FACULTAD DE INGENIERÍA

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Aprendizaje Automático para Datos en Grafos

Aprendiendo Grafos desde Datos

Marcelo Fiori
Muy basado en transparencias de **Gonzalo Mateos**

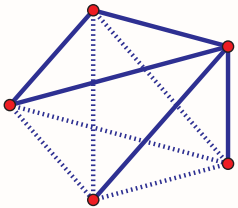
mfiori@fing.edu.uy
http://www.fing.edu.uy/~mfiori/

Octubre 2022

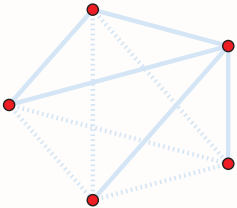
Recordemos

- Tres problemas canónicos de inferencia de topología del grafo [Kolaczyk'09]
  - (i) Predicción de enlaces
  - (ii) Association network inference ← Énfasis de estas clases
  - (iii) Tomographic network topology inference

Association network inference



Original graph



Association network inference

- Supongamos que solo observamos la señal en el grafo $\mathbf{x} = [x_1, \ldots, x_N]^\top$; y
- Asumimos (i, j) definido por un 'nivel de asociación' no trivial entre x_i, x_j
- **Objetivo:** predecir estado de aristas para todos los pares de vértices $\mathcal{V}^{(2)}$

Recordemos

- Redes de correlación
  - $\Rightarrow$ Test de hipótesis usando la correlación empírica
- Correlaciones parciales
  - $\Rightarrow$ Similar, pero condicionando a otros nodos para evitar influencias indirectas

Undirected Gaussian graphical models

- Supongamos que las variables $\{x_i\}_{i\in\mathcal{V}}$ tienen distribución gaussiana multivariada
  - $\Rightarrow$ Consideramos $\rho_{ij|\mathcal{V}\setminus\{i,j\}}$ condicionada respecto a todos los otros nodos ($m = N-2$)

Teorema

Bajo este supuesto gaussiano, los nodos $i, j \in \mathcal{V}$ tienen correlación parcial

$$\rho_{ij|\mathcal{V}\setminus\{i,j\}} = 0$$

si y solo si x_i y x_j son condicionalmente independientes dados $\{x_k\}_{k\in\mathcal{V}\setminus\{i,j\}}$

- **Def:** el grafo de independencia condicional $G(\mathcal{V}, \mathcal{E})$ tiene conjunto de aristas

$$\mathcal{E} = \left\{(i,j) \in \mathcal{V}^{(2)} : \rho_{ij|\mathcal{V}\setminus\{i,j\}} \neq 0\right\}$$

  - $\Rightarrow$ Es un caso especial y popular de redes de correlaciones parciales
- También conocido como Gaussian Markov random field (GMRF)

Covariance selection

- Sea $\mathbf{\Sigma}$ la matriz de covarianza de $\mathbf{x} = [x_1, \ldots, x_N]^\top$
  Def: la matriz de precisión es $\mathbf{\Theta} := \mathbf{\Sigma}^{-1}$ con entradas θ_{ij}
- Resultado clave: Para GMRFs, las correlaciones parciales pueden ser expresadas como

$$\rho_{ij|V\setminus\{i,j\}} = -\frac{\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}}$$

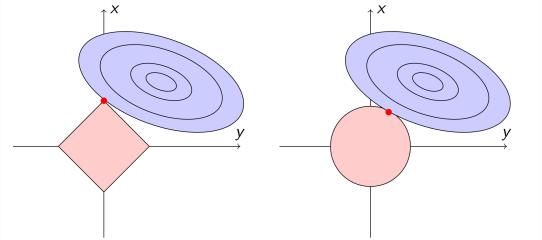
  $\Rightarrow$ Entradas no nulas en $\mathbf{\Theta} \Leftrightarrow$ Aristas en el grafo G
- El problema de inferir G desde $\mathcal{X}$ es conocido como covariance selection [Dempster'74]
  $\Rightarrow$ Métodos clásicos son 'network-agnostic,' y testean

$$H_0 : \rho_{ij|\mathcal{V}\setminus\{i,j\}} = 0 \quad \text{versus} \quad H_1 : \rho_{ij|\mathcal{V}\setminus\{i,j\}} \neq 0$$

- En general no escalable, y $P \ll N$ entonces la estimación de $\hat{\mathbf{\Sigma}}$ es desafiante

Digresión: Sparsity y la norma ℓ_1

- Consideremos la minimización de una función cuadrática de $\boldsymbol{\theta}$ como en mínimos cuadrados
- Q: ¿Cuál es el efecto de una restricción de norma ℓ_1, i.e., $\|\boldsymbol{\theta}\|_1 = \sum_i |\theta_i| \leq \tau$?



$\Rightarrow$ Si conjuntos de nivel intersecan restricciones en un vértice$\rightarrow$ soluciones esparsas

- Estimador Lasso permite hacer selección de variables [Tibshirani'94]

$$\hat{\boldsymbol{\theta}}_{Lasso} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \boldsymbol{\theta})^2, \text{ s. to } \|\boldsymbol{\theta}\|_1 \leq \tau$$

Graphical Lasso

- Estimador de máxima verosimilitud de $\mathbf{\Theta}$ con regularización que promueve esparsidad [Yuan-Lin'07]

$$\hat{\mathbf{\Theta}} \in \arg\max_{\mathbf{\Theta} \succeq \mathbf{0}} \left\{ \log\det \mathbf{\Theta} - \operatorname{trace}(\hat{\mathbf{\Sigma}}\mathbf{\Theta}) - \lambda \|\mathbf{\Theta}\|_1 \right\}$$

  $\Rightarrow$ Efectivo cuando $P \ll N$, da lugar a modelos interpretables

  $\Rightarrow$ Métodos escalables usan coordinate-descent [Friedman et al'08]

- Garantía de performance: Graphical lasso con $\lambda = 2\sqrt{\frac{\log N}{P}}$ satisface

$$\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}_0\|_2 \leq \sqrt{\frac{d_{\text{máx}}^2 \log N}{P}} \quad \text{w.h.p.}$$

  $\Rightarrow$ donde $\mathbf{\Theta}_0$ es el ground-truth, y grado máximo $d_{\text{máx}}$

- Consistencia del soporte para $P = \Omega(d_{\text{máx}}^2 \log N)$ [Ravikumar et al'11]

GMRFs con restricciones de Laplaciano

- Graphical model selection with Laplacian constraints $\mathbf{\Theta} = \mathbf{L}$
  - Entradas fuera de diagonal $\theta_{ij} = L_{ij} = -A_{ij} \leq 0 \Rightarrow$ GMRF atractivo
  - El Laplaciano es singular ($\mathbf{L1} = \mathbf{0}$) $\Rightarrow$ GMRF impropio
- Estimar un GMRF propio agregando algo en la diagonal [Lake-Tenembaum'07]

$$\begin{aligned} \underset{\mathbf{\Theta} \succeq \mathbf{0}, \gamma \geq 0}{\text{máx}} & \left\{ \log\det \mathbf{\Theta} - \operatorname{trace}(\hat{\mathbf{\Sigma}}\mathbf{\Theta}) - \lambda \|\mathbf{\Theta}\|_1 \right\} \\ \text{s. to } & \mathbf{\Theta} = \mathbf{L} + \gamma \mathbf{I} \\ & \mathbf{L1} = \mathbf{0},\ L_{ij} \leq 0,\ i \neq j \end{aligned}$$

$\Rightarrow$ Se interpreta γ^{-1} como la varianza de fluctuaciones isotrópicas gaussianas

- Favorece grafos sobre los cuales las señales son suaves (más sobre esto la clase siguiente)

$$\operatorname{trace}(\hat{\mathbf{\Sigma}}\mathbf{L}) \propto \sum_{p=1}^{P} \mathbf{x}_p^\top \mathbf{L} \mathbf{x}_p = \sum_{p=1}^{P} \operatorname{TV}(\mathbf{x}_p)$$

Covariance selection meets linear regression

- **Idea:** Estimar de forma independiente vecindades $\mathcal{N}_i := \{j : (i,j) \in \mathcal{E}\}$, $i \in \mathcal{V}$
- El valor esperado condicional de x_i dado $\mathbf{x}_{\setminus i} := [x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N]^\top$ es

$$\mathbb{E}\left[x_i \,\middle|\, \mathbf{x}_{\setminus i}\right] = \mathbf{x}_{\setminus i}^\top \boldsymbol{\beta}^{(i)} \qquad \text{(es lineal)}$$

- Las entradas de $\boldsymbol{\beta}^{(i)}$ se pueden expresar en términos de $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$, a saber

$$\beta_j^{(i)} = -\frac{\theta_{ij}}{\theta_{ii}}$$

  $\Rightarrow$ valor no nulo $\beta_j^{(i)} \Leftrightarrow$ valor no nulo θ_{ij} en $\boldsymbol{\Theta} \Leftrightarrow$ Arista (i,j) in G

  $\Rightarrow$ En otras palabras, $\text{supp}(\boldsymbol{\beta}^{(i)}) := \{j : \beta_j^{(i)} \neq 0\} \equiv \mathcal{N}_i$

- Sugiere inferencia de G via regresión de mínimos cuadrados (LS), pues

$$\boldsymbol{\beta}^{(i)} = \arg\min_{\boldsymbol{\beta}} \mathbb{E}\left[(x_i - \mathbf{x}_{\setminus i}^\top \boldsymbol{\beta})^2\right], \quad i \in \mathcal{V}$$

Neighborhood-based sparse regression

- Recorrer los nodos $i \in \mathcal{V}$ y estimar $\hat{\mathcal{N}}_i = \text{supp}(\hat{\boldsymbol{\beta}}^{(i)})$, donde

$$\hat{\boldsymbol{\beta}}^{(i)} \in \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^{N-1}} \left\{ \sum_{p=1}^{P} (x_{pi} - \mathbf{x}_{p,\setminus i}^{\top} \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}$$

  $\Rightarrow$ Problemas lasso separados por cada nodo

- No hay garantía que $\hat{\beta}_j^{(i)} \neq 0$ implica $\hat{\beta}_i^{(j)} \neq 0$ y vice versa

  $\Rightarrow$ Combina información en $\hat{\mathcal{N}}_i$ y $\hat{\mathcal{N}}_j$ para forzar simetría

  $\Rightarrow$ Regla OR: $(i,j) \in \mathcal{E}$ si $\beta_j^{(i)} \neq 0$ o $\beta_i^{(j)} \neq 0$. De igual forma, regla AND

- Consistencia del soporte para ambas reglas [Meinshausen-Bühlmann'06]
  - con cierta elección de λ, esparsidad de $\boldsymbol{\Theta}_0$, y relación de tamaño y muestras $P \ll N$

Mapa conceptual para GMRF model selection

Testing partial correlations

For each $(i,j) \in \mathcal{V} \times \mathcal{V}$, test the hypothesis

$$H_0 : \rho_{ij|\mathcal{V}\setminus ij} = 0 \quad \text{versus} \quad H_1 : \rho_{ij|\mathcal{V}\setminus ij} \neq 0$$

$$\rho_{ij|\mathcal{V}\setminus ij} = -\frac{\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}} \;\rightarrow\; \rho_{ij|\mathcal{V}\setminus ij} \neq 0 \Leftrightarrow \theta_{ij} \neq 0$$

Covariance selection

Infer non-zero entries $\theta_{ij} \neq 0$ of the precision matrix

$$\boldsymbol{\Theta} := \boldsymbol{\Sigma}^{-1}$$

$$\beta_j^{(i)} = -\frac{\theta_{ij}}{\theta_{ii}} \;\rightarrow\; \beta_j^{(i)} \neq 0 \Leftrightarrow \theta_{ij} \neq 0$$

Neighborhood-based regression

For each $i \in \mathcal{V}$, infer non-zero regression coefficients $\beta_j^{(i)} \neq 0$ in

$$\boldsymbol{\beta}^{(i)} = \arg\min_{\boldsymbol{\beta}} \mathbb{E}\left[(x_i - \mathbf{x}_{\setminus i}^T \boldsymbol{\beta})^2\right]$$

Resumen comparativo

- Paralelizable neighborhood-based regression (NBR)
  - $\Rightarrow$ Verosimilitud condicionales por nodo $i \in \mathcal{V}$
  - $\Rightarrow$ no fuerza $\mathbf{\Theta} \succeq \mathbf{0}$
  - $\Rightarrow$ Tiende a ser más rápido computacionalmente
- Graphical Lasso minimiza una verosimilitud global (regularizada)

$$\mathcal{L}(\mathbf{\Theta}; \mathcal{X}) = \log\det \mathbf{\Theta} - \operatorname{trace}(\hat{\mathbf{\Sigma}}\mathbf{\Theta})$$

  - $\Rightarrow$ Tiende a ser estadísticamente más eficiente
- Método NBR se puede adaptar para graphical models discretos o mixtos
  - $\Rightarrow$ *Ising-model selection* para $\mathbf{x} \in \{-1, +1\}^N$ [Ravikumar'10]