A Flexible Generative Framework for Graph-based Semi-supervised Learning

Jiaqi Ma, Weijing Tang, Ji Zhu, Qiaozhu Mei

Presenter: Arshdeep Sekhon

https://qdata.github.io/deep2Read

Motivation

- tasks where the relational information is stored in a graph structure with the data samples as nodes
- two types of graph-based semi-supervised learning.
  - graph-based regularization methods : smooth predictions/features of connected nodes
  - graph neural networks: aggregated into a hidden representation
- This Paper: modeling the joint distribution of the data, graph, and labels with generative models has

Why model graph using a generative model?

- succinct underlying structures of the graph data
- an observed graph is often noisy
- more general relationship among features, outcomes, and the graph

Related Work in detail: Graph based Regularization

$$\sum_i \ell_i + \sum_{i,j} w_{i,j} R(f_i, f_j) \tag{1}$$

Here, R a regularization function of features or labels. Most commonly, it is set as a graph Laplacian regularizer

Method:Generation Process

- the setup is regular semi supervised deep learning
- $\boldsymbol{Y} = [\boldsymbol{Y}_{obs}, \boldsymbol{Y}_m iss]$
- infer $\boldsymbol{Y}_m iss$ based on $(X, \boldsymbol{Y}_o bs, G)$.
- assume the graph is generated based on the node features and outcomes
- This paper: $p(X, Y, G) = p(G|X, Y)p(Y|X)p(X)$

Method:Inference

- To infer labels of nodes: $p_\theta(Y_{miss}|X, Y_{obs}, G)$
- But this is intractable
- Approximate by true posterior $q_\phi(Y_{miss}|X, Y_{obs}, G)$
- ELBO: $logp(Y_{obs}, G|X)E_{q_\phi(Y_{miss}|X,Y_{obs},G)}(logp_\theta(Ymiss, Yobs, G|X)logq(Ymiss|X, Y_{obs}, G$
- defined as $L_{ELBO}(\theta, \phi; X, Y_{obs}, G)$
- $\hat{\theta}, \hat{\phi} = argmin_\theta, \hat{\phi} L_{ELBO}(\theta, \phi; X, Y_{obs}, G)$

Instantiations of the Generative Model

- $p_\theta(Y|X)$ in the generative model: multi-layer perceptron
- $p_\theta(G|X, Y)$ latent space model and stochastic block models
- Further assume independence of edges: $p_\theta(G|X, Y) = \pi p_{\theta_{i,j}}(e_{i,j}|X, Y)$.

Instantiation of LSM

- nodes lie in a latent space and the probability of ei,j only depends the representation of nodes i and j
- logistic regression model

$$p_\theta(e_{i,j} = 1 | x_i, y_i, x_j, y_j) = \sigma([(Ux_i)^T, y_i^T, (Ux_j)^T, y_j^T]w) \quad (2)$$

Instatiation with SBM

- SBM : C types of nodes and each node i has a (latent) type variable z_i, here same as node label
- $e_{i,j}|y_i, y_j = Ber(p_0)$ if $y_i = y_j$
- $e_{i,j}|y_i, y_j = Ber(p_1)$ if $y_i \neq y_j$

Instantiations of the Approximate Posterior Model

- $q_\phi(Y_{miss}|X, Y_{obs}, G)$
- use GCN or GAT

Training the model

- additional supervised loss to better train the approximate posterior model (similar to conditional VAE)
- Negative Sampling: only calculate the probabilities of the edges observed in the graph and a set of "negative edges" randomly sampled from the (i, j) pairs where edges do not exist.

Results: Benchmark

	Cora	Pubmed	Citeseer
MLP	0.583 ± 0.009	0.734 ± 0.002	0.569 ± 0.008
GCN	0.815 ± 0.002	$\mathbf{0.794} \pm 0.004$	0.718 ± 0.003
GAT	0.825 ± 0.005	0.785 ± 0.004	0.715 ± 0.007
LSM_GCN	$\underline{0.825} \pm 0.002$*	0.779 ± 0.004	$\underline{0.744} \pm 0.003$*
LSM_GAT	$\underline{\mathbf{0.829}} \pm 0.003$	0.776 ± 0.007	$\underline{0.731} \pm 0.005$*
SBM_GCN	$\underline{0.822} \pm 0.002$*	0.784 ± 0.006	$\underline{\mathbf{0.745}} \pm 0.004$*
SBM_GAT	$\underline{0.829} \pm 0.003$	0.774 ± 0.004	$\underline{0.740} \pm 0.003$*

The upper block lists the discriminative baselines. The lower block lists the proposed variants of G3NN

Missing Edge Settings

Table 3: Classification accuracy under the missing-edge setting. The **bold** marker, the <u>underline</u> marker, the asterisk (*) marker, and the (±) error bar share the same definitions in Table 2.

	Cora	Pubmed	Citeseer
MLP	0.583 ± 0.009	0.734 ± 0.002	0.569 ± 0.008
GCN	0.665 ± 0.007	0.746 ± 0.004	0.652 ± 0.005
GAT	0.682 ± 0.004	0.744 ± 0.006	0.642 ± 0.004
LSM_GCN	<u>0.711</u> ± 0.005*	**<u>0.766</u>** ± 0.006*	<u>0.704</u> ± 0.002*
LSM_GAT	<u>0.710</u> ± 0.007*	<u>0.766</u> ± 0.004*	<u>0.691</u> ± 0.005*
SBM_GCN	**<u>0.718</u>** ± 0.004*	<u>0.762</u> ± 0.005*	**<u>0.716</u>** ± 0.004*
SBM_GAT	<u>0.716</u> ± 0.007*	<u>0.761</u> ± 0.005*	<u>0.709</u> ± 0.008*

Reduced Label Settings

drop half of the training labels for each class compared to the standard benchmark setting

	Cora	Pubmed	Citeseer
MLP	0.498 ± 0.004	0.674 ± 0.005	0.493 ± 0.010
GCN	0.750 ± 0.003	$\mathbf{0.724} \pm 0.005$	0.666 ± 0.003
GAT	0.771 ± 0.004	0.711 ± 0.006	0.675 ± 0.005
LSM_GCN	$\underline{0.777} \pm 0.002$*	0.709 ± 0.003	$\underline{0.691} \pm 0.005$*
LSM_GAT	$\underline{0.792} \pm 0.004$*	0.699 ± 0.003	$\underline{0.691} \pm 0.004$*
SBM_GCN	$\underline{0.780} \pm 0.002$*	0.710 ± 0.004	$\underline{\mathbf{0.703}} \pm 0.006$*
SBM_GAT	$\underline{\mathbf{0.796}} \pm 0.008$*	0.699 ± 0.003	$\underline{0.698} \pm 0.003$*