Estatística de Redes Sociais

Antonio Galves

Módulo 2

7a. aula

Grafos aleatórios e redes sociais.

Estimação de máxima verossimilhança em $G(N, p)$.

Recapitulando: o grafo não dirigido de Erdős-Rényi

- Conjunto de vértices: $V = \{1, \ldots, N\}$.
- Os valores das entradas da matriz

$$M = (M(v, v') : v, v' \in V, v < v')$$

são variáveis aleatórias, independentes e identicamente distribuídas (i.i.d.) com

$$\mathbb{P}\{M(v, v') = 1\} = p \text{ e } \mathbb{P}\{M(v, v') = 0\} = 1 - p\,,$$

e $M(v, v) = 0$, para todo $v \in V$.
- Essa classe de grafos será designada com a notação $G(N, p)$.

Verossimilhança de uma realização

- Seja M a matriz de adjacência de uma realização de $G(N, p)$.
- Exemplo: $N = 5$

$$M = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ & 0 & 0 & 1 & 0 \\ & & 0 & 1 & 0 \\ & & & 0 & 1 \\ & & & & 0 \end{pmatrix}$$

- Na linha v, coluna v', com $v < v'$, colocamos o valor $M(v, v')$.
- Queremos calcular a probabilidade de obtermos um grafo com essa matriz de adjacência na classe $G(5, p)$, com $p \in [0, 1]$ conhecido.

Verossimilhança de um grafo de matriz M

- $$M = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ & 0 & 0 & 1 & 0 \\ & & 0 & 1 & 0 \\ & & & 0 & 1 \\ & & & & 0 \end{pmatrix}$$

$$\mathbb{P}_p \begin{pmatrix} M(1,2) = M(1,3) = M(2,4) = M(3,4) = M(4,5) = 1, \\ M(1,4) = M(1,5) = M(2,3) = M(2,5) = M(3,5) = 0 \end{pmatrix} =$$

$$= \mathbb{P}_p(M(1,2) = 1)\mathbb{P}_p(M(1,3) = 1) \ldots \mathbb{P}_p(M(2,5) = 0)\mathbb{P}_p(M(3,5) = 0)$$

- Na última passagem usamos a independência das variáveis aleatórias $M(v, v')$, $v < v'$.

Verossimilhança de um grafo de matriz M

- Substituímos cada probabilidade pelo seu valor.

$$\mathbb{P}_p(M(v,v') = \epsilon(v,v')) = \begin{cases} p, \text{ se } \epsilon(v,v') = 1, \\ 1-p, \text{ se } \epsilon(v,v') = 0. \end{cases}$$

- Ou seja,

$$\mathbb{P}_p(M(v,v') = \epsilon(v,v')) = p^{\epsilon(v,v')}(1-p)^{1-\epsilon(v,v')}.$$

- Portanto,

$$\mathbb{P}_p(M(1,2) = 1)\mathbb{P}_p(M(1,3) = 1)\ldots\mathbb{P}_p(M(2,5) = 0)\mathbb{P}_p(M(3,5) = 0)$$

$$= p^{\mathcal{N}_G(1)}(1-p)^{\mathcal{N}_G(0)}, \text{ onde}$$

$$\mathcal{N}_G(1) = \sum_{v<v'} M(v,v') \quad \text{e} \quad \mathcal{N}_G(0) = \sum_{v<v'} \Big(1 - M(v,v')\Big).$$

Em linguagem de gente

- $\mathcal{N}_G(1)$ conta quantos pares de vértices (v, v') com $v < v'$ estão ligados por arestas no grafo não dirigido G.
- $\mathcal{N}_G(0)$ conta quantos pares de vértices (v, v') com $v < v'$ não estão ligados por arestas no grafo não dirigido G.
- Portanto, o cálculo anterior diz que a probabilidade de um grafo G na classe $G(N, p)$ ter matriz de adjacência M vale

$$p^{\sum_{v<v'} M(v,v')}(1-p)^{\sum_{v<v'}[1-M(v,v')]} = p^{\mathcal{N}_G(1)}(1-p)^{\mathcal{N}_G(0)}.$$

- Na literatura estatística,

$$\mathbb{P}_p(M(v,v') = \epsilon(v,v'), v < v', v, v' \in V)$$

é chamado de verossimilhança do grafo $(M(v,v') = \epsilon(v,v'), v < v', v, v' \in V)$ na classe $G(N,p)$.

QUIZ

Se G é um grafo não dirigido com N vértices, quanto vale

$$\mathcal{N}_G(1) + \mathcal{N}_G(0)?$$

R E S P O S T A

- Observe que $\mathcal{N}_G(1) + \mathcal{N}_G(0)$ é o número total de pares (v, v'), com $v \leq v'$, sendo $v, v' \in V = \{1, ..., N\}$.
- Então o que estamos perguntando é: qual é o número total de pares (v, v'), com $v \leq v'$ que podemos formar com elementos do conjunto $V = \{1, ..., N\}$?
- Em outras palavras, $\mathcal{N}_G(1) + \mathcal{N}_G(0)$ é o número total de subconjuntos distintos com cardinal 2 que podemos formar com elementos do conjunto $V = \{1, ..., N\}$.
- Em conclusão,

$$\mathcal{N}_G(1) + \mathcal{N}_G(0) = \binom{N}{2} = \frac{N(N-1)}{2}.$$

- Este QUIZ sugere uma maneira simples de controlar a correção do cálculo de $\mathcal{N}_G(1)$ e $\mathcal{N}_G(0)$, quando G é muito grande.

Estimando o parâmetro p

- Seja M a matriz de adjacência de uma realização de $G(N, p)$ com p desconhecido.
- Exemplo: $N = 5$

$$M = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ & 0 & 0 & 1 & 0 \\ & & 0 & 1 & 0 \\ & & & 0 & 1 \\ & & & & 0 \end{pmatrix}$$

- Na linha v, coluna v', com $v < v'$, colocamos o valor $M(v, v')$.
- Sabendo que M é a matriz de adjacência de um grafo gerado por $G(5, p)$, com p desconhecido, queremos estimar o valor do parâmetro p.

Estimador de máxima verossimilhança

- Seja $M = (M(v, v') : v < v', v, v' \in V)$ a matriz de adjacência de um grafo não dirigido tendo vértices em $V = \{1, \ldots, N\}$.
- Para todo $q \in [0, 1]$ seja

$$P_q(M) = q^{\mathcal{N}_M(1)}(1-q)^{\mathcal{N}_M(0)}, \text{ onde}$$

$$\mathcal{N}_M(1) = \sum_{v<v'} M(v, v') \quad \text{e} \quad \mathcal{N}_M(0) = \sum_{v<v'} \Big(1 - M(v, v')\Big).$$

- Se M é uma matriz de adjacência de um grafo tendo como vértices $V = \{1, \ldots N\}$, então $P_q(M)$ é exatamente a sua verossimilhança na classe $G(N, q)$.
- Dado M, o estimador de máxima verossimilhança $\hat{p}_N$ é definido como o valor de $q \in [0, 1]$ que maximiza $P_q(M)$.

Estimador de máxima verossimilhança

- Formalmente,

$$\hat{p}_N = \text{argmax}\{P_q(M) : q \in [0,1]\}.$$

- Vamos reescrever a definição de $\hat{p}_N =$ usando o valor de $P_q(M)$.

$$\hat{p}_N = \text{argmax}\{q^{\mathcal{N}_M(1)}(1-q)^{\mathcal{N}_M(0)} : q \in [0,1]\}.$$

- Fazer contas com produtos e potências não é muito cômodo e numericamente é inviável pois o produto

$$q^{\mathcal{N}_M(1)}(1-q)^{\mathcal{N}_M(0)}$$

é um número positivo muito pequeno.

Reescrevendo $\hat{p}_N$

- Vamos calcular o valor de $q \in [0,1]$ que maximiza $f(q) = \log(P_q(M))$.
- $$\hat{p}_N = \text{argmax}\left\{f(q) : q \in [0,1]\right\}.$$
- $$f(q) = \log\left(q^{\mathcal{N}_M(1)}(1-q)^{\mathcal{N}_M(0)}\right) = \log(q)\mathcal{N}_M(1) + \log(1-q)(\mathcal{N}_M(0))$$
- Essa nova definição é possível, porque a função $r \to \log(r)$ é estritamente crescente.

Calculando $\hat{p}_N$

- Vamos calcular a derivada $\frac{d}{dq} f(q)$:
-

$$\frac{d}{dq} f(q) = \mathcal{N}_M(1) \frac{d}{dq} \log(q) + \mathcal{N}_M(0) \frac{d}{dq} \log(1-q) = \frac{\mathcal{N}_M(1)}{q} - \frac{\mathcal{N}_M(0)}{1-q}.$$

- $\hat{p}_N$ é o valor de q que anula essa derivada. Ou seja,

$$\frac{\mathcal{N}_M(1)}{\hat{p}_N} = \frac{\mathcal{N}_M(0)}{1-\hat{p}_N}.$$

- Isso acontece quando

$$\hat{p}_N = \frac{\mathcal{N}_M(1)}{\mathcal{N}_M(1) + \mathcal{N}_M(0)} = \frac{\mathcal{N}_M(1)}{\binom{N}{2}}.$$

Confirmando que $\hat{p}_N$ é o ponto de máximo

- Para isso, vamos calcular a segunda derivada.
- $$\frac{d^2}{dq^2} f(q) = \frac{d}{dq}\left(\frac{\mathcal{N}_M(1)}{q} - \frac{\mathcal{N}_M(0)}{1-q}\right) = -\frac{\mathcal{N}_M(1)}{q^2} - \frac{\mathcal{N}_M(0)}{(1-q)^2} < 0.$$
- Ou seja, $\hat{p}_N$ que anula a primeira derivada é efetivamente um ponto de máximo.

DESAFIO

- Vamos definir uma sequência de grafos $(G_N : N \geq 2)$ na classe de Erdős-Rényi, tendo todos o mesmo valor do parâmetro $p \in (0,1)$ e G_N tendo como conjunto de vértices $V_N = \{1, ..., N\}$. Seja M_N a matriz de adjacência do grafo G_N.
- Finalmente vamos supor que esses grafos estão encaixados no seguinte sentido: para todo $v < v'$, $v, v' \in V_N$, $M_{N+k}(v, v') = M_N(v, v')$ para todo $k \geq 0$.
- Para cada $N \geq 2$, calculamos o estimador $\hat{p}_N$ com base na matriz M_N. A sequência $(\hat{p}_N : N \geq 2)$ é convergente? Caso seja, quanto vale o limite

$$\lim_{n \to \infty} \hat{p}_N ?$$

Exercícios

1. Seja G um grafo gerado aleatoriamente na classe $G(101, 0.3)$. Calcule $\mathbb{E}(D(1))$. Usando o Teorema-Limite Central, calcule aproximadamente a probabilidade $\mathbb{P}(D(1) > 40)$.
2. Seja G um grafo gerado aleatoriamente na classe $G(100, 0.3)$. Calcule $\mathbb{E}(\mathcal{N}_G(1))$ e usando o Teorema-Limite Central, calcule aproximadamente a probabilidade $\mathbb{P}(\mathcal{N}_G(1) > 1683)$.
3. Escreva um código para simular a sequência do desafio G_N de grafos em $G(N, p)$ com $p \in (0, 1)$ fixado. Use esse código para simular os grafos e tentar identificar o limite $\lim_{n\to\infty} \hat{p}_N$.
4. Construa uma sequência $(G_t : t \geq 0)$ de grafos *rico fica mais rico*, com $V_0 = \{1, ..., 10\}$, e com M_0, tal que para todo par de vértices $v < v'$, $M_0(v, v') = 1$, se e somente se, v e v' pertencerem ambos a $\{1, ..., 5\}$, ou ambos a $\{6, ..., 10\}$. O que essa condição inicial implica sobre $(G_t : t \geq 0)$?