Assignment-IV

NOTE: Both Questions are to be implemented in RAPIDS Container using cudf, cupy,a nd cuml

Q1: (Regularization – Ridge Regression using LSE Fit)

I. Generate a dataset with two columns x (input variable) and y (output variable). The values of x are angles in radians from 60° to 360° with a step size of 4°. y values are sine of angle x and add some random noise to these values (using random.normal).
II. Split the dataset in train and test; fit a ridge regression using least square error fit on the train set using a particular value of regularization parameter (λ).
III. Tune the value of λ using concept of validation sets i.e., Divide the train set further into train and validation sets and consider different values of λ. For each value of λ, fit a ridge regressor (using LSE) on train set and test its performance on validation set. Choose the value of λ which gave best performance on validation set.
IV. For the best value of λ, train the ridge regressor on the entire training set (train + validation set) and test the performance on test set.
V. Use the inbuilt function for Ridge Regression in cuml to fit and predict the values of output variable using the best value of λ.

Q2: Based on PPMI matrix

Download the IMDB Movie Review dataset from the following link:

https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

IMDB dataset having 50K movie reviews for natural language processing or Text analytics. This is a dataset for binary sentiment classification containing substantially more data than previous benchmark datasets.

I. Load the dataset in a DataFrame.
II. Drop the sentiment column and consider the first 1000 reviews
III. Make the corpus of first 1000 reviews.
IV. Convert the corpus into binary BOW vector of size mXn, where m=1000 (number of reviews documents) and n is the number of unique terms obtained from the 1000 documents. Each ijth entry of the vector is a binary value which is 1 if the jth term is present in ith review else 0. (Without using in-built function)
V. Compute the co-occurrence matrix of order nXn where each ijth entry of matrix is number of documents in which both ith and jth terms co-occur. (Use binary BOW vector to compute it).
VI. Compute PPMI matrix where PPMI between two words a and b is given by:

$$PPMI(a,b) = max\left(log\left(\frac{n(a,b) * |D|}{n(a) * n(b)}\right), 0\right)$$

where n(a,b) is the number of documents in which both words a and b co-occur (from co-occurrence matrix), n(a) and n(b) is number of documents in which terms a and b occur respectively (from BOW vector); |D| is total number of documents (=1000 in this case).