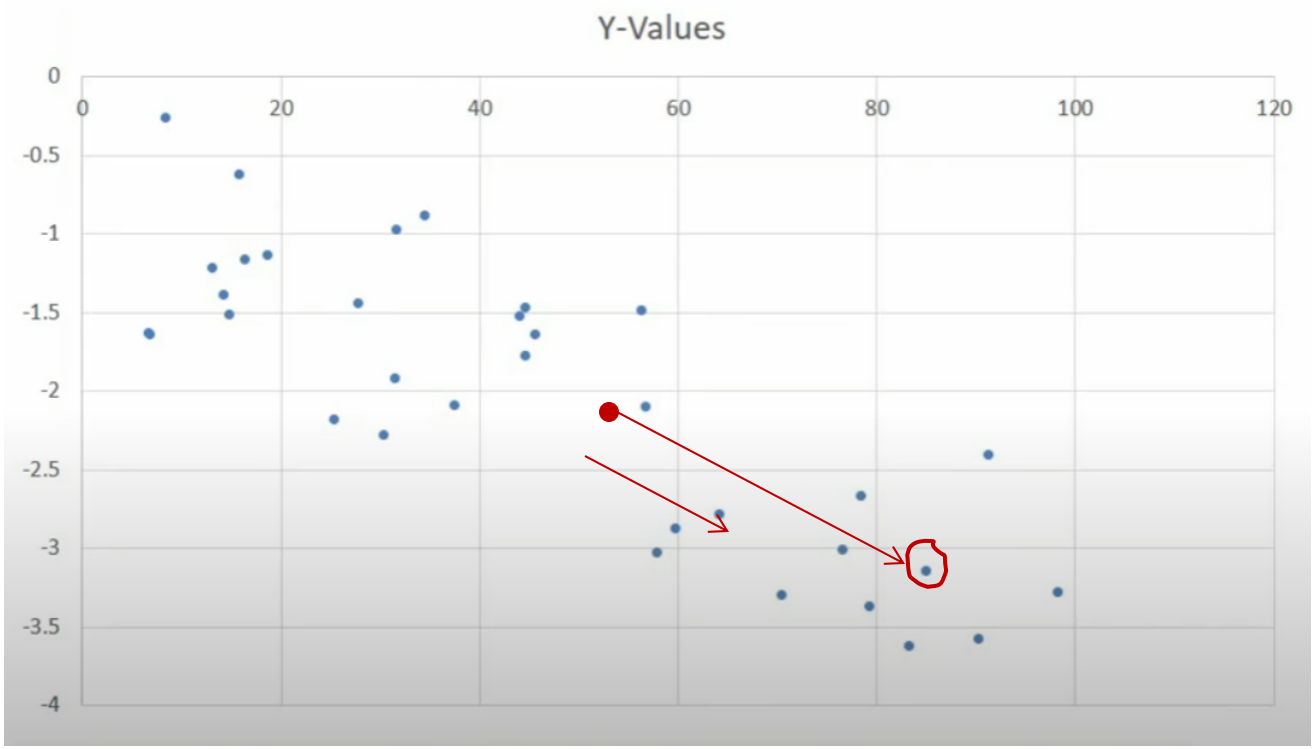
Graphic representation



Graphic representation

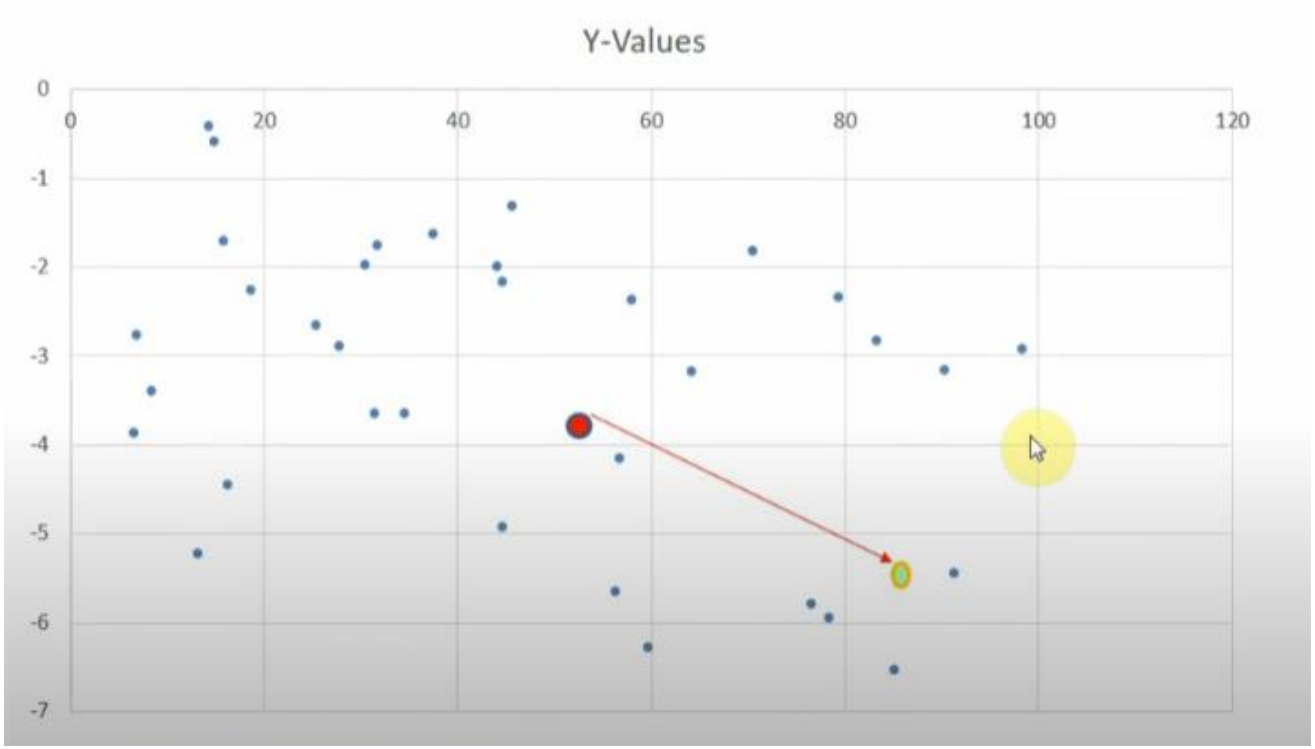


Graphic representation



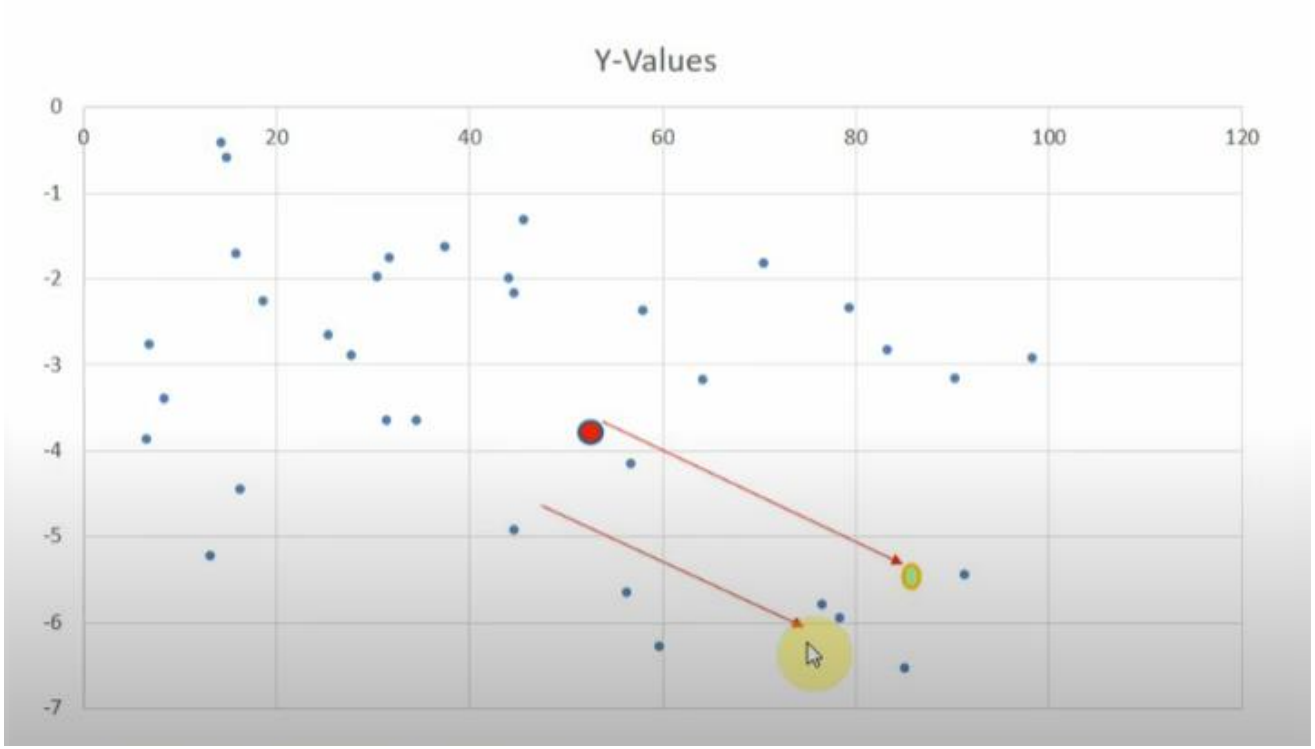
Graphic representation

- When data is uncorrelated



Graphic representation

- When data is uncorrelated



Summary

- Metric distance measure
- Mahalanobis distance
- Institution behind MD

THANK YOU