

Graph learning code: <https://github.com/alikhz1376/LLFG/tree/master>

VGAE architecture

N = Number of nodes

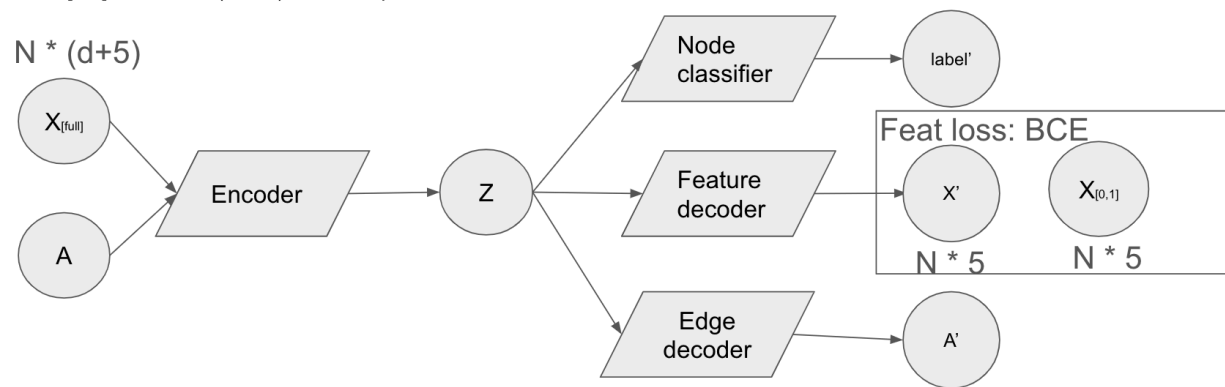
d = feature dimension

5 = number of selected important features

$X_{[full]}$ = has all the original features from PyGeo + binarized(0 or 1) version of the important features

X' = Reconstructed X

$X_{[0,1]}$ = binarized(0 or 1) 5 most important features



VGAE framework setting:

```
Namespace(epoch_number=101, Vis_step=20, lr=0.001, dataset='citeseer', hemogenize=False,
num_of_comunities=64, encoder_layers='64', use_feature=True, DropOut_rate=0,
batch_norm=True, split_the_data_to_train_test=True, decoder_type='MultiRelational_SBM',
encoder_type='RGCN_Encoder', num_node=-1, downstreamTasks={'linkPrediction',
'nodeClassification'})
```

Datasets:

Dataset	#Nodes	#edges	#node types	#edge types	#features	#Labels
ACM (Paper, author, subject)	8,993	18,929	3	2	1902	3 (the conferences the papers published)

<u>IMDB</u> (actor, movie, director)	11,616	19,120	3	2	3066	3 (genres of movies)
<u>Cora</u> (papers)	2,708	5,429	1	1	1,433	7(subject of papers)
<u>Citeseer</u> (papers)	3,327	4,732	1	1	3,703	6(publications topic)
<u>Photos</u> (co purchased photos)	7,650	238,162	1	1	745	8(product category)
<u>Computers</u> (co purchased computers)	13,752	491,722	1	1	767	10(product category)

You can find the sql version of these datasets on:

```
mysql -h database-1.cxcqxpvnbnwo.us-east-2.rds.amazonaws.com -P 3306 -uadmin -pnewPassword
```

The preprocessing steps for these datasets are found at:

<https://github.com/alikhz1376/LLFG/blob/master/preprocess.py>

And

https://github.com/alikhz1376/LLFG/blob/master/imdb_preprocess.py

Configurations for running factor base([config.cfg](#)):

```
# Database Configurations
dbaddress = mysql://database-1.cxcqxpvnbnwo.us-east-2.rds.amazonaws.com
dbname = cora
dbusername = admin
dbpassword = newPassword
```

```
dbtemporarytablesize = 4
dbcollation = latin1_swedish_ci

# FactorBase Configurations
AutomaticSetup = 1
ComputeKLD = 0
Continuous = 0
LinkAnalysis = 0
LinkCorrelations = 0
UseLocal_CT = 0
SkipParameterLearning = 0

# Counting strategy for generating the contingency tables.
# 0 - Precount
# 1 - Ondemand
# 2 - Hybrid
CountingStrategy = 0

# Logging levels for console output.
# off - No console output.
# info - Basic console output.
# runtimeDetails - Additional runtime information.
# debug - Detailed information.
LoggingLevel = debug
```

Steps of running the code:

1) Run preprocess/imdb_preprocess.py to create the sql version of the database with the 5 most important features

2) Run factorbase using:

```
java -Dconfig=config.cfg -jar -Xmx64G factorbase-1.0-SNAPSHOT-parmis.jar
```

For more information on factorbase visit [factorbase git](#) and [other instructions](#)

3) Optional BN files: If you want to see the bayesian network generated from factorbase you can use the UBC tool:

Download the tool from [here](#)

And the run:

```
java -jar bayes.jar
```

Then go click file/load from file/ and then select the generated bif_****.xml that the factorbase has outputted. I have attached a list of BN files that I got from factorbase [here](#)

4) Run [rule learning](#):

Python RuleLearning.py