GRAPH CONVOLUTIONAL NETWORKS

CMU 11441/11641/11741: ML FOR TEXT & GRAPH MINING
Due date: 12/3/2021, 11:59 PM EST

Instructions

- Allowed libraries: This assignment involves implementing graph convolutional networks. You are not allowed to use any libraries that implement GCNs out of the box (like Pytorch-geometric). It is allowed to use autodiff libraries like Pytorch/Tensorflow.
  We highly recommend using Python + Pytorch for this assignment.
- Statement of Assurance
  1. Did you receive any help whatsoever from anyone in solving this assignment? No
  2. Did you give any help whatsoever to anyone in solving this assignment? No
  3. Did you find or come across code that implements any part of this assignment? No

1 GCN Review (30 points)

Q1. What is the big-O time complexity of the computation expressed in Equation ?? in terms of $|V|$, $|E|$, d, k, and L? Your expression should not contain any other term. Assume $d < k$.

- We need to do computation for L layers
- Each computation can be seen as:
  - Matrix product (O(k^2)), which needs to be done O(|E|/|V|) times (average number of neighbors) for each node : O(|V| |E|/|V| k^2) = O(|E| k^2).
  - Also need to apply σ, which is O(|V|)
- This totals: $O(L \times (|E| \times k^2 + |V|))$

Q2. What is the space complexity of the computation expressed in Equation ?? in terms of $|V|$, $|E|$, d, k, and L (assume intermediate terms are saved)? Your expression should not contain any other term.

Node features (final only): $O(|V| \times k)$

Considering the Intermediate computations (L Wh products): $O(L \times |V| \times k)$

Weight matrices: $O(L \times k^2)$

Adjacency list for neighbors: O(|E|)

Total: $O(L \times k \times (|V| + k) + |E|)$

2 Graph Exploration (20 points)

Graph	Karate	Cora	Citeseer
Max in-degree	18	169	100
Min in-degree	2	2	1
Average in-degree	5.58	4.89	3.73
# nodes	34	2708	3312
# edges	190	13264	12384
Node feature dim	34	1433	3703

Table 1: Graph statistics

3 Node classification

3.1 Implementation (60 points)

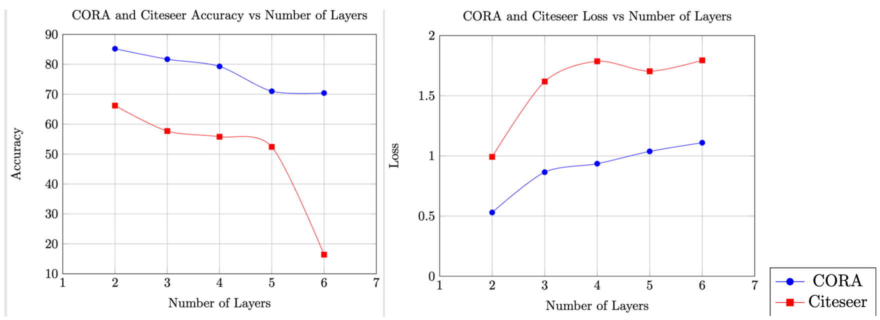
Graph	Accuracy %	Loss
KARATE	100	0
CORA	85.2	0.529
CITESEER	66.2	0.992

Table 2: Node classification results

3.2 Varying L (20 points)

For both CORA and CITESEER, modify the GNN to include *L*= 3,4,5,6 layers and plot the loss and accuracy vs. *L*. Summarize your observations in 2-3 lines.

Test Accuracy and Loss plots for CORA and Citeseer



In general, accuracy lowers and loss increases with L. This may indicate that more complex models are harder to train, hence more epochs or more data could benefit models with larger L. Looking at train VS validation loss/acc in the std output, larger models had an harder time generalizing, motivating harder regularization.

3.3 Topological features vs. inbuilt features (20 points)

Graph	Accuracy %	Loss
Cora	85.2	0.529
Cora_topo	41.8	1.553
Cora_plus_topo	66.0	0.988
Citeseer	66.2	0.992
Citeseer_topo	25.7	1.756
Citeseer_plus_topo	63.6	1.144

Table 2: Effect of topological features

Looking at the table, topological features alone have weak performance. Combining them with the original features did not improve the results. This second setup may require more testing and hyperparameter tuning as the input features now have a higher dimension.

4 Link prediction

4.1 Training data for link prediction (20 points)

A.

Graph	# Positive edges	# Negative edges
KARATE	190	190
CORA	13264	13264
CITESEER	12384	12384

Table 3: Training data statistic for link prediction

B. How is the training data for link prediction created? Please explain in 2-3 lines.

First, split the set of all edges (vi, vj) into train, validation, and test splits. All the edges present in the training set are naturally considered positive instances. Negatives are sampled by randomly shuffling the nodes of the positive edges.

4.2 Implementation (80 points)

Graph	Accuracy %	Loss
KARATE	51.34	1.008
CORA	93.25	0.179
CITESEER	94.95	0.152

Table 4: Link Prediction Results

5 Graph classification

5.1 Graph Statistics (10 points)

Graph	MUTAG	ENZYMES
Num graphs	141	360
Avg. num nodes	18.85	33.27
Avg. num edges	94.04	221.19
Node feature dim	8	22

Table 5: Graph statistics for the graph classification datasets

5.2 Implementation (90 points)

Graph	MUTAG			ENZYMES		
	P	R	F1	P	R	F1
Mean-pooling	63	76	64	34	37	33
Max-pooling	84	83	83	37	42	37
Last-node pooling	71	77	73	35	36	35

Table 6: Graph classification results. Please use macro-averages to report the precision, recall, and F1 score for ENZYMES.

References