cairo.thws
Center für Künstliche Intelligenz und Robotik
Technische Hochschule Würzburg-Schweinfurt

Fakultät Informatik und Wirtschaftsinformatik

Prof. Dr. rer. nat. habil.
Frank-Michael Schleif
Philipp Väth

date:

name(s) & matr. no.:

Exercises

Learning of structured data
– Sheet 2 –

The second portfolio sheet is about a real-world scenario, where you should leverage your knowledge related to proximity functions to embed and further process given data. All exercises should be completed in order and the solution consists of python code and answers to questions.

Exercise1: 'Data set'

Create a data set import script

- Load the protein sequence data set in the *.fasta* format. The numbers in the square brackets correspond to class labels.
  Hint: Search for the python package 'biopython'
- Utilize the Smith-Waterman algorithm to calculate an alignment score for two sequences
  Hint: There are a lot of pre-existing implementations, you don't have to implement it from scratch!

Exercise2: 'Embedding'

Use the loaded data to embed the protein sequences into a 100-dimensional vector space. Use the 100 largest eigenvalues for the embedding.

- Construct a pair-wise similarity matrix using the Smith-Waterman alignment
- Implement an embedding into a 100-dimensional euclidean vector space
- Hint: you are allowed to assume symmetry (which is almost true)

Exercise3: 'Evaluation'

- Evaluate the resulting embedding depending on your group (1-4,5-8 or 9-11) and provide a brief explanation of the algorithm you used. You may choose an algorithm yourself but we provide recommendations below.
- Try out different hyperparameter options, specifically for the Smith-Waterman alignment. You can also try different embedding dimensions and normalizations of the obtained vectors.

Group 1-4 (Classification) Classify the protein sequences and calculate the accuracy score w.r.t. the given class labels.
Recommendation: GMLVQ (click). Provide a full modeling and evaluation workflow. Other options are e.g. k-NN, Support Vector Machine, Random Forest

Group 5-8 (Clustering) Cluster the data, use an evaluation measure like Normalize Mutual Information (NMI) and provide a visualization e.g. by employing the proximity matrix. Save the plot.
Recommendation: Affinity Propagation (click). Provide a full modeling and evaluation workflow. Also possible k-means or so.

Group 9-11 (Embedding) Visualize the data in 2-D and color the points / nodes based on the label assignments. Please save the plot.
Recommendation: U-Map (click). Provide a full modeling and evaluation workflow. Modify some of the meta-parameters of the algorithm in a meaningful way. Alternative t-SNE, Fast-Multi Scale Neighbor Embedding (click).

submit:

-