Data Mining:
Final Stage Report

Yimeng Lu • 12.05.2017

Link Prediction

Methods tried

- Unsupervised
- Supervised
- Matrix factorization

Evaluation

- Cross validation
- Top k hit rates

Graph construction

Paper citation graph

- Directed
- Sparse (#Edges ~ ½#Nodes)
- Consists of unconnected components: use largest component
- Sample subgraph

(Randomly) Sampled subgraph

- About 10% of the nodes
- #Nodes: 14010
- #Edges: 6422
- Training data: 10% of this sampled subgraph

Unsupervised method

Steps:

- Compute similarities between nodes
- Remove edges already there
- Sort similarities
- Evaluate

```
32769.0:35713.0:1.0
32769.0:35762.0:1.0
32769.0:36143.0:1.0
32769.0:37421.0:1.0
2.0:117464.0:1.0
2.0:118839.0:1.0
14.0:81324.0:1.0
14.0:83784.0:1.0
14.0:81047.0:1.0
19.0:3479.0:1.0
19.0:94800.0:1.0
19.0:51019.0:1.0
19.0:92595.0:1.0
19.0:92999.0:1.0
19.0:92247.0:1.0
19.0:54193.0:1.0
26.0:5608.0:1.0
27.0:4003.0:1.0
27.0:37980.0:1.0
27.0:8042.0:1.0
```

Supervised: Classification

Utilize information in papers

- Generate feature vector from title
- Use features between 2 nodes to generate feature of the edge
- Run classification methods

Result

- Random forest: 75.59%
- MLPClassifier: 75.58%

Matrix Factorization

Idea

- Use low rank matrices W and H to represent P, adjacent matrix of the graph
- Minimizing with regularizer

Parameter choosing

- rank(P): about 1000
- W, H: 14010*100
- T: weights constructed using elements in P (balance the training data)

$$\min \|P - WH^T\|_{TF}^2 + \frac{\lambda}{2}(\|W\|_F^2 + \|H\|_F^2)$$

Matrix Factorization-cont.

Other parameters

- Learning rate: 0.003
- \lambda: 1e-5

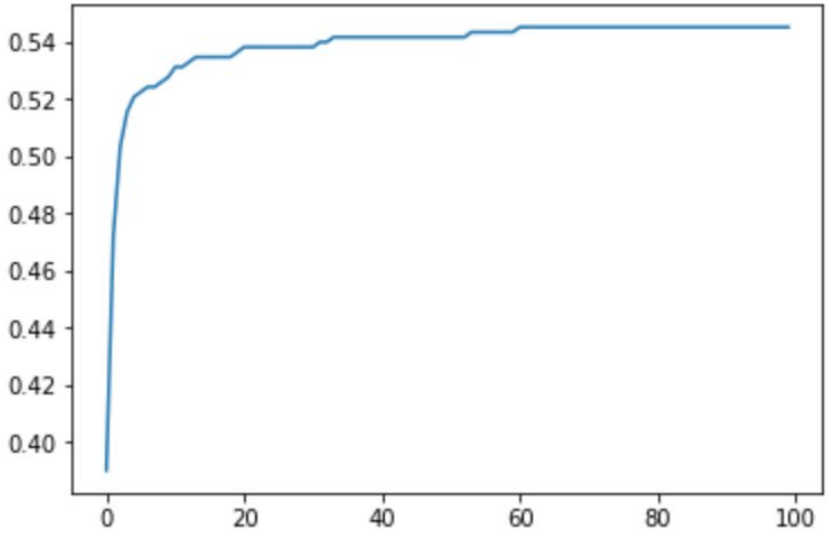
Optimization

- TensorFlow
- Loss: 4953->0.38 in 100 iterations of AdamOptimizer

$$\min \|P - WH^T\|_{TF}^2 + \frac{\lambda}{2}(\|W\|_F^2 + \|H\|_F^2)$$

Matrix Factorization-Evaluation

Top K hit rate



- Hitrate(1) = 38%
- Reaches 55% when K=100

Thank you