Shiraz University

Neural Networks and Deep Learning (Spring 2021)

Homework #2: Kohonen Self-Organizing Map (SOM)

Due date: 24th Ordibehesht 1400

In this homework, you have to train and test an SOM network to do cluster analysis of a news collection, from the BBC news website corresponding to stories in five topical areas from 2004-2005. This dataset is a collection of 2225 news document, categorized into 5 classes of ‘business’, ‘entertainment’, ‘politics’, ‘sport’, and ‘tech’.

Text Clustering using SOM

Text clustering is an unsupervised process, used to separate a document collection into some clusters on the basis of the similarity relationship between documents in the collection. Suppose $D = \{d_1, \dots, d_N\}$ be a collection of N documents to be clustered. The task is to divide D into k clusters $C_1, \dots, C_k$ where $C_1 \cup \dots \cup C_k = D$ and $C_i \cap C_j = \emptyset$, for $i \neq j$.

SOM text clustering can be done in two main phases. The first phase is document preprocessing, which uses Vector Space Model (VSM) to generate a numeric vector for each text document. In the next phase, SOM is applied on the document vectors to obtain document clusters.

Phase 1: Document Preprocessing

By means of VSM, each document d_i can be represented by an n-dimensional feature vector $\boldsymbol{v}_i =< v_{i1}, \dots, v_{in} >$, where v_{ij} is a representation of term t_j in document d_i and n is the number of distinct terms in the document collection D.

An approach for computing v_{ij} is the Term Frequency - Inverse Document Frequency (TF-IDF) weighting scheme. This method computes v_{ij} for term t_j in document d_i as:

$$v_{ij} = \log(1 + tf_{ij}) \times \log(\frac{N}{df_j})$$

where tf_{ij} is the frequency of term t_j in document d_i, and df_j is the number of documents in D containing term t_j.

Read ‘bbc-text.csv’ file and for each document:

1. Remove all non-letter characters from the documents.

2. Extract all words of the document and remove the short words (length ≤ 2).
3. Remove all stop words (e.g., ‘a’, ‘and’, ‘what’, …), given in file ‘stopwords.txt’.
4. Compute the feature vector for each document, using TF-IDF weighting scheme.

Phase 2: SOM Clustering

a) Winner-takes-all approach

1. Using all documents, build an SOM with one neuron for each class.
2. Depict the SOM-hits plot.
3. Compute and report the confusion matrix.

b) On-center, off-surround approach

1. Using all documents, build an SOM with 3×3 neurons.
2. Depict the SOM-hits plot.
3. Compute the Euclidean distance of all documents to their winner neurons and sum up the distances.
4. Repeat steps 1-3 for 4×4 and 5×5 topologies.
5. Report and discuss the overall distances of three topologies.

Notes:

- Pay extra attention to the due date. It will not extend.
- Be advised that submissions after the deadline would not grade.
- Prepare your full report in PDF format and include the figures and results.
- You can use any library for SOM in Matlab or Python.
- Submit your assignment using a zipped file with the name of “StdNum_FirstName_LastName.zip” to soroushmehrpou@gmail.com with "NNDL-Spring 2021-HW#2" subject.