Estatística de Redes Sociais

Antonio Galves

Aula 8

Módulos 1 e 2: complementos e revisão

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 $G(N, p)$.

Q U I Z

- Um grafo $G(10, 0.3)$ foi gerado e sua verossimilhança calculada. Porém, o papel onde estava escrito o valor da verossimilhança foi mal encaminhado e talvez, perdido. O fato é que a pessoa encarregada de arquivar o resultado encontrou em sua mesa uma folha de papel com um possível valor da verossimilhança calculada.
- O valor escrito na folha era $(0.3)^{20}(0.7)^{20}$.
- Será esse o valor perdido da verossimilhança do grafo?

Generalização: $G(N, p)$ com duas comunidades

- O conjunto de vértices V é agora formado por duas comunidades $V^{(1)}$ e $V^{(2)}$.
- $V = V^{(1)} \cup V^{(2)}$, com $V^{(1)} \cap V^{(2)} = \emptyset$.
- As v.a. $(M(v, v') : v < v', v, v' \in V)$ são independentes, porém a distribuição de $M(v, v')$ depende das comunidades as quais pertencem v e v'.

$$\mathbb{P}(M(v, v') = 1) = \begin{cases} p_{1,1}, \text{ se } \{v, v'\} \subset V^{(1)}, \\ p_{2,2}, \text{ se } \{v, v'\} \subset V^{(2)}, \\ p_{1,2}, \text{ se } \{v, v'\} \cap V^{(1)} \neq \emptyset \text{ e } \{v, v'\} \cap V^{(2)} \neq \emptyset \end{cases}$$

- Grafos assim construídos serão denotados $G(V^{(1)}, V^{(2)}, \mathbf{p})$, onde $\mathbf{p} = (p_{1,1}, p_{2,2}, p_{1,2})$.

Verossimilhança de $G(V^{(1)}, V^{(2)}, \mathbf{p})$

- Suponhamos que $(M(v,v') = \epsilon(v,v') : v < v', v, v' \in V)$ é a matriz de adjacência de um grafo $G(V^{(1)}, V^{(2)}, (p_{1,1}, p_{2,2}, p_{1,2}))$.
- Vamos calcular o log da verossimilhança desse grafo

$$l(\mathbf{p}) = \log\left(\mathbb{P}_{\mathbf{p}}(M(v,v') = \epsilon(v,v') : v < v', v, v' \in V)\right) =$$

$$\sum_{v<v',v,v'\in V} \log(\mathbb{P}_{\mathbf{p}}(M(v,v') = \epsilon(v,v'))).$$

$$= \sum_{i\leq j, i,j\in\{1,2\}} \left[\mathcal{N}_{i,j}(1)\log(p_{i,j}) + \mathcal{N}_{i,j}(0)\log(1-p_{i,j})\right],$$

onde

$$\mathcal{N}_{i,j}(1) = \sum_{v\in V^{(i)}, v'\in V^{(j)}} M(v,v') \text{ e}$$

$$\mathcal{N}_{i,j}(0) = \sum_{v\in V^{(i)}, v'\in V^{(j)}} [1 - M(v,v')]$$

Em linguagem de gente

- $\mathcal{N}_{i,j}(1)$ conta quantos pares de vértices (v, v') com $v < v'$, $v \in V^{(i)}$ e $v' \in V^{(j)}$ estão ligados por arestas no grafo não dirigido $G(V^{(1)}, V^{(2)}, \mathbf{p})$.
- $\mathcal{N}_{i,j}(0)$ conta quantos pares de vértices (v, v') com $v < v'$,$v \in V^{(i)}$ e $v' \in V^{(j)}$ não estão ligados por arestas no grafo não dirigido $G(V^{(1)}, V^{(2)}, \mathbf{p})$.

QUIZ

- Quanto vale $N_{1,1}(1) + N_{1,1}(0)$?
- Quanto vale $N_{1,2}(1) + N_{1,2}(0)$?
- Quanto vale

$$\sum_{i \leq j, i, j \in \{1,2\}} \Big[N_{i,j}(1) + N_{i,j}(0) \Big] ?$$

RESPOSTAS

- $$N_{1,1}(1) + N_{1,1}(0) = \binom{|V^1|}{2}$$
- $$N_{1,2}(1) + N_{1,2}(0) = |V^1||V^2|$$
- $$\sum_{i \leq j, i,j \in \{1,2\}} \Big[N_{i,j}(1) + N_{i,j}(0) \Big] = \binom{|V^1| + |V^2|}{2}$$

Exercício

- Seja M a matriz de adjacência de uma realização de $G(V^{(1)}, V^{(2)}, \mathbf{p},)$ com $V^{(1)} = \{1, 2, 3\}$, $V^{(2)} = \{4, 5, 6\}$ e $\mathbf{p} = (p_{1,1} = \alpha, p_{2,2} = \beta, p_{1,2} = \gamma)$, sendo α, β, γ três parâmetros fixados no intervalo $[0, 1]$.

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

- Calcule o logaritmo da verossimilhança dessa matriz.
- Calcule os valores de α, β e γ que maximizam essa verossimilhança.

Respostas

- Arestas entre vértices em $V^{(1)}$: $\{\{1,2\},\{1,3\}\}$. $\mathcal{N}_{1,1}(1) = 2$.
- Arestas ausentes em $V^{(1)}$: $\{\{2,3\}\}$. $\mathcal{N}_{1,1}(0) = 1$.
- Arestas entre vértices em $V^{(2)}$: $\{\{4,5\}\}$. $\mathcal{N}_{2,2}(1) = 1$.
- Arestas ausentes em $V^{(2)}$: $\{\{4,6\},\{5,6\}\}$. $\mathcal{N}_{1,1}(0) = 2$.
- Arestas entre vértices de $V^{(1)}$ e $V^{(2)}$:
  $\{\{1,4\},\{2,4\},\{2,5\},\{2,6\},\{3,6\}\}$. $\mathcal{N}_{1,2}(1) = 5$.
- Arestas ausentes entre vértices de $V^{(1)}$ e $V^{(2)}$:
  $\{\{1,5\},\{1,6\},\{3,4\},\{3,5\}\}$. $\mathcal{N}_{1,2}(0) = 4$.
- Observe que

$$2 + 1 + 1 + 2 + 5 + 4 = 15 = \binom{6}{2}.$$

Respostas

- Logo $l(\mathbf{p})$ é igual a

$$2\log(\alpha)+\log(1-\alpha)+\log(\beta)+2\log(1-\beta)+5\log(\gamma)+4\log(1-\gamma).$$

- Para encontrar os valores de $\hat{\alpha}$, $\hat{\beta}$ e $\hat{\gamma}$ que maximizam $l(\mathbf{p})$ temos que derivar $l(\mathbf{p})$ em relação a α, β e γ, e igualar a zero:
- $$\frac{\partial}{\partial\alpha}l(\mathbf{p})\bigg|_{\alpha=\hat{\alpha}} = \frac{2}{\hat{\alpha}} - \frac{1}{1-\hat{\alpha}} = 0. \text{ Logo, } \hat{\alpha} = \frac{2}{3}.$$
- $$\frac{\partial}{\partial\beta}l(\mathbf{p})\bigg|_{\beta=\hat{\beta}} = \frac{1}{\hat{\beta}} + \frac{2}{1-\hat{\beta}} = 0. \text{ Logo, } \hat{\beta} = \frac{1}{3}.$$
- $$\frac{\partial}{\partial\gamma}l(\mathbf{p})\bigg|_{\gamma=\hat{\gamma}} = \frac{5}{\hat{\gamma}} - \frac{4}{1-\hat{\gamma}} = 0. \text{ Logo, } \hat{\gamma} = \frac{5}{9}.$$

O que essa conta nos ensina?

- Quando calculamos o logaritmo da verossimilhança do grafo na classe $G(V^{(1)}, V^{(2)}, \mathbf{p})$, os termos correspondentes às arestas entre vértices de $V^{(1)}$, às arestas entre vértices de $V^{(2)}$ e às arestas ligando vértices de $V^{(1)}$ e $V^{(2)}$, aparecem separados na soma.
- Em consequência, quando buscamos os valores de α, β e γ que igualam as derivadas a zero, temos o mesmo cálculo que tinhamos quando havia só uma comunidade.
- Ou seja, quando temos duas comunidades, calculamos separadamente os estimadores de máxima verossimilhança para $p_{1,1}$, $p_{2,2}$ e $p_{1,2}$.

Estimadores de máxima verossimilhança para $p_{i,j}$

- Para $i, j \in \{1, 2\}$, $i \leq j$,

$$\hat{p}_{i,j} = \frac{\mathcal{N}_{i,j}(1)}{\mathcal{N}_{i,j}(0) + \mathcal{N}_{i,j}(1)}.$$

D E S A F I O

- Refaça a modelagem do grafo $G(N, p)$ e a derivação dos estimadores de máxima verossimilhança no caso em que temos C comunidades, com $C \geq 2$.
- Ou seja, o conjunto de vértices

$$V = V^{(1)} \cup \ldots \cup V^{(C)} \text{ e } V^{(c)} \cap V^{(c')} = \emptyset, \text{ se } c \neq c';$$

- as variáveis aleatórias $M(v, v')$ são independentes e para todo $v \neq v'$

$$\mathbb{P}(M(v, v') = 1) = p_{c,c'}, \text{ se } v \in V^{(c)} \text{ e } v' \in V^{(c')}.$$

Um exercício com o Teorema-Limite Central

- **Exercício:** 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 $\mathbb{P}(D(1) > 40)$.
- **Resolução:**

$$\mathbb{E}(D(1)) = \mathbb{E}\left(\sum_{v=2}^{101} M(1,v)\right) = \sum_{v=2}^{101} \mathbb{E}\left(M(1,v)\right) = 100 \times 0.3 = 30.$$

- O Teorema-Limite Central diz que

$$\mathbb{P}\left(\frac{D_N(1) - 30}{\sqrt{100 \times 0.3 \times 0.7}} > t\right) \approx \frac{1}{\sqrt{2\pi}} \int_t^{+\infty} e^{-\frac{1}{2}s^2} ds.$$

- O valor da integral do lado direito dessa equação pode ser encontrado em uma tabela de $N(0,1)$.

Um exercício com o Teorema-Limite Central

- Observamos que

$$\mathbb{P}\left(\frac{D(1)-30}{\sqrt{100\times 0.3\times 0.7}} > t\right) = \mathbb{P}\left(\frac{D(1)-30}{4.6} > t\right)$$

-

$$= \mathbb{P}\left(D(1) > 30 + 4.6\times t\right).$$

- Portanto, para calcular um valor aproximado para $\mathbb{P}(D(1) > 40)$, basta encontrar o valor de t para o qual

$$30 + 4.6\times t = 40$$

- Ou seja,

$$t = \frac{40-30}{4.6} = 2.17.$$

Olhando a tabela de $N(0,1)$

- Sabemos que

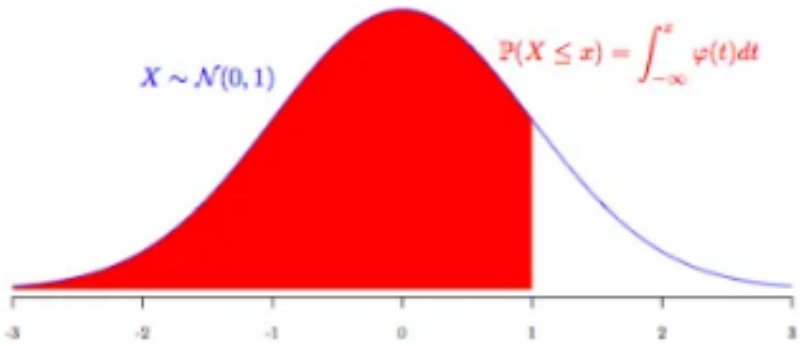
$$\frac{1}{\sqrt{2\pi}}\int_{2.17}^{+\infty} e^{-\frac{1}{2}s^2}ds = 1 - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{2.17} e^{-\frac{1}{2}s^2}ds.$$

- A tabela de $N(0,1)$ nos diz que

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{2.17} e^{-\frac{1}{2}s^2}ds \approx 0.98505.$$

- Logo

$$\mathbb{P}(D(1) > 40) \approx 1 - 0.9850 = 0.015.$$



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857