CS 768

Scribbing Assignment - Lecture 3

1 Necessity for Train Test split

In Link Prediction our task is to rank the non-edges as the graph evolves. In most cases we only get a snapshot a graph and we can't wait for the graph to evolve to see which turns out to be edges to evaluate the model we had trained. Therefore we have to evaluate the Link Predictor using the current snapshot of the graph itself and to do this we split the graph into test and train edges and non edges.

2 Splitting into train and test

To obtain a train and test set from the given graph we we choose a fraction of edges from graph in training set and the rest in the test set and similarly for the non edges. Non edges are also split into train and test so that we don't use any information about the non edges that we are going to evaluate against later. Therefore while training the link predictor using the training set we have to make sure that we use only the labels of the non edges other than the ones in test set and not use any information of the no edges in test set.

It is not possible to not use information form the test set. For example when we construct features for each node f_u to train the LP we have to know the relationship between 2 nodes and thus we have to incorporate the fact that it is not an edge and thus information leaks from test set. But since the fraction of non edges in test set is much smaller than that is training set we can neglect the effects caused by this in features f_u.

3 Applications of Aggregrate Sampling

- Predicting which two authors are going to collaborate in a particular topic from citation graphs.
- Predicting which k connections are potentially lethal in a social network.
- Understanding reaction mechanisms.
- Predicting interactions in Protein Protein Interaction (PPO) network

4 Applications of node specific sampling

- Used when we want to do well across all nodes
- Recommending k users to an user to send friend requests in Facebook
- Recommending m movies to an user in Netflix