Aprendizaje Automático para Datos en Grafos

Laboratorio 4

Graciana Castro
4.808.848-2
gcastro@fing.edu.uy

Julian O'Flaherty
6.285.986-9
julian.o.flaherty@fing.edu.uy

1. Introducción

2. Inferencia de topología en datos sintéticos

2.1. Generación del grafo sintético

Para generar un grafo sintético, utilizaremos el siguiente algoritmos:

1. Sorteamos N puntos de forma uniforme en el cuadrado $[0, 1] \times [0, 1]$. Esos serán los vértices de nuestro grafo.
2. Para cada par de puntos i y j que sorteamos antes, tomamos como peso de la arista

$$w_{ij} = \begin{cases} 0 & \text{si } i = j \\ e^{-\frac{d(i,j)}{2\sigma^2}} & \text{si } i \neq j \end{cases}$$

   donde $d(i, j)$ es la distancia euclidea en $\mathbb{R}^2$ y σ es un parámetro fijo.
3. Descartamos las aristas cuyo peso w_{ij} sea menor que un número fijo $r > 0$.

En la figura 1 se muestra un grafo sintético generado con el algoritmo anterior para $N = 50$, $\sigma = 0{,}5$ y $r = 0{,}6$.

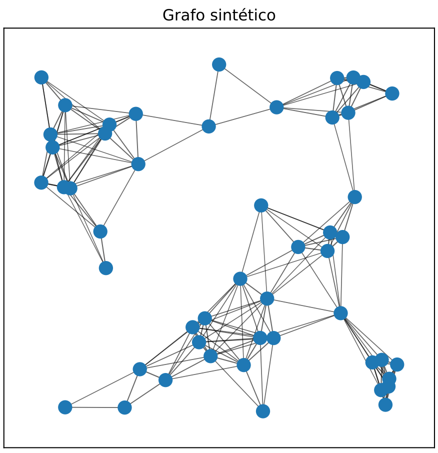


Figura 1: Grafo sintético generado con $N = 50$, $\sigma = 0{,}5$ y $r = 0{,}6$.

A cada nodo del grafo sintetico, le asignaremos una señal $x_i \sim \mathcal{N}(0, \mathbf{L}^\dagger)$, donde $\mathbf{L}^\dagger$ es la pseudoinversa de la matriz laplaciana $\mathbf{L}$ del grafo. En la figura 2 se muestra un sample de señales donde cada columna j es la señal asociada al nodo j del grafo. La cantidad de filas, o el tamaño de la muestra, es el parámetro n_{samples} del algoritmo, por lo que si $n_samples$ es menor que N, la matriz de covarianza empírica no es invertible.

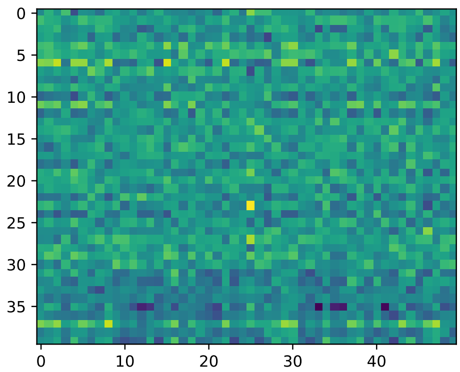


Figura 2: Sample de señales asignadas a los nodos del grafo sintético.

2.2. Estimación de estructura con Graphical Lasso

Estimaremos la estructura del grafo a partir de la matriz de datos utilizando el Graphical Lasso. Este método busca al estimador de máxima verosimilitud de la matriz de precisión $\mathbf{\Theta}$ que cumple:

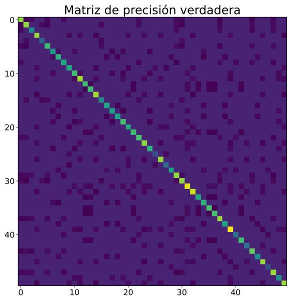
$$\hat{\mathbf{\Theta}} = \arg\max_{\mathbf{\Theta} \succeq 0} \left\{ \log\det \mathbf{\Theta} - \operatorname{trace}(\hat{\mathbf{\Sigma}}\mathbf{\Theta}) - \lambda \|\mathbf{\Theta}\|_1 \right\} \tag{1}$$

Una de las propiedades importantes del Graphical Lasso, es que cuando $\lambda = 2\sqrt{\frac{\log N}{P}}$,

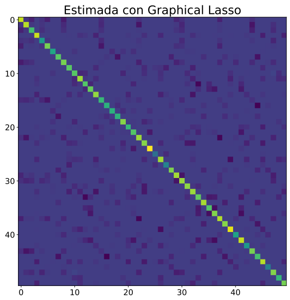
$$\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}_0\|_2 \leq \sqrt{\frac{d_{\text{máx}}^2 \log N}{P}} \quad \text{w.h.p.}$$

En la figura 3 se puede observar la estimación obtenida con *Graphical Lasso* comparada con la estimación empírica cuando el número de samples utilizado n_{samples} es menor que la cantidad de nodos N. En este caso, la estimación empírica resulta mala, dado que la matriz de covarianza empírica no es invertible. El método de *Graphical Lasso* obtiene un resultado muy similar a la matriz de precisión verdadera, donde el valor de λ se calcula mediante cross validation.

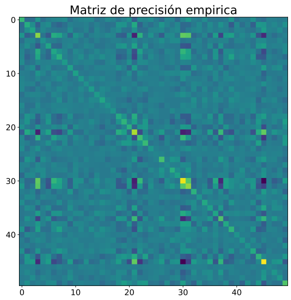
En la figura 4 observamos el efecto del variar el valor de lambda sobre la matriz de precisión estimada. A medida que aumenta la regularización, la matriz va perdiendo zeros tendiendo a una matriz diagonal, lo cual es esperable dado que el valor de lambda esta asociado a la norma 1 de la matriz de precisión.



(a) Precisión verdadera



(b) Estimación mediante Graphical Lasso



(c) Estimación Graphical Lasso

Figura 3: Comparación entre la matriz de precisión verdadera, la estimación mediante Graphical Lasso y la estimación obtenida con empíricamente ($\Theta = \Sigma^{-1}$), cuando $n_{\text{samples}} = 40$ y $N = 50$

2.3. Estimación de estructura con Meinshausen y Bühlmann

2.4. Estimación de estructura con Kalofolias

2.5. Comparación y visualización de los grafos aprendidos

2.6. Discusión

3. Inferencia de topología en MNIST

3.1. Descripción del dataset y preprocesamiento

3.2. Construcción de la matriz de features y t-SNE

3.3. Grafo aprendido con Meinshausen y Bühlmann

3.4. Grafo aprendido con Kalofolias y umbralización

3.5. Clustering espectral sobre grafos aprendidos

3.5.1. Métricas de evaluación

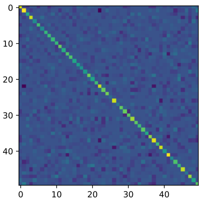
3.5.2. Resultados y comparación

3.6. Visualización en el plano t-SNE por método

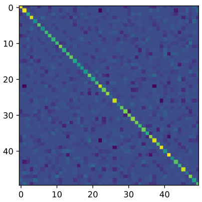
3.7. Análisis de aciertos y errores por método

4. Conclusiones

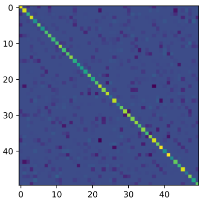
5. Referencias



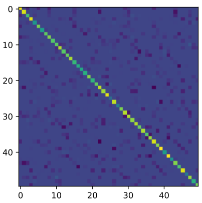
(a) $\lambda = 0{,}01$



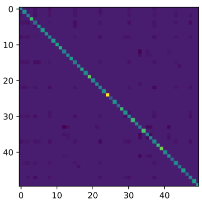
(b) $\lambda = 0{,}02$



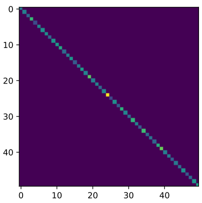
(c) $\lambda = 0{,}05$



(d) $\lambda = 0{,}1$

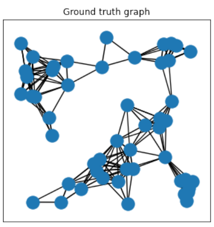


(e) $\lambda = 0{,}5$

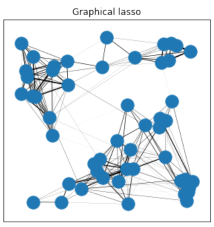


(f) $\lambda = 1$

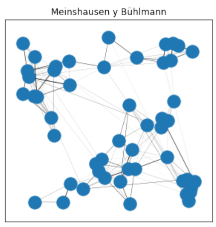
Figura 4: Estimación mediante Graphical Lasso para distintos valores de λ. La matriz de precisión verdadera se muestra en la figura 3a. No se muestran valores menores a 0,01 porque hay poca diferencia notable, y a partir de 0,005 el sistema queda mal condicionado y no puede ser resuelto.



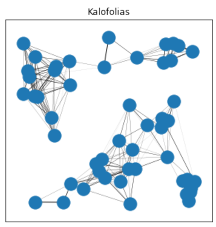
(a) Grafo verdadero



(b) Graphical Lasso



(c) Meinshausen y Bühlmann



(d) Kalofolias

Figura 5: Comparación visual de los grafos: verdadero y aprendidos por distintos métodos.