6.2.2022

Flowchart

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

1. Data Normalization
   1. Purpose – Scaling each feature to achieve comparable distances.
   2. Input – Raw tabular, numeric dataset
   3. Process - Min-Max scaler is used to normalize the data to range [0,1].
   4. Output – Tabular, numeric dataset, with values scaled to [0,1].
2. Distances Calculation
   1. Purpose – Acquiring a numeric measure, indicating how each class is different than the other, in each feature.
   2. Process – Calculating JM, Hellinger, or Wasserstein distances between classes in each feature's distribution.
   3. Output – Distances Matrices:
      1. Define F = number of features, C = number of classes
      2. F Matrices of size CxC, containing numerical values
3. Matrix Flattening
   1. Purpose – Pre-processing for diffusion matrix calculation
   2. Process – "flattening" the three dimensional distances matrices from shape FxCxC to two dimensional Fx
   3. Output – Flattened distances matrices
4. Diffusion Maps Calculation
   1. Purpose – Perform feature reduction on the flattened distances matrix, in order to select the features of interest.
   2. Process – Diffusion maps algorithm
   3. Outputs
      1. Eigen vector
      2. Eigen value
      3. Coordinates of following Eigen vectors
      4. Ranking – first Eigen vector
5. Feature Selection
   1. Purpose – Select the features which separate the classes most effectively
   2. Process –
      1. Top K features of first Eigen vector ("ranking")
      2. Ranked + K means using first two coordinates
      3. K mediods using first two coordinates
      4. Farthest features from 0 using first two coordinates
   3. Output – Dataset containing only the K selected features