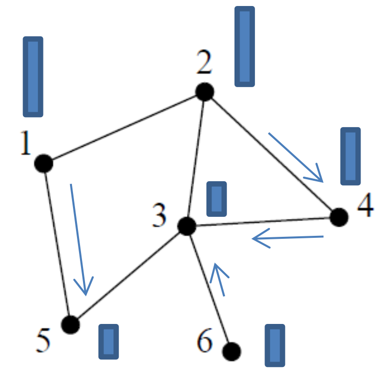
Laplacianos y aprendizaje semisupervisado

Difusión en redes

Pensamos en un **proceso difusivo** en el que el flujo desde el *nodo-j* al *nodo-i* es **proporcional** a la diferencia de material: *flujo*$_{j\to i} \sim C * (x_j - x_i)$

cte de difusión



$$\frac{dx_i}{dt} = C\sum_{j=1}^{N} A_{ij}\,(x_j - x_i)$$

$$= C\sum_{j=1}^{N} A_{ij}x_j - C\underbrace{\sum_{j=1}^{N} A_{ij}}_{k_i}x_i$$

$$= C\sum_{j=1}^{N} A_{ij}x_j - Ck_ix_i$$

$$= C\sum_{j=1}^{N} A_{ij}x_j - C\delta_{ij}k_jx_j$$

$$\frac{dx_i}{dt} = -C\sum_{j=1}^{N}\underbrace{(\delta_{ij}k_j - A_{ij})}_{[L]_{ij}}x_j$$

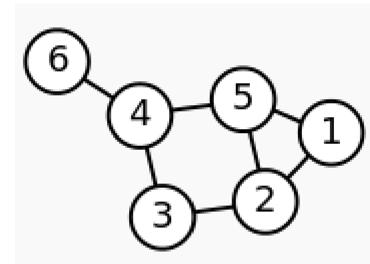
$$\frac{d\boldsymbol{x}}{dt} = -C.L\,\boldsymbol{x}$$

$$\frac{d\boldsymbol{x}}{dt} = C\nabla^2\boldsymbol{x}$$

Ec de difusion de calor

Difusión en redes

Pensamos en un proceso difusivo en el que el flujo desde el *nodo-j* al *nodo-i* es **proporcional** a la diferencia de material: *flujo*$_{j\text{->}i}$ ~ $C * (x_j - x_i)$

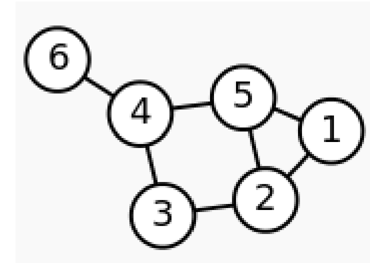


cte de difusión

$$\frac{dx_i}{dt} = -C\sum_{j=1}^{N}\underbrace{(\delta_{ij}k_j - A_{ij})}_{[L]_{ij}}x_j \qquad\qquad L = D - A$$

Degree matrix	Adjacency matrix	Laplacian matrix
$\begin{pmatrix} 2&0&0&0&0&0\\ 0&3&0&0&0&0\\ 0&0&2&0&0&0\\ 0&0&0&3&0&0\\ 0&0&0&0&3&0\\ 0&0&0&0&0&1 \end{pmatrix}$	$\begin{pmatrix} 0&1&0&0&1&0\\ 1&0&1&0&1&0\\ 0&1&0&1&0&0\\ 0&0&1&0&1&1\\ 1&1&0&1&0&0\\ 0&0&0&1&0&0 \end{pmatrix}$	$\begin{pmatrix} 2&-1&0&0&-1&0\\ -1&3&-1&0&-1&0\\ 0&-1&2&-1&0&0\\ 0&0&-1&3&-1&-1\\ -1&-1&0&-1&3&0\\ 0&0&0&-1&0&1 \end{pmatrix}$

Propiedades



Laplacian matrix

$$\begin{pmatrix} 2 & -1 & 0 & 0 & -1 & 0 \\ -1 & 3 & -1 & 0 & -1 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ -1 & -1 & 0 & -1 & 3 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix}$$

$$L = D - A$$

- *L* es simétrica
- *L* es semidefinida-positiva (i.e. autovalores $\lambda_i \geq 0, i = 1 \ldots N$)
- Suma por columnas y por filas de *L* es cero
- Siempre $\lambda = 0$ es autovalor de L con autovector $\boldsymbol{v_0}$=(1,1,…,1), y que $L\,\boldsymbol{v_0} = 0\,\boldsymbol{v_0} = 0$
- *L* es una matriz singular (i.e. no inversible)

Propiedades

$$L = D - A$$

Laplaciano **combinatorio**

- *L* es simétrica
- *L* es semidefinida-positiva (i.e. autovalores $\lambda_i \geq 0$, $i = 1 \ldots N$)
- Suma por columnas y por filas de *L* es cero
- Siempre $\lambda = 0$ es autovalor de L con autovector $\boldsymbol{v_0}=(1,1,\ldots,1)$, ya que $L\,\boldsymbol{v_0} = 0\,\boldsymbol{v_0} = 0$
- *L* es una matriz singular (i.e. no inversible)
- Si el grafo tiene *p* componentes, de tamaños $n_1, n_2, \ldots, n_p$

$$\mathbf{L} = \begin{pmatrix} \boxed{L_1} & 0 & \cdots \\ 0 & \boxed{L_2} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}.$$

Notar que ahora habrá p autovectores de *L* asociados a $\lambda = 0$ de la forma

$$(\underbrace{1, \ldots, 1}_{n_1}, 0,0,\ldots,0)$$

$$(\underbrace{0, \ldots, 0}_{n_1}, \underbrace{1, \ldots, 1}_{n_2}, 0,0,\ldots,0)$$

$$(\underbrace{0, \ldots, 0}_{n_1}, \underbrace{0, \ldots, 0}_{n_2}, \underbrace{1, \ldots, 1}_{n_3}, 0,0,\ldots,0)$$

Propiedades

$$L = D - A$$

Laplaciano **combinatorio**

- *L* es simétrica
- *L* es semidefinida-positiva (i.e. autovalores $\lambda_i \geq 0, i = 1 \dots N$)
- Suma por columnas y por filas de *L* es cero
- Siempre $\lambda = 0$ es autovalor de L con autovector $\boldsymbol{v_0}=(1,1,\dots,1)$, ya que $L\, \boldsymbol{v_0} = 0\, \boldsymbol{v_0} = 0$
- *L* es una matriz singular (i.e. no inversible)
- Si el grafo tiene *p* componentes, de tamaños $n_1, n_2, \dots, n_p$, existen p autovectores de *L* asociados a $\lambda = 0$

Corolario:

- el segundo autovalor de L es $\lambda_2 \neq 0$ sii el grafo posee una única componente
- λ_2 se denomina **conectividad algebraica** de la red

Autovectores de L

(a) a linear unweighted graph with two segments

$$L\boldsymbol{\phi}_i = \lambda_i \boldsymbol{\phi}_i$$

autovector i-esimo

$\boldsymbol{\phi}_i$ define un campo escalar sobre el grafo

Notar: autovectores de bajo índice están asociados a **campos más suaves** sobre el grafo

Caminatas al azar…y difusión

$$p_i(t+1) = \sum_{j=1}^{N} \frac{1}{k_j} a_{ij}\, p_j(t)$$

$p_i(t)$ probabilidad de encontrar a Juan en el nodo-*i*, en el paso temporal *t*

- Esta misma ecuación aplica a procesos de tipo **difusivos** en una red, donde en cada paso temporal toda la cantidad de material que se encuentra en el nodo-i es repartida y enviada a sus vecinos

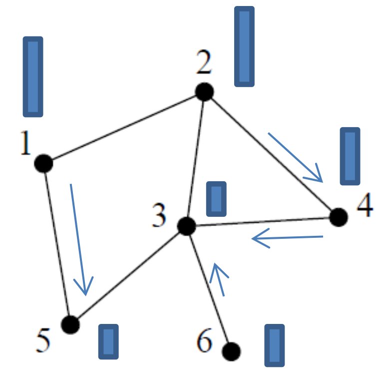
$$x_i(t+1) = \sum_{j=1}^{N} \frac{1}{k_j} a_{ij}\, x_j(t)$$

$x_j(t)$ cantidad de material que a tiempo *t* se encuentra en el nodo-*j*

el recurso x_j se reparte *extensivamente*

Difusión *random-walk* en redes

A cada paso, la cantidad x_i de cada nodo se altera



en cada nodo hay una cantidad x_i de material

$$x_i(t+1) = \sum_{j=1}^{N} \frac{1}{k_j} a_{ij}\, x_j(t)$$

$$\Delta x_i = x_i(t+1) - x_i(t) = \sum_{j=1}^{N} \frac{1}{k_j} a_{ij}\, x_j(t) - x_i(t)$$

$$= \sum_{j=1}^{N} \frac{1}{k_j} a_{ij}\, x_j(t) - \delta_{ij}\, x_j(t)$$

$\delta_{ij} = 1 \; sii \; i = j$

$$= -\sum_{j=1}^{N} \underbrace{\left(\delta_{ij} - \frac{1}{k_j} a_{ij}\right)}_{[L^{rw}]_{ij}} x_j(t)$$

Laplaciano *random-walk*

$$\Delta \boldsymbol{x} = -L^{rw} \boldsymbol{x} \qquad \frac{d\boldsymbol{x}}{dt} = -L^{rw} \boldsymbol{x} \qquad \frac{d\boldsymbol{x}}{dt} = \nabla^2 \boldsymbol{x}$$

Laplacianos

Entonces, vimos dos tipos de procesos difusivos:

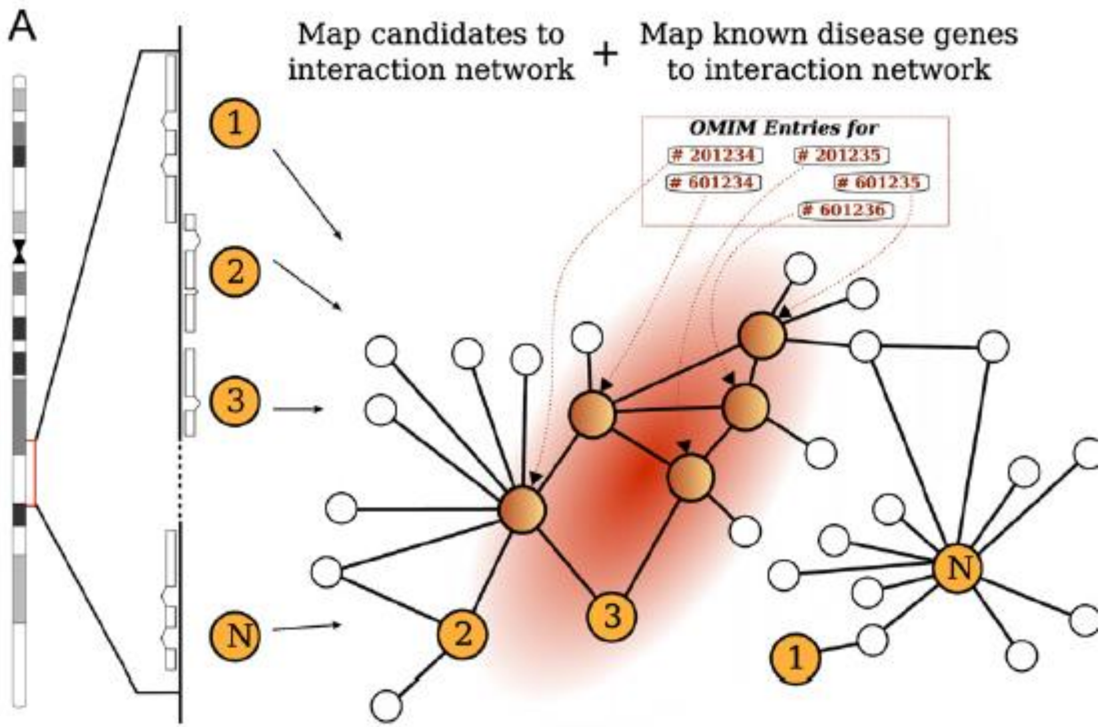
$$\frac{dx_i}{dt} = -C\sum_{j=1}^{N}\overbrace{\left(\delta_{ij}k_j - a_{ij}\right)}^{[L]_{ij}}x_j \qquad L = D - A$$

$$\frac{dx_i}{dt} = -\sum_{j=1}^{N}\overbrace{\left(\delta_{ij} - \frac{1}{k_j}a_{ij}\right)}^{[L^{rw}]_{ij}}x_j(t) \qquad L^{rw} = I - D^{-1}A = D^{-1}L$$

$$L^{sym} \equiv D^{-1/2}LD^{-1/2} = I - D^{-1/2}LD^{-1/2}$$

L : Laplaciano combinatorio o no-normalizado
L^{rw}: Laplaciano random-walk
L^{sym}: Laplaciano normalizado o simetrico

Priorización de nuevas asociaciones gen/enfermedad



Algoritmos para **propagar sentido de pertenecia** al conjunto de interés:

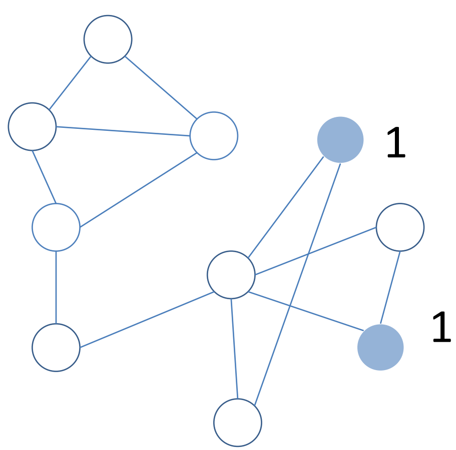
Random Walk with Restart (Kholer 2009)
Net-Rank (Chen 2009)
Net-Propagation (Vanunu 2010)
Functional Flow (Navieba 2005)

Lo podemos pensar también como un proceso de **difusión con fuentes**

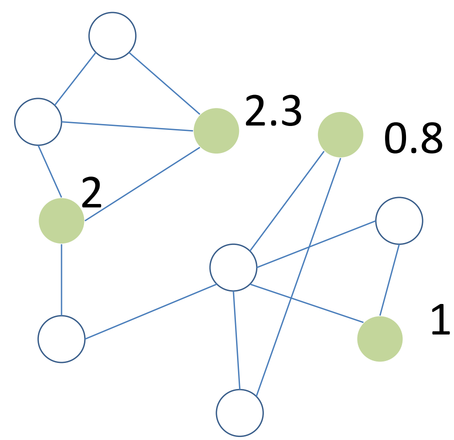
Kolher et al The American Journal of Human Genetics 2008

Difusión en redes y aprendizaje semi-supervisado

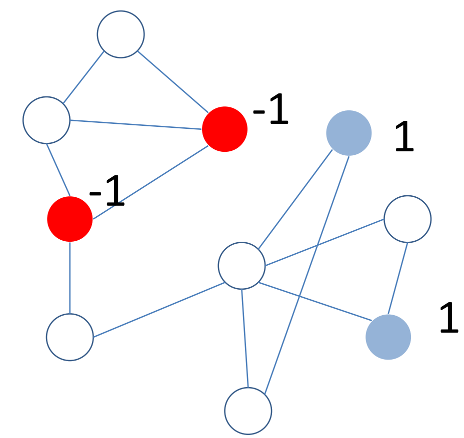
Tengo información parcial asociada a un subconjunto de nodos (etiquetas, o valores reales) y quiero utilizarla para inferir propiedades de los no-etiquetados Cada nodo va a propagar su etiqueta de manera iterativa hasta converger



Problema de priorización:
l nodos etiquetados con valor 1 y *N-l* nodos con valor 0.



Problema de regresión:
l nodos etiquetados con valores reales y *N-l* nodos con valor 0.

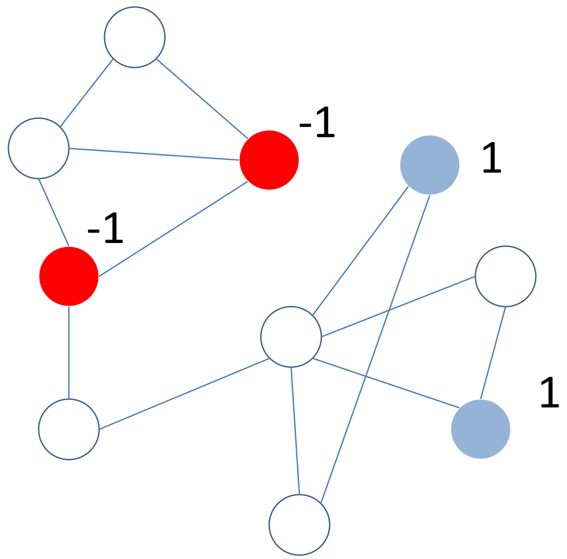


Problema de clasificación:
l nodos etiquetados con valor 1 (azul) o -1 (rojo) y *N-l* nodos con valor 0.

El problema de aprendizaje **semi-supervisado** consiste en encontrar un etiquetado de los nodos del grafo consistente con

I. El etiquetado inicial (incompleto)
II. La geometria inducida por la estructura de la red

Difusión en redes y aprendizaje semi-supervisado



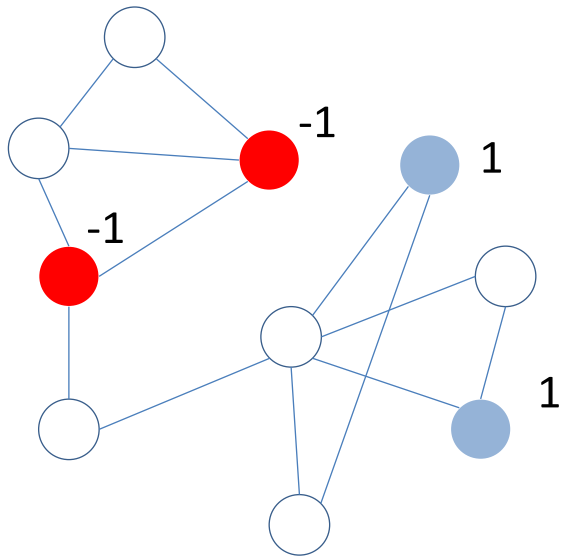
Algoritmo 1 Label Propagation (Zhu 2002)

- Computo de matriz de adyacencia W
- Computo de matriz diagonal D: $D_{ii} \leftarrow \sum_j w_{ij}$
- Inicializo $\hat{Y}^{(t=0)} \leftarrow (y_1, \ldots, y_l, 0,0, \ldots, 0)$
- Itero hasta convergencia
  1. $\hat{Y}^{(t+1)} \leftarrow D^{-1} W \hat{Y}^{(t)}$
  2. $\hat{Y}_l^{(t+1)} \leftarrow Y_l$
- Etiqueta del nodo-i resulta $sign(\hat{y}_i{}^{(\infty)})$

Sea el grafo G,

- nodos 1,2,…,l etiquetados no trivialmente segun $Y_l = (y_1, \ldots, y_l)$
- nodos l+1,…,N etiquetados con valor 0
- Queremos propagar la información por la red y estimar el vector de etiquetas asintótico: $\hat{Y} = (\hat{Y}_l, \hat{Y}_u)$

Difusión en redes y aprendizaje semi-supervisado



Algoritmo 2 Label Propagation

- Computo de matriz de adyacencia W, se fija $w_{ii}=0$
- Computo de matriz diagonal D: $D_{ii} \leftarrow \sum_j w_{ij}$
- Elijo $\epsilon > 0$ y $\alpha \in (0,1)$
  $$\mu \leftarrow \frac{\alpha}{1-\alpha} (0, +\infty)$$
- Computo la matriz diagonal $A_{ii} \leftarrow I_{[l]}(i) + \mu D_{ii} + \mu\varepsilon$
- Inicializo $\hat{Y}^{(t=0)} \leftarrow (y_1, \dots, y_l, 0,0, \dots, 0)$
- Itero hasta convergencia
  $$\hat{Y}^{(t+1)} \leftarrow A^{-1}(\mu W \hat{Y}^{(t)} + \hat{Y}^{(0)})$$
- Etiqueta del nodo-i resulta $sign({\hat{y}_i}^{(\infty)})$

para nodo etiquetado

$$\hat{y}_i^{(t+1)} \leftarrow \frac{\sum_j \mathbf{W}_{ij} \hat{y}_j^{(t)} + \frac{1}{\mu} y_i}{\sum_j \mathbf{W}_{ij} + \frac{1}{\mu} + \epsilon}$$

para nodo sin etiquetar

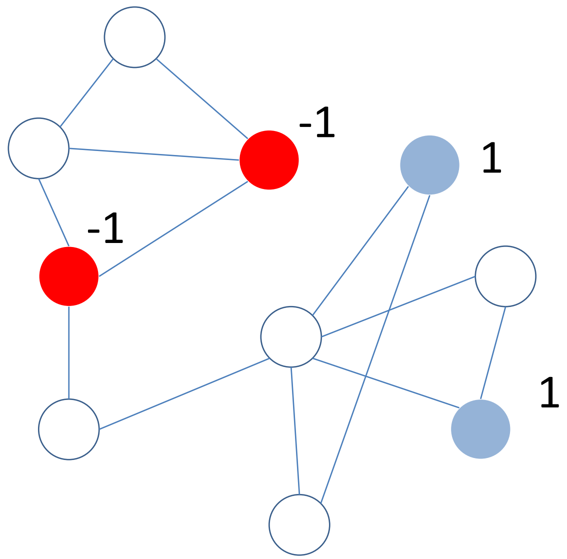
$$\hat{y}_i^{(t+1)} \leftarrow \frac{\sum_j \mathbf{W}_{ij} \hat{y}_j^{(t)}}{\sum_j \mathbf{W}_{ij} + \epsilon}$$

Diferencias con Algo 1

- Se fija $w_{ii}=0$
- Se permite $\hat{Y}_l \neq Y_l$
- Se considera un termino de regularización ϵ

$I_{[l]}$:matriz diagonal con 1's en los primeros l elementos y 0 el resto

Difusión en redes y aprendizaje semi-supervisado



Algoritmo 3 Label Propagation (Zhou 2004)

- Computo de matriz de adyacencia W, se fija $w_{ii}=0$
- Computo de matriz diagonal D: $D_{ii} \leftarrow \sum_j w_{ij}$
- Computo Laplaciano simetrico

$$\mathcal{L} \leftarrow D^{-1/2} W D^{-1/2}$$

- Inicializo $\hat{Y}^{(t=0)} \leftarrow (y_1, \dots, y_l, 0,0, \dots, 0)$
- Elijo $\alpha \in [0,1)$
- Itero hasta convergencia

Difusión con fuentes → $$\hat{Y}^{(t+1)} \leftarrow \alpha \mathcal{L} \hat{Y}^{(t)} + (1-\alpha) \hat{Y}^{(0)}$$

- Etiqueta del nodo-i resulta $sign({\hat{y}_i}^{(\infty)})$

Regularización en grafos

El problema de aprendizaje supervisado consiste en encontrar un etiquetado de los nodos del grafo consistente con

I. El etiquetado inicial (incompleto)
II. La geometria inducida por la estructura de la red

etiquetados

Sea $\hat{Y} = (\hat{Y}_l, \hat{Y}_u)$

no-etiquetados

penalizamos apartamiento de etiquetado original

Consistencia con etiquetado inicial:

$$\sum_{i=1}^{l} (\hat{y}_i - y_i)^2 = \|\hat{Y}_l - Y_l\|^2$$

Consistencia con geometría de los datos:

$$\begin{aligned}
\frac{1}{2}\sum_{i,j=1}^{n} \mathbf{W}_{ij}(\hat{y}_i - \hat{y}_j)^2 &= \frac{1}{2}\left(2\sum_{i=1}^{n}\hat{y}_i^2\sum_{j=1}^{n}\mathbf{W}_{ij} - 2\sum_{i,j=1}^{n}\mathbf{W}_{ij}\hat{y}_i\hat{y}_j\right) \\
&= \hat{Y}^\top(\mathbf{D} - \mathbf{W})\hat{Y} \\
&= \hat{Y}^\top L\hat{Y}
\end{aligned}$$

penalizamos cambios **bruscos** de etiquetado

Regularización en grafos

El problema de aprendizaje supervisado consiste en encontrar un etiquetado de los nodos del grafo consistente con

I. El etiquetado inicial (incompleto)
II. La geometria inducida por la estructura de la red

penalizamos apartamiento de etiquetado original

penalizamos cambios **bruscos** de etiquetado

$$\mathcal{H}(\hat{Y}) = \sum_{i=1}^{l} (\hat{y}_i - y_i)^2 + \frac{1}{2} \sum_{i,j=1}^{N} w_{ij} \left(\hat{y}_i - \hat{y}_j\right)^2$$

$$= \left\|\hat{Y}_l - Y_l\right\|^2 + \hat{Y}^T L \hat{Y}$$

Se trata de encontrar el etiquetado $\hat{Y}$ que minimice $\mathcal{H}(\hat{Y})$

En la practica se suele agregar un término de regularización para romper la simetría entre etiquetas

$$\mathcal{H}(\hat{Y}) = \left\|\hat{Y}_l - Y_l\right\|^2 + \mu \hat{Y}^T L \hat{Y} + \mu\epsilon \left\|\hat{Y}_l\right\|^2$$

Regularización en grafos

Se trata de encontrar el etiquetado $\hat{Y}$ que minimice $\mathcal{H}(\hat{Y})$

$$\mathcal{H}(\hat{Y}) = \left\|\hat{Y}_l - Y_l\right\|^2 + \mu \hat{Y}^T L \hat{Y} + \mu\epsilon\left\|\hat{Y}_l\right\|^2$$

$$\left\|S\hat{Y} - SY\right\|^2$$

$S = I_{[l]}$ matriz diagonal con 1's en los primeros l elementos y 0 el resto

$$\frac{1}{2}\frac{\partial\mathcal{H}(\hat{Y})}{\partial\,\hat{Y}} = S\left(\hat{Y} - Y\right) + \mu\, L\,\hat{Y} + \mu\epsilon\hat{Y} = (S + \mu\, L + \mu\epsilon I)\,\hat{Y} - SY = 0$$

$$\Rightarrow \hat{Y} = (S + \mu\, L + \mu\epsilon I)^{-1} SY$$

- Las etiquetas *optimas* pueden obtenerse invirtiendo una matriz
- Esta matriz depende únicamente del Laplaciano de la red (no de las etiquetas)
- La manera en que las etiquetas se propagan depende exclusivamente de la estructura del grafo

Regularización en grafos

Hay otras elecciones para la función a minimizar:

$$\mathcal{H}(\hat{Y}) = \left\|S\hat{Y} - SY\right\|^2 + \mu\hat{Y}^T L\hat{Y} + \mu\epsilon\left\|\hat{Y}_l\right\|^2$$

Zhou 2004

$$\mathcal{H}'(\hat{Y}) = \underbrace{\left\|\hat{Y} - SY\right\|^2}_{\left\|\hat{Y}_l - Y_l\right\|^2 + \left\|\hat{Y}_u\right\|^2} + \frac{\mu}{2}\sum_{i,j} w_{ij}\left(\frac{\hat{y}_i}{\sqrt{D_{ii}}} - \frac{\hat{y}_j}{\sqrt{D_{jj}}}\right)$$

Dos diferencias entre $\mathcal{H}\ (\hat{Y})$y $\mathcal{H}'(\hat{Y})$:

I. El primer término busca ajustar bien las etiquetas conocidas pero **además** favorece etiquetas 0 para nodos no-etiquetados inicialmente
II. Las etiquetas están normalizadas por la raiz cuadrada del grado cuando se computa similaridad entre vecinos.

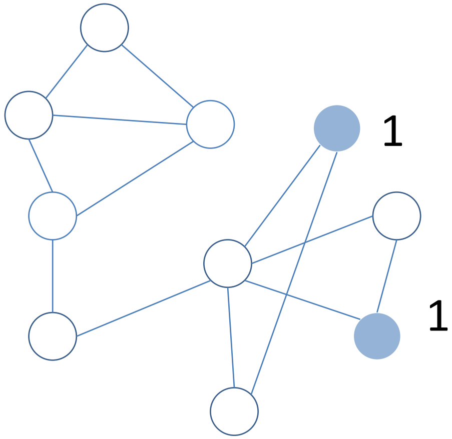
Se puede ver que

$$\mathcal{H}'(\hat{Y}) = \left\|\hat{Y} - SY\right\|^2 + \mu(D^{-1/2}\hat{Y})^T L(D^{-1/2}\hat{Y})$$

Vínculo entre regularización y propagación de etiquetas

$$\mathcal{H}(\hat{Y}) = \left\|S\hat{Y} - SY\right\|^2 + \mu \hat{Y}^T L \hat{Y} + \mu\epsilon\left\|\hat{Y}_l\right\|^2$$

$$\Rightarrow \hat{Y}^* = (S + \mu\, L + \mu\epsilon I)^{-1} SY$$



Metodo iterativo de Jacobi para invertir una matriz:

$$Mx = b$$

$$x_i^{(t+1)} = \frac{1}{M_{ii}}\left(b - \sum_{j \neq i} M_{ij} x_j^{(t)}\right)$$

En nuestro caso $x \equiv \hat{Y}, b \equiv SY, M = S + \mu\, L + \mu\epsilon I$ y la iteración de Jacobi resulta en

para nodo etiquetado

$$\hat{y}_i^{(t+1)} \leftarrow \frac{\sum_j \mathbf{W}_{ij} \hat{y}_j^{(t)} + \frac{1}{\mu} y_i}{\sum_j \mathbf{W}_{ij} + \frac{1}{\mu} + \epsilon}$$

para nodo sin etiquetar

$$\hat{y}_i^{(t+1)} \leftarrow \frac{\sum_j \mathbf{W}_{ij} \hat{y}_j^{(t)}}{\sum_j \mathbf{W}_{ij} + \epsilon}$$

Algoritmo 2 de propagación de etiquetas (!)

Regularización y propagación de etiquetas

$$\mathcal{H}'(\hat{Y}) = \left\|\hat{Y} - SY\right\|^2 + \mu(D^{-1/2}\hat{Y})^T L(D^{-1/2}\hat{Y})$$

$$\frac{1}{2}\frac{\partial\mathcal{H}'(\hat{Y})}{\partial\,\hat{Y}} = \hat{Y} - SY + \mu\,(\hat{Y} - \mathcal{L}\hat{Y})$$

$$\Rightarrow \hat{Y}^* = \left((1+\mu)I - \mu\,\mathcal{L}\right)^{-1}SY$$

La iteracion de Jacobi para resolver $\frac{\partial\mathcal{H}'(\hat{Y})}{\partial\,\hat{Y}} = 0$

$$\hat{Y}^{(t+1)} \leftarrow \alpha\mathcal{L}\hat{Y}^{(t)} + (1-\alpha)\hat{Y}^{(0)}$$

con $\mu = \frac{\alpha}{1-\alpha}$

Algoritmo 3 de propagación de etiquetas (!)

Conclusiones

- Vimos ejemplos donde es posible establecer un link entre una metodologia de propagacion de etiquetas y un proceso de minimizacion de una cantidad que da cuenta de
  - el ajuste a la informacion inicial (parcial)
  - suavidad del campo de etiquetas sobre la geometria de los datos

Refs

- Capitulo 11 de Semi-supervised learning, Chapelle et al, *Label propagation and quadratic criterion.*