UMAP

It is a dimensionality reduction technique, takes high dimensional data and projects it to a lower-dim space.

To perform UMAP:

1) Calculate probabilities of the points, i.e. one point j being a meaningful neighbours of i. Steps to do that:
   i) Calculate pairwise Euclidian distances
   ii) Select nearest neighbours.
   iii) Calculate probabilities
   iv) Symmetrize probabilities.

WORKFLOW

list of files $\xrightarrow{\text{embed}}$ list of file vectors $\xrightarrow[\text{for all}]{(N \times 5)\ \text{euclidian}}$ $(N \times N)$ distance matrix $\left(\sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}\right)$

$\downarrow$ k-nearest neighbours

reduced file vectors $(N \times \text{dim})$ $\xleftarrow{\text{lower dimensionality}}$ $(N \times N)$ probability matrix $\xleftarrow[\text{for k-nearest}]{\text{get probabilities}}$ {node: [(point, dist)...]}

2) Perform dimensionality reduction using obtained probability matrix.

$$\text{Loss} = \sum_{i,j} \left[p'_{ij} \log q_{ij} + (1 - p'_{ij}) \log (1 - q_{ij}) \right]$$

(p'_{ij}: high dim symmetric probabilities; q_{ij}: low dim probability)

where $q_{ij} = \dfrac{1}{1 + d_{ij}^2}$ (reduced dim distance)

$$\therefore \frac{\partial \text{loss}}{\partial \text{embedding of point } i} = \sum_j 2 \cdot (p'_{ij} - q_{ij}) \cdot \frac{x_i - x_j}{(1 + d_{ij}^2)}$$

Update embedding using sgd and above gradient.