Lecture 6 Summary - CS 768

Abhinav Kumar (180050003)

September 2020

1 Introduction to Link Prediction Methods

1.1 Unsupervised Learning Based

This subset of prediction techniques includes heurisitic based methods etc. which have no updation of scores unlike supervised methods i.e. same score will be obtained on multiple running of the algorithm. e.g. Katz measure, Adamic-Adar

1.2 Supervised Learning Methods

This includes techniques using which prediction scores are refined over multiple iterations of the model on the dataset. At high level, it involves designing features automatically/ manually which then serves as the dataset for training of the model and the scores are usually 0-1 based denoting whether there is an edge or non-edge for link-prediction task.

2 Unsupervised Methods

2.1 Preferential Attachment

This heurisitic is based on the idea that if two nodes or even one odd have high degree then they have higher chance of forming a link in the future.
The score for edge between nodes u and v is given by-

$$s(u, v) = d_{\mathrm{u}} * d_{\mathrm{v}} \tag{1}$$

Another possible scoring heuristic based on the extrovertness of the node is

$$s(u, v) = d_{\mathrm{u}} + d_{\mathrm{v}} \tag{2}$$

But in most of the practical graphs, degree of nodes usually follow power law which is achieved when (1) is used hence (1) is known as preferential attachment measure.

2.2 Common Neighbour

This heuristic proposes that more the number of common nodes between node u and node v, more likely is an edge to appear in the future between them. The score for edge between node u and node v is given by-

$$s(u, v) = |N(u) \cap N(v)| \tag{3}$$

where N(u) is the set of neighbours of node u

But the above score gives us the impression that more the common neighbours among 2 nodes, higher will be the latent similarity which may not be true. For e.g. if 2 nodes denote 2 person having affinity to cricket then they will have many common neighbours related to the field of cricket namely players etc. This heuristic is although found to perform well on datasets where very low number of nodes are celebrity kind.

The above discussion calls for a good measure of celebritiness of a node.

Some proxies of celebritiness of a node are-
a) *Degree of a node*- Higher the degree of a node, more likely the concerned node is celebrity.

Using above measure, we can put some threshold on the degree of a node which will decide whether the node is celebrity or not. Common Neighbour measure can be modified as follows

$$s(u, v) = |N(u) \cap N(v) \cap \{II(d_u < d^*)\}| \tag{4}$$

where d* is the threshold.

2.3 Resource Allocation

Continuing above idea of celebritiness of a node, another heuristic for score of edge can be

$$s(u, v) = \sum_{w \epsilon N(u) \cap N(v)} \frac{1}{c(w)} \tag{5}$$

where c(w) is the celebrity score of node w

2.4 Adamic-Adar

Similar in spirit to resource allcation measure-

$$s(u, v) = \sum_{w \epsilon N(u) \cap N(v)} \frac{1}{\log(d(w))} \tag{6}$$

where d(w) is the degree of node w

2.5 Jacquard Coefficient

The heuristics discussed above are much sensitive to slight change in the degree of nodes.
A relatively more stable scoring heuristics is-

$$s(u, v) = \frac{|N(u) \cap N(v)|}{(u) \cup N(v)|} \tag{7}$$

Above discussed heuristics capture signals only from atmax 2 hops distant nodes (1 hop for CN and JC, 2 hop for AA and RA). So these methods work good only in the dense parts of the graphs.

2.6 Katz Measure

$$s(u, v) = \sum_{l=1}^{\infty} \beta^l * A_{uv}^l \tag{8}$$

$$S = \sum_{l=1}^{\infty} \beta^l A^l \tag{9}$$

where S is the score matrix whose entry at u,v is equal to s(u,v) and A is the adjacency matrix of the graph

The above sum converges iff $\beta < \rho(A)$ where $\rho(A)$ is the spectral radius of A
If above condition is satisfied then closed form equivalent of (9) is

$$S = [(I - \beta A)^{-1} - I] \tag{10}$$