Econometrics on Networks

Special Quarter: Data Science & Online Markets

June 14, 2018

Northwestern University

Introduction

Ising Model

Given a graph $G = (V, E)$ whose nodes are associated with external field $(\theta_u)_{u \in V}$ and edges with parameters $(\theta_{u,v})_{(u,v) \in E}$, the Ising model defines a probability distribution over $x \in \{\pm 1\}^n$

PDF of Ising model

$$p(x) = \frac{1}{Z(\theta)} \cdot \exp\left(\sum_i \theta_i x_i + \sum_{(i,j)} x_i x_j \theta_{i,j}\right)$$

θ_i on each vertex is also known as external fields.

Problem 1 : Approximating Ising Distribution

Problem setup

- Unknown Graph
- Polynomial samples from the stationary distribution
- No external fields
- $\theta^*_{ij} > 0 \quad \forall i, j$
- Goal: Recover an Ising model whose distribution is close in TV distance (or SKL) to the Ising distribution sampled from.

Motivation: Lower bounds to recover the graph

- Given i.i.d samples $x_1, .., x_N$, ideally, we would like to recover the graph structure and the associated edge weights.

Motivation: Lower bounds to recover the graph

- Given i.i.d samples $x_1, .., x_N$, ideally, we would like to recover the graph structure and the associated edge weights.
- However, there is a information-theoretic lower bound - we need atleast $O(\exp(d))$ samples, where d is the maximum degree. See [Santhanam and Wainwright, 2012]

Problem Objective

This calls for a relaxation of our original goal. Can we instead just find an approximation of the original Ising distribution $p^*(x)$, i.e find another Ising model whose distribution $p(x)$ satisfies $SKL(p||p^*) \leq \epsilon$, where SKL refers to symmetric KL divergence. Such a goal comes under *proper learning*.

Symmetric KL Divergence

$$SKL(p||p^*) = \sum_{i \in V} (\theta_i - \theta_i^*) \left(\mathbb{E}_p[x_i] - \mathbb{E}_{p^*}[x_i] \right) + \sum_{(i,j)} (\theta_{ij} - \theta_{ij}^*) \left(\mathbb{E}_p[x_i x_j] - \mathbb{E}_{p^*}[x_i x_j] \right) \leq \epsilon$$

Sufficient condition

Sufficient condition:

Ignoring external fields, the following simple condition that $\forall ij \in E, \left|\theta_{ij} - \theta^*_{ij}\right| \leq \frac{\epsilon}{n^2}$ guarantees that $SKL(p||p^*) \leq \epsilon$

Prior work

A recent paper by [Klivans and Meka, 2017] addresses this problem.

Approach: They pose the problem as learning a generalized linear model (GLM): Given Ising samples with conditional mean functions of the form $P(x_v|x_{-v}) = \sigma(\vec{\theta^*}_v.x_{-v})$, we want to recover $\vec{\theta^*}_v$ for all v.

Algorithm: Their algorithm termed *Sparsitron* is based on multiplicative updates method but uses a novel loss function. This method is used to achieve small squared loss and a further assumption that the given Ising model is δ-unbiased makes strong recovery (parameters) possible.

Continuing..

Results: To achieve $||\theta - \theta^*||_\infty \leq \frac{\epsilon}{n^2}$, they require:

Sample-complexity: N $= O(n^8 \lambda^2 exp(\lambda)/\epsilon^4).(\log(n^3/\rho\epsilon))$, where λ upper bounds $||\theta_v^*||_1$ for all v and ρ is the probability of failure.

In the high-temperature regime, under the assumption that $||\theta_v^*||_1 = 1$ for all v, this effectively gives $\tilde{O}\left(\frac{n^8}{\epsilon^4}\right)$

However, information-theoretically, finding a SKL approximation only requires $O(n^2/\epsilon^2)$ samples in the high-temperature regime

Can we do better?

Our approach: Logistic Regression

For a given vertex v, the marginal distribution given all the other labels x_{-v} is

$$p(x_v|\vec{\theta_v}, x_{-v}) = \frac{\exp(\sum_{u,u\neq v} \theta_{v,u} \cdot x_u \cdot x_v)}{\exp(\sum_{u,u\neq v} \theta_{v,u} \cdot x_u \cdot x_v) + \exp(-\sum_{u,u\neq v} \theta_{v,u} \cdot x_u \cdot x_v)}$$

Hence if we have N iid samples, and if $\vec{x_v} \in \{\pm 1\}^N$ is the labels for v, and $X_{-v} \in \{\pm 1\}^{n-1\times N}$

Likelihood

$$L_v(\vec{\theta_v}) = p(\vec{x_v}|\vec{\theta_v}, X_{-v}) = \prod_{i=1}^{N} \frac{\exp(2\sum_{u,u\neq v} \theta_{v,u} \cdot x_u^{(i)} \cdot x_v^{(i)})}{1 + \exp(2\sum_{u,u\neq v} \theta_{v,u} \cdot x_u^{(i)} \cdot x_v^{(i)})}$$

Which would serve as the likelihood function for logistic regression.

Negative log-likelihood minimization

Just like in logistic regression, we perform negative log likelihood (NLL) minimization

Negative log likelihood

$$NLL_v(\vec{\theta_v}) = -2\sum_{i=1}^{N} x_v^{(i)}\vec{\theta_v}^T X_{-v}^{(i)} + \sum_{i=1}^{N} \log\left(1 + \exp(2x_v^{(i)}\vec{\theta_v}^T X_{-v}^{(i)})\right)$$

Gradient

$$\nabla NLL = -2X_{-v}(I - D)x_v$$

where D is a diagonal matrix with $D_{i,i} = \frac{\exp(2x_v^{(i)}\vec{\theta_v}^T X_{-v}^{(i)})}{1+\exp(2x_v^{(i)}\vec{\theta_v}^T X_{-v}^{(i)})}$

Hessian

Hessian

$$\nabla^2 NLL = X_{-v} D^{'} X_{-v}^T$$

where D' is a diagonal matrix with $D'_{i,i} = \frac{4e^{2x_v^{(i)} \vec{\theta}_v^T x_{-v}^{(i)}}}{\left(1+e^{2x_v^{(i)} \vec{\theta}_v^T x_{-v}^{(i)}}\right)^2}$

Notice that this is a PSD matrix and thus, the objective is convex. This allows us to exploit convex optimization tools (aka Gradient Descent).

Technical Questions

- Understanding spectral properties of the Hessian so as to check for strong convexity and smoothness. This might be of independent interest.
- Under what conditions is the NLL objective function Lipschitz?
- Study how minimizing the NLL objective might give us strong recovery guarantees in the parameter space.
- What set of assumptions about the temperature and the underlying graph are needed for the above?
- How does regularization help?

Problem 2 : Learning the sign form one sample

Problem setup

- Unknown Graph
- One sample
- No external fields
- $\theta_{ij} = \theta \quad \forall i, j$ i.e. edge parameter is the same for all edges.
- Goal: Recover the sign of θ

Hurdles

Can't do it for general Graphs (even with constant probability)

For example :

- Star Graph $K(1, n)$
- Complete Bipartite Graph where one side is constant sized : $K(c, n)$
- Path on cliques

We can learn on $K(n, n)$ graphs!

Approaches and Experiments

Given these results : Try to solve the problems on restricted graph classes.

Idea 1 : $G(n, p)$ graphs.

Idea 2 : Graphs that are 'well connected'.

Motivation : If there are a lot of pieces of the graph that don't interact much with each other, then$|\sum_{i=1}^{n} X_i|$ is close to zero in both cases $\theta > 0$ and $\theta < 0$. For example a path of cliques.

Preliminary Experiments

In random $G(n, p)$ graphs, for $\theta < 0$, $\sum X_i$ is close to zero and for $\theta > 2/np$, $|\sum X_i|$ is close to n.

As a possible explanation, $\theta \leq 1/np$ is the 'high temperature' regime or Dobrushin's condition is satisfied. There might be a transition behaviour at this point.

Random Cluster(FK) model

- Introduced by Fortuin and Kastelyn, for $\theta > 0$, this is another way of sampling from the Ising model.
- Sample a subset $S \subseteq E$ of edges of the input graph G with probability proportional to

$$p^{|S|}(1-p)^{|E|-|S|}q^{k(S)}$$

  where $k(S)$ is the number of connected components in (V, S).
- Assign each connected component a label uniformly at random from $1 \dots q$.
- They prove, this distribution is exactly same as the Potts model for q states. Specifically for $q = 2$ this is exactly the Ising distribution.

Using the FK model : Further Approaches

For the restricted classes of graphs mentioned above :

- Prove a lower bound on the size of the largest component.
- Prove that the other components are somewhat small.
- Conclude that in expectation, $|\sum X_i|$ size of largest component.
- Conclude separately that for $\theta < 0$, $|\sum X_i|$ is close to zero. (Some preliminary experimental and theoretical evidence suggesting this.)
- Deduce sign of θ depending on whether $\sum X_i$ is small or large.

Problem 3 : Logistic Regression on Networks

Model

Ising model and features

- Graph $G(V, E)$ with $n \times n$ adjacency matrix A (Known).

Model

Ising model and features

- Graph $G(V, E)$ with $n \times n$ adjacency matrix A (Known).
- Features $\vec{y}_v \in \mathbb{R}^m$ (Known, m is constant).

Model

Ising model and features

- Graph $G(V, E)$ with $n \times n$ adjacency matrix A (Known).
- Features $\vec{y}_v \in \mathbb{R}^m$ (Known, m is constant).
- Parameters $\beta \in \mathbb{R}_+$ and $\vec{\theta} \in \mathbb{R}^m$ (Unknown).

Model

Ising model and features

- Graph $G(V, E)$ with $n \times n$ adjacency matrix A (Known).
- Features $\vec{y}_v \in \mathbb{R}^m$ (Known, m is constant).
- Parameters $\beta \in \mathbb{R}_+$ and $\vec{\theta} \in \mathbb{R}^m$ (Unknown).

Given an assignment $\vec{\sigma} : V \to \{-1, +1\}$, we define the following probability distribution (on the 2^n configurations):

$$\Pr[\{]\tau = \sigma\} = \frac{\exp\left(\sum_{v \in V}(\vec{\theta}^\top \cdot \vec{y}_v)\sigma_v + \beta\vec{\sigma}^\top A\vec{\sigma}\right)}{Z(G)} \tag{1}$$

where $Z(G)$ is the partition function of the system (i.e., the renormalization factor).

Model

Ising model and features

- Graph $G(V, E)$ with $n \times n$ adjacency matrix A (Known).
- Features $\vec{y}_v \in \mathbb{R}^m$ (Known, m is constant).
- Parameters $\beta \in \mathbb{R}_+$ and $\vec{\theta} \in \mathbb{R}^m$ (Unknown).

Given an assignment $\vec{\sigma} : V \to \{-1, +1\}$, we define the following probability distribution (on the 2^n configurations):

$$\Pr[\{]\tau = \sigma\} = \frac{\exp\left(\sum_{v \in V}(\vec{\theta}^\top \cdot \vec{y}_v)\sigma_v + \beta \vec{\sigma}^\top A \vec{\sigma}\right)}{Z(G)} \quad (1)$$

where $Z(G)$ is the partition function of the system (i.e., the renormalization factor).

Problem: We would like to infer $\vec{\theta}, \beta$

Naive approach for bounded degree (Δ) graphs

- Find an independent set I of size at least $\frac{n}{\Delta}$.

Naive approach for bounded degree (Δ) graphs

- Find an independent set I of size at least $\frac{n}{\Delta}$.
- Each $v \in I$, conditioned on its neighbors has the following marginal:

$$\Pr[\{]\tau_v = 1 | \sum_{u \in N(v)} \sigma_u = c\} = \frac{1}{1 + e^{-2(\vec{\theta},\beta)^\top (\vec{y}_v, c)}}. \tag{2}$$

Naive approach for bounded degree (Δ) graphs

- Find an independent set I of size at least $\frac{n}{\Delta}$.
- Each $v \in I$, conditioned on its neighbors has the following marginal:

$$\Pr[\{]\tau_v = 1 | \sum_{u \in N(v)} \sigma_u = c\} = \frac{1}{1 + e^{-2(\vec{\theta},\beta)^\top (\vec{y}_v, c)}}. \qquad (2)$$

- To infer β, θ, we reduce the problem to classic logistic regression (independent "samples"). $\vec{\theta}$ can be inferred but the signal for β might be lost.

Main challenges!

- Go beyond bounded degree graphs (dense).

Naive approach for bounded degree (Δ) graphs

- Find an independent set I of size at least $\frac{n}{\Delta}$.
- Each $v \in I$, conditioned on its neighbors has the following marginal:

$$\Pr[\{]\tau_v = 1 | \sum_{u \in N(v)} \sigma_u = c\} = \frac{1}{1 + e^{-2(\vec{\theta}, \beta)^\top (\vec{y}_v, c)}}. \qquad (2)$$

- To infer β, θ, we reduce the problem to classic logistic regression (independent "samples"). $\vec{\theta}$ can be inferred but the signal for β might be lost.

Main challenges!

- Go beyond bounded degree graphs (dense).
- Use the information from all the vertices.

Naive approach for bounded degree (Δ) graphs

- Find an independent set I of size at least $\frac{n}{\Delta}$.
- Each $v \in I$, conditioned on its neighbors has the following marginal:

$$\Pr[\{]\tau_v = 1 | \sum_{u \in N(v)} \sigma_u = c\} = \frac{1}{1 + e^{-2(\vec{\theta},\beta)^\top(\vec{y}_v,c)}}. \tag{2}$$

- To infer β, θ, we reduce the problem to classic logistic regression (independent "samples"). $\vec{\theta}$ can be inferred but the signal for β might be lost.

Main challenges!

- Go beyond bounded degree graphs (dense).
- Use the information from all the vertices.
- Beat the $\sqrt{\frac{\Delta}{n}}$ consistency that will occur.

Questions?

Thank You!

References i

Klivans, A. and Meka, R. (2017).
Learning graphical models using multiplicative weights.
arXiv preprint arXiv:1706.06274.

Santhanam, N. P. and Wainwright, M. J. (2012).
Information-theoretic limits of selecting binary graphical models in high dimensions.
IEEE Transactions on Information Theory, 58(7):4117–4134.