Mathematics of Image and Data Analysis
Math 5467

Graph-based semi-supervised learning

Instructor: Jeff Calder
Email: jcalder@umn.edu

http://www-users.math.umn.edu/~jwcalder/5467

Last time

- Intro to Machine Learning

Today

- Graph-based semi-supervised learning.

Graph

Recall: Semi-supervised learning uses both labeled and unlabeled data to learn.

One way to use the unlabeled data is to build a graph, which is encoded into a weight matrix W.

- $W(i, j)$ is the similarity between i and j ($W(i, j) \geq 0$).
- We assume W is symmetric $W = W^T$. $W(i,j) = W(j,i)$
- Often can choose Gaussian weights (recall spectral clustering)

$$W(i, j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}.$$

Some datasets already have graph structure (citation databases, network problems, etc.).

Example graph

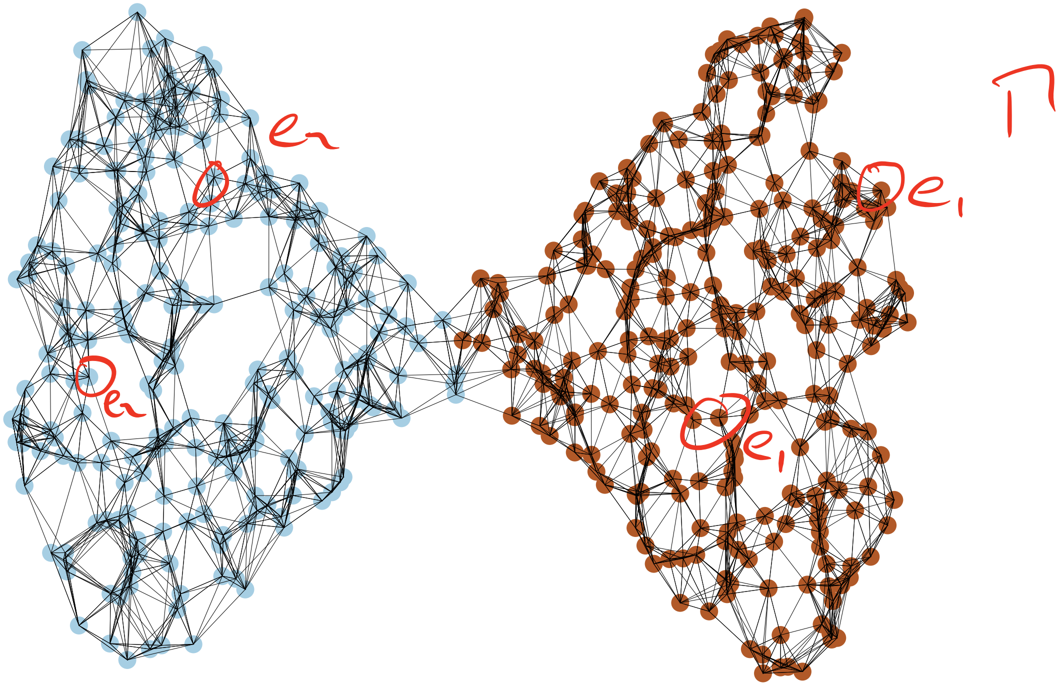

Figure 1: An example of a k-nearest neighbor graph.

Graph-based semi-supervised learning

Let $I_m = \{1, 2, \ldots, m\}$ denote the indices of all our datapoints.

We asssume there is a subset of the nodes $\Gamma \subset I_m$ that are asssigned label vectors from the one-hot vectors

$$S_k = \{e_1, e_2, \ldots, e_k\}.$$ one-hot vectors

We can treat the labels as a function $g : \Gamma \to S_k$, where $g(i)$ is the label of node $i \in \Gamma$.

$g(i) \in S_k \subseteq \mathbb{R}^k$

Task: Extend labels from the subset Γ to the rest of the graph is a meaningful way.

Laplacian regularization

It is common in practice to take the *semi-supervised smoothness assumption*, which stipulates that the learned labels should vary as smoothly as possible, and in particular, should not change rapidly within high density regions of the graph, which are likely to be clusters with the same label.

Laplacian regularized learning imposes the semi-supervised smoothness assumption by minimizing the function

$W(i,j) > 0 \Longleftrightarrow$ edge $i \longleftrightarrow j$

$$E(u) = \frac{1}{4}\sum_{i=1}^{m}\sum_{j=1}^{m} W(i,j)\|u(i) - u(j)\|^2 \tag{1}$$

over labeling functions $u : I_m \to \mathbb{R}^k$, subject to $u = g$ on Γ, that is, that the known labels are correct.

Gradient descent

$u(i) \in \mathbb{R}^k$

To minimize E we use gradient descent. Define the inner product for $u, v : I_m \to \mathbb{R}^k$ by

$$\langle u, v\rangle = \sum_{i=1}^{m} \underbrace{d(i)}_{\text{degree.}} \overbrace{u(i)^T v(i)}^{\text{Dot product}}, \qquad (2)$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{m} W(i,j)\, u(i)^T v(i)$$

where $d : I_m \to \mathbb{R}$ are the degrees, given by $d(i) = \sum_{j=1}^{m} W(i, j)$. The induced norm is

$$\|u\|^2 = \langle u, u\rangle = \sum_{i=1}^{m} d(i)\|u(i)\|^2.$$

We claim that $\nabla E(u) = d^{-1}Lu. = L_{rw} u.$

Recall:

$$\left.\frac{d}{dt}\right|_{t=0} E(u+tv) = \langle \underbrace{\nabla E(u)}_{\text{Def. of gradient}}, v\rangle$$

Gradient depends on choice of inner product.

$$E(u) = \frac{1}{4}\sum_{i=1}^{m}\sum_{j=1}^{m} W(i,j)\|u(i) - u(j)\|^2$$

$$\left.\frac{d}{dt}\right|_{t=0} E(u+tv) = \frac{1}{4}\sum_{i=1}^{m}\sum_{j=1}^{m} W(i,j)\frac{d}{dt}\|u(i)+tv(i) - (u(j)+tv(j))\|^2$$

$$= \frac{1}{4}\sum_{i=1}^{m}\sum_{j=1}^{m} W(i,j)\left.\frac{d}{dt}\right|_{t=0} \underbrace{\|u(i)-u(j)+t(v(i)-v(j))\|^2}_{\cancel{\|u(i)-u(j)\|^2} + 2t(u(i)-u(j))^T(v(i)-v(j)) + \cancel{t^2\|v(i)-v(j)\|^2}}$$

$t=$

(*)

$$= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} w(i,j)\,(u(i)-u(j))^T(v(i)-v(j)).$$

$$= \quad \text{want} = \langle \nabla E(u), v\rangle = \sum_{i=1}^{m} d(i)\,\nabla E(u)(i)\,v(i)$$

$$= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} w(i,j)\,(u(i)-u(j))^T v(i)$$

$$- \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} w(i,j)\,(u(i)-u(j))^T v(j)$$

swap $i \to j$

$$= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} w(i,j)\,(u(i)-u(j))^T v(i)$$

$$- \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} w(i,j)\,(u(j)-u(i))^T v(i)$$

$$= \sum_{i=1}^{m} d(i) \left[\frac{1}{d(i)} \underbrace{\sum_{j=1}^{m} w(i,j)\,(u(i)-u(j))^{T}}_{Lu(i)} \right] v(i)$$

$L =$ Unnormalized Graph Laplacian

$$L_{rw}\, u(i) = \frac{1}{d(i)}\, L u(i)$$

= Random walk graph Laplacian

$$= \sum_{i=1}^{m} d(i)\, L_{rw} u(i)\, v(i) = \langle L_{rw} u, v \rangle \quad (**)$$

Hence, by definition $\nabla E(u) = L_{rw} u$. $\blacksquare$

$$(*) = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} w(i,j) \left(u(i) - u(j)\right)^T \left(v(i) - v(j)\right)$$

$$(**) = \langle L_{rw} u, v \rangle$$

$$(*) \ \underline{u = v} \quad (*) = 2E(u).$$

$$E(u) = \frac{1}{2} \langle L_{rw} u, u \rangle.$$

Gradient descent

Using gradient descent to minimize E amounts to the iteration

$$u_{k+1} = u_k - dt\nabla E(u_k). \tag{3}$$

$$u_{k+1} = u_k - dt \cdot L_{rw} u_k \quad \text{on } I_m \setminus \Gamma$$

$$u_{k+1}(x) = g(x) \quad \text{for } x \in \Gamma$$

Lemma 1. *If $0 < dt \leq 1$ then for all $k \geq 1$ and $1 \leq i \leq m$ we have*

$$\|u_k(i)\| \leq \max_{1\leq i\leq m} \|u_0(i)\|. \tag{4}$$

$$u_0(i) = g(i)$$
$$i \in \Gamma$$

Proof:

$$u_{k+1}(i) = u_k(i) - \frac{dt}{d(i)} \sum_{j=1}^{m} w(i,j)\,(u_k(i) - u_k(j))$$

$$= u_k(i) - \frac{dt}{\cancel{d(i)}} u_k(i) \overbrace{\cancel{\sum_{j=1}^{m} w(i,j)}}^{d(i)}$$

$$+ \frac{dt}{d(i)} \sum_{j=1}^{m} w(i,j)\, u_k(j)$$

$$= (1-dt)\, u_k(i) + \frac{dt}{d(i)} \sum_{j=1}^{m} w(i,j)\, u_k(j)$$

Triangle inequality

$$\|u_{k+1}(i)\| \le |1-dt|\, \|u_k(i)\| + \left\| \frac{dt}{d(i)} \sum_{j=1}^{m} w(i,j)\, u_k(j) \right\|$$

$$\le |1-dt|\, \|u_k(i)\| + \frac{dt}{d(i)} \sum_{j=1}^{m} w(i,j)\, \|u_k(j)\|$$

$$\le |1-dt|\, \|u_k(i)\| + \frac{dt}{\cancel{d(i)}} \max_j \|u_k(j)\| \cancel{\sum_{j=1}^{m} w(i,j)}$$

$$\max_{1\le i\le m} \|u_0(i)\|.$$

$d(i)$

$$= |1-dt| \, \|u_k(i)\| + dt \max_j \|u_k(j)\|$$

$$\le \underbrace{\left(|1-dt| + dt\right)}_{\le 1 \text{ for stability}} \max_j \|u_k(j)\|$$

$$|1-dt| + dt \le 1$$

① $1 - dt + dt \le 1$ $\quad (dt \le 1)$

② $dt - 1 + dt \le 1$ $\quad (dt \ge 1)$

② $2\,dt \le 2 \Rightarrow dt \le 1$

In this case

$$\max_i \|u_{k+1}(i)\| \le \max_j \|u_k(j)\|$$

$$\le \dots \le \max_j \|u_0(j)\| \qquad \blacksquare$$

Convergence?

We may wish to go beyond stability and instead prove convergence of the iterations as $k \to \infty$ to a solution of the equation $\nabla E = 0$, that is

$$\frac{1}{d_i}\begin{cases} Lu(i) = 0, & \text{if } i \in I_m \setminus \Gamma \\ u(i) = g(i), & \text{otherwise.} \end{cases} \tag{5}$$

$\nabla E = 0$

- Depends on whether the graph is connected.
- If the graph is connected, (5) has a unique solution and gradient descent converges.
- If the graph is not connected, then (5) can have multiple solutions, and gradient descent will converge to one, but which one is dependent on initialization.

Convergence proof: let u solve (5).

$L_{rw} u = 0$ on $I_m \setminus \Gamma$

$$u_{k+1} = u_k - dt \cdot L_{rw} u_k$$

$$- \left(u = u - dt\, L_{rw} u \right)$$

$$u_{k+1} - u = u_k - u - dt\, L_{rw}(u_k - u)$$

Take inner product with $u_k - u$ on both sides

$$\langle u_{k+1} - u, u_k - u \rangle = \| u_k - u \|^2 - dt \langle L_{rw}(u_k - u), u_k - u \rangle$$

$E(u) = \frac{1}{2} \langle L_{rw} u, u \rangle.$

$$= \|u_k - u\|^2 - \frac{dt}{2} E(u_k - u).$$

<u>Note</u>: $u_k(i) = u(i) = g(i)$ on Γ

$u_k(i) - u(i) = 0$ on Γ

<u>Claim</u>: If $w(i) = 0$ on Γ

then there exists $\lambda > 0$ such

that $E(w) \geq \lambda \|w\|^2$ (*)

(λ depends only on the graph).

Proof $\quad \lambda = \min\limits_{\substack{w \in A \\ w \neq 0}} \dfrac{E(w)}{\|w\|^2}$

$$A = \left\{ w : I_m \longrightarrow \mathbb{R}^K : \begin{array}{l} w(i) = 0 \\ i \in \Gamma \end{array} \right\}.$$

Show $\lambda > 0$.

By $(*)$ we have $\quad E(u_K - u) \geq \lambda \|u_K - u\|^2$

$$\langle u_{k+1} - u, u_k - u \rangle \le \|u_k - u\|^2 - \frac{dt}{2} E(u_k - u)$$

$$\le \|u_k - u\|^2 - \frac{dt}{2} \lambda \|u_k - u\|^2$$

$$= \left(1 - \frac{dt}{2}\lambda\right) \|u_k - u\|^2$$

LHS

$$= \langle u_{k+1} - u, u_k + u_{k+1} - u_{k+1} - u \rangle$$

$$= \|u_{k+1} - u\|^2 + \langle u_{k+1} - u, u_k - u_{k+1} \rangle$$

So, we set

$$\|u_{k+1} - u\|^2 \leq \left(1 - \frac{dt}{2}\lambda\right)\|u_k - u\|^2 + \underbrace{\langle u_{k+1} - u, u_{k+1} - u_k\rangle}_{A}$$

Cauchy schwarz

$$|A| \leq \|u_{k+1} - u\| \cdot \|u_{k+1} - u_k\|$$

$$\leq \frac{1}{2}\|u_{k+1} - u\|^2 + \frac{1}{2}\|u_{k+1} - u_k\|^2$$

Cauchy Inequality $ab \le \frac{1}{2}a^2 + \frac{1}{2}b^2$

$(a-b)^2 \ge 0$

$$\frac{1}{2}\|u_{k+1}-u\|^2 \le \left(1-\frac{dt}{2}\lambda\right)\|u_k-u\|^2 + \underbrace{\frac{1}{2}\|u_{k+1}-u_k\|^2}_{\text{bit stuck}}$$

Fixed proof: Recall

$$u_{k+1} - u = u_k - u - dt\, L_{rw}(u_k - u)$$

Set $e_k = u_k - u$. Then

$$e_{k+1} = e_k - dt\, L_{rw} e_k$$

Take norms on both sides:

$$\|e_{k+1}\|^2 = \|e_k - dt\, L_{rw} e_k\|^2$$

$$= \|e_k\|^2 - 2dt\langle L_{rw} e_k, e_k\rangle + dt^2 \|L_{rw} e_k\|^2$$

Recall:

$$E(u) = \frac{1}{2}\langle L_{rw}u, u\rangle. \quad \text{and} \quad \underline{E}(w) \geq \lambda \|w\|^2 \text{ if } w(i) = 0 \text{ for } i \in \Gamma.$$

$\lambda > 0$

Since $\underbrace{u_k(i) - u(i)}_{e_k(i)} = 0$ for $i \in \Gamma$, we have

$$\langle L_{rw} e_k, e_k \rangle = 2E(e_k) \geq 2\lambda \|e_k\|^2$$

Hence

$$\|e_{k+1}\|^2 \leq \|e_k\|^2 - 4\lambda dt \|e_k\|^2 + dt^2 \|L_{rw} e_k\|^2$$

$$= (1 - 4\lambda dt)\|e_k\|^2 + dt^2 \|L_{rw} e_k\|^2$$

Note that

$$\|L_{rw} e_k\|^2 = \sum_{i=1} \left\| \frac{1}{d(i)} \sum_{j=1} w(i,j)\,(e_k(i) - e_k(j)) \right\|$$

$$\le \sum_{i=1}^{m} \frac{1}{d(i)} \sum_{j=1}^{m} w(i,j)\, \|e_k(i) - e_k(j)\|^2$$

$$\le \sum_{i=1}^{m} \frac{1}{d(i)} \sum_{j=1}^{m} w(i,j) \left(\|e_k(i)\| + \|e_k(j)\| \right)^2$$

$$\le 2 \sum_{i=1}^{m} \sum_{j=1} \frac{w(i,j)}{d(i)} \left(\|e_k(i)\|^2 + \|e_k(j)\|^2 \right)$$

$$= 2 \sum_{i=1}^{m} \|e_k(i)\|^2 + 2 \sum_{j=1}^{m} \|e_k(j)\|^2 \sum_{i=1}^{m} \frac{w(i,j)}{d(i)}$$

$$= 2(1 + C)\|e_k\|^2, \quad \text{where } C = \max_j \sum_{i=1}^{m} \frac{w(i,j)}{d(i)}$$

Therefore

$$\|e_{k+1}\|^2 \le \left(1 - 4\lambda dt + 2(1+c)dt^2\right) \|e_k\|^2$$

Choose dt small enough so that

$$2(1+c)\,dt^2 \le 2\lambda\, dt$$

$$\longrightarrow \quad dt \le \frac{\lambda}{1+c}$$

Then $\|e_{k+1}\|^2 \le (1-2\lambda\, dt)\|e_k\|^2$

By induction $\longrightarrow \quad \|e_k\|^2 \le (1-2\lambda\, dt)^k \|e_0\|^2$

Linear convergence rate.

Clasification of MNIST digits

We use a k-nearest neighbor graph with $k = 10$ and weights given by

$$W(i,j) = \exp\left(-\frac{4\|x_i - x_j\|^2}{d_k(x_i)^2}\right),$$

where $d_k(x_i)$ is the distance to the k^{th} nearest neighbor. The matrix is then symmetrized $W = W + W^T$.

10	**20**	**40**	**80**	**160**
85.4 (4.4)	91.7 (1.2)	93.4 (0.5)	94.3 (0.3)	94.8 (0.1)

Table 1: Laplace learning on MNIST with 10, 20, 40, 80, and 160 labels per class. The average (standard deviation) classification accuracy over 100 trials is shown.

GraphLearning (.ipynb)