Semi-Supervised Learning by Higher Order Regularization

Xueyuan Zhou[1] and Mikhail Belkin[2]

[1]Computer Science, University of Chicago

[2]Computer Science and Engineering, The Ohio State University

AISTATS 2011, Ft. Lauderdale, FL, USA

Outline

- Semi-Supervised Learning
- Graph Laplacian Regularization
- Problem of Graph Laplacian Regularization
- Solution: Iterated Laplacian Regularization
- Experiments
- Summary

Semi-Supervised Learning

- Semi-Supervised Learning (SSL)
  - A few labeled examples, with an additional very large unlabeled example set (partially observed data)
- Goal
  - Inductive inference: estimate an unknown function over the whole domain
  - Transductive inference: estimate the values of an unknown function at particular points

Semi-Supervised Learning

- Semi-Supervised Learning (SSL)
  - Given $X_L = \{x_1, x_2, \cdots, x_l\}$
    $Y_L = \{y_1, y_2, \cdots, y_l\}$
    $X_U = \{x_{l+1}, x_{l+2}, \cdots, x_{l+u}\}$
    $|X_L| << |X_U|, X = X_L \bigcup X_U, n = l + u$
  - SSL: find a function *f(x)* on the whole domain (inductive), or on the unlabeled set $Y_U = \{y_{l+1}, y_{l+2}, \cdots, y_{l+u}\}$ (transductive)

Semi-Supervised Learning

- Semi-Supervised Learning (SSL)
  - Given $X_L = \{x_1, x_2, \cdots, x_l\}$
    $Y_L = \{y_1, y_2, \cdots, y_l\}$
    $X_U = \{x_{l+1}, x_{l+2}, \cdots, x_{l+u}\}$
    $|X_L| << |X_U|, X = X_L \bigcup X_U, n = l + u$
  - SSL: find a function *f(x)* on the whole domain (inductive), or on the unlabeled set $Y_U = \{y_{l+1}, y_{l+2}, \cdots, y_{l+u}\}$ (transductive)
  - Model: $x_i \in \Omega \subset \mathbb{R}^d$, ($\Omega$ can be a manifold $\mathcal{M}$)
    $x_i \sim p(x), (x_i, y_i) \sim p(x, y)$
    $0 < a \leq p(x) \leq b < +\infty$

Semi-Supervised Learning

- Examples
  - Internet search queries/text/image classification
    - x:query/text document/image, y:predefined class
  - Estimation of functions on social network: Linkedin
    - x: user profile, y: e.g., interested in machine learning job
  - Information retrieval: image ranking as regression
    - x: images, y: ranking score
  - etc

Graph Laplacian Regularization

- A popular algorithm: find a "smooth" function
  - Similar x ➔ similar y

$$\min_{f \in \mathbb{R}^n} \sum_{x_i, x_j \in X} w(x_i, x_j)(f(x_i) - f(x_j))^2$$
$$s.t. \quad f(X_L) = Y_L$$

  - Similarity weight $w(x, y)$
  - Alternatively $\min_{f \in \mathbb{R}^n} \sum_{x_i \in X_L} (f(x_i) - y_i)^2 + \mu \sum_{x_i, x_j \in X} w(x_i, x_j)(f(x_i) - f(x_j))^2$

Graph Laplacian Regularization

- A popular algorithm: find a "smooth" function
  - Similar x ➔ similar y

$$\min_{f \in \mathbb{R}^n} \sum_{x_i, x_j \in X} w(x_i, x_j)(f(x_i) - f(x_j))^2$$
$$s.t. \quad f(X_L) = Y_L$$

  - Similarity weight $w(x, y)$
  - Alternatively $\min_{f \in \mathbb{R}^n} \sum_{x_i \in X_L} (f(x_i) - y_i)^2 + \mu \sum_{x_i, x_j \in X} w(x_i, x_j)(f(x_i) - f(x_j))^2$

- Estimation on an undirected graph G(V,E)
  - $V = X$
  - $E = \{e_{ij} : (x_i, x_j)\}, w(x_i, x_j) = \frac{1}{t^{d/2}} e^{-\frac{\|x_i - x_j\|^2}{t}}$
  - E.g., *k*NN graphs

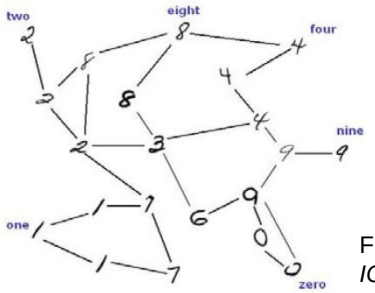


Figure from, X. Zhu et al., *ICML* 2003.

Graph Laplacian Regularization

- Graph Laplacian Regularizer as a Semi-norm

$$\frac{1}{2}\sum_{x_i,x_j\in X} w(x_i,x_j)(f(x_i)-f(x_j))^2 = f^T L f$$

where $L = D - W$ is called the graph Laplacian

Graph Laplacian Regularization

- Graph Laplacian Regularizer as a Semi-norm

$$\frac{1}{2}\sum_{x_i,x_j\in X} w(x_i,x_j)(f(x_i)-f(x_j))^2 = f^T L f$$

where $L = D - W$ is called the graph Laplacian

- Analysis in An Asymptotic Setting
  - Fixed labeled set, increasing unlabeled points
- Limit as *n* increases, *t* decreases

$$\frac{1}{n^2 t^{(d/2+1)}} f^T L f \overset{a.s.}{\to} c\int_\Omega \|\nabla f(x)\|^2 p^2(x)dx$$

Graph Laplacian Regularization

- SSL with Infinite Unlabeled Data and Fixed Labeled Data

$$\begin{aligned} \min \quad & \int_{\Omega} \|\nabla f(x)\|^2 p^2(x) dx \\ s.t. \quad & f(X_L) = Y_L \end{aligned}$$

  - "Smooth": small penalty/semi-norm, ***intuitively***
  - Use *sign(f)* to obtain the class in classification
  - Same analysis to regularized Least Squares

Graph Laplacian Regularization

- SSL with Infinite Unlabeled Data and Fixed Labeled Data

$$\begin{aligned} \min \quad & \int_{\Omega} \|\nabla f(x)\|^2 p^2(x) dx \\ s.t. \quad & f(X_L) = Y_L \end{aligned}$$

  - "Smooth": small penalty/semi-norm, ***intuitively***
  - Use *sign(f)* to obtain the class in classification
  - Same analysis to regularized Least Squares

  - Tip of the iceberg: $\min_{f \in ?}$

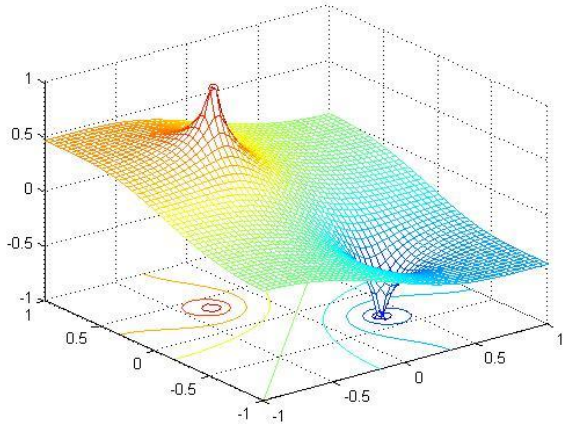
Problem of Graph Laplacian Regularization

- In SSL, the more unlabeled data, the better results we expect

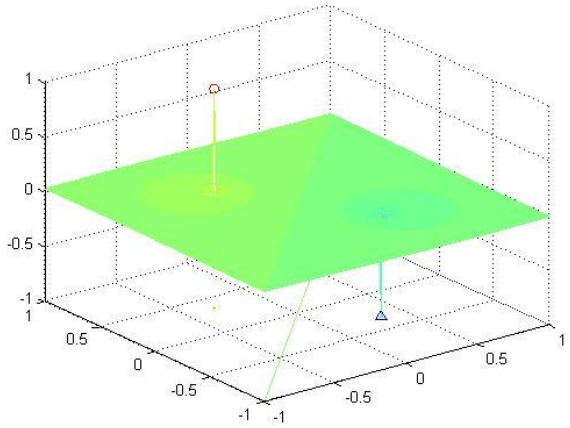
Problem of Graph Laplacian Regularization

- In SSL, the more unlabeled data, the better results we expect
- However...

Less unlabeled data



More unlabeled data

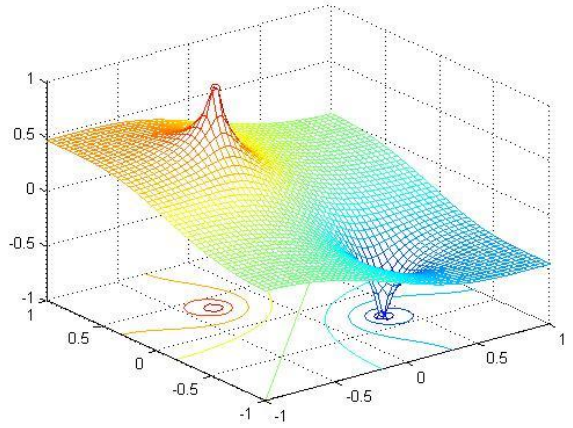


"Indicator" functions of labeled points.

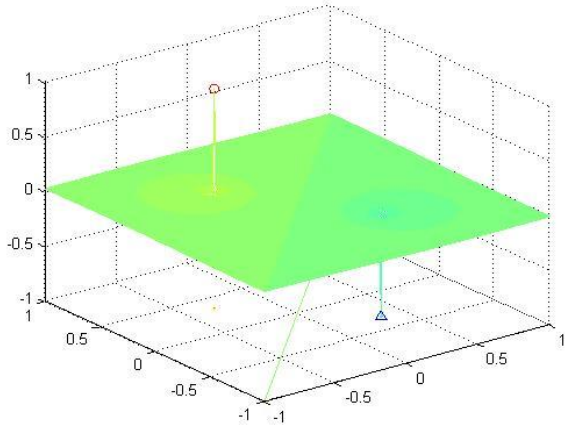
Problem of Graph Laplacian Regularization

- In SSL, the more unlabeled data, the better results we expect
- However...

Less unlabeled data



More unlabeled data



"Indicator" functions of labeled points.

The more unlabeled data we have, the less stable the classifier gets, and the worse the results become.
The opposite of the intuition "smooth" !

Problem of Graph Laplacian Regularization

- Reason: is the solution "smooth" ?

$$I(f) = \int_{\Omega} \|\nabla f(x)\|^2 dx = \int_{\Omega} \sum_{i=1}^{d} |\frac{\partial f(x)}{\partial x_i}|^2 dx \approx d(\frac{1}{h})^2 hh \cdots h = O(h^{d-2})$$

  - For $d > 2$ as $h \to 0$, $I(f) \to 0$, but $|\nabla f(x)| \to +\infty$, same for $d = 2, \ldots$; this means for the solutions (*indicator functions*), $I(f) = 0$ (*the penalty is small !*)
  - Indicator functions: nearby points have *quite different* function values, so not "smooth", contrary to the intuition!

Problem of Graph Laplacian Regularization

- Reason: is the solution "smooth" ?

$$I(f) = \int_{\Omega} \|\nabla f(x)\|^2 dx = \int_{\Omega} \sum_{i=1}^{d} |\frac{\partial f(x)}{\partial x_i}|^2 dx \approx d(\frac{1}{h})^2 hh \cdots h = O(h^{d-2})$$

  - For $d > 2$ as $h \to 0$, $I(f) \to 0$, but $|\nabla f(x)| \to +\infty$, same for $d = 2, \ldots$; this means for the solutions (*indicator functions*), $I(f) = 0$ (*the penalty is small !*)
  - Indicator functions: nearby points have *quite different* function values, so not "smooth", contrary to the intuition!
  - The semi-normed space is too rich when $d \geq 2$

$$\{f : f \in L_2, \int_{\Omega} \|\nabla f(x)\|^2 dx < +\infty\}$$

  - Note: in $d = 1$, no such problem.

Solution: Iterated Laplacian Regularization

- Problem: Semi-normed (solution) space is too large
  - Sobolev Space

$$H^m(\Omega) = \{f : D^\alpha f \in L_2(\Omega), \forall \alpha, s.t. |\alpha| \leq m\}, D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}$$

$$\{f : f \in L_2, \int_\Omega \|\nabla f(x)\|^2 dx < +\infty\} \leftrightarrow H^1(\Omega)$$

Solution: Iterated Laplacian Regularization

- Problem: Semi-normed (solution) space is too large
  - Sobolev Space

$$H^m(\Omega) = \{f : D^\alpha f \in L_2(\Omega), \forall \alpha, s.t. |\alpha| \leq m\}, D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}$$

$$\{f : f \in L_2, \int_\Omega \|\nabla f(x)\|^2 dx < +\infty\} \leftrightarrow H^1(\Omega)$$

- Solution: Shrink the solution space

$$H^j(\Omega) \subset H^k(\Omega), j > k$$

  - Need *a Sobolev norm of higher order* to shrink the solution space; the Sobolev embedding theorem tells more.

Solution: Iterated Laplacian Regularization

- Problem: Semi-normed (solution) space is too large
  - Sobolev Space

$$H^m(\Omega) = \{f : D^\alpha f \in L_2(\Omega), \forall \alpha, s.t. |\alpha| \leq m\}, D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}$$

$$\{f : f \in L_2, \int_\Omega \|\nabla f(x)\|^2 dx < +\infty\} \leftrightarrow H^1(\Omega)$$

- Solution: Shrink the solution space

$$H^j(\Omega) \subset H^k(\Omega), j > k$$

  - Need *a Sobolev norm of higher order* to shrink the solution space; the Sobolev embedding theorem tells more.
- Key: How to implement this idea in a practical algorithm

Solution: Iterated Laplacian Regularization

- How to implement Sobolev Norm
  - Original definition: $\|D^m f\|^2$, *not easy to implement* for random samples

Solution: Iterated Laplacian Regularization

- How to implement Sobolev Norm
  - Original definition: $\|D^m f\|^2$, *not easy to implement* for random samples
  - *Easy to implement* using iterated Laplacian semi-norm

$$I_s^d(f) = \int_\Omega f(x)\Delta^s f(x)dx = \sum_{i=1}^{\infty} |\hat{f}_i|^2 \lambda_i^s, s \geq 0$$

$$\Delta[\Delta^{s-1}\phi_i] = \lambda_i^s \phi_i$$

  - Equivalent definition using Fourier basis: eigenfunctions of a Laplacian form L_2 basis

$$f \in L_2, f = \sum_{i=1}^{\infty} \hat{f}_i \phi_i, \hat{f}_i = \langle f, \phi_i \rangle, \Delta\phi_i = \lambda_i \phi_i$$

  - Note, λ_i is increasing

Solution: Iterated Laplacian Regularization

- Iterated Laplacian Semi-normed Space

$$D_s(\Omega) = \{f \in L_2(\Omega) : I_s^d(f) < \infty\}$$

$$D_s(\Omega) \subset H^s(\Omega)$$

  - Alternative way of smoothness controlled by *s*, compared to C^k functions controlled by *k*.
  - From point-wise functions to L_2 functions.

Solution: Iterated Laplacian Regularization

- Least Squares by Higher Order Regularization Using the iterated graph Laplacian (least squares or interpolation)

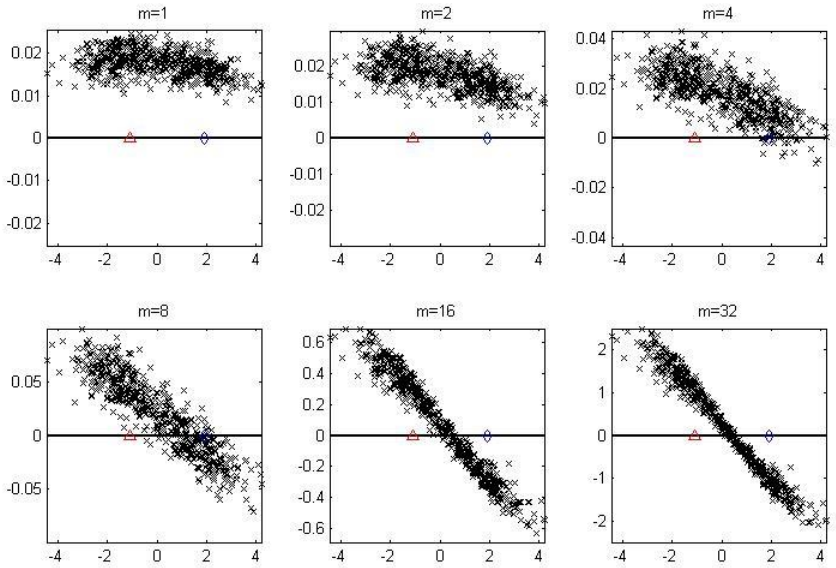
$$\min_{f \in \mathbb{R}^n} \sum_{x_i \in X_L} (f(x_i) - y_i)^2 + \mu f^T L^m f$$

- Limit of $I^d_{m,n}(f) = f^T L^m f$
  - For uniform density, $I^d_m(f) = \int_\Omega f(x) \Delta^m f(x) dx$
  - For nonuniform density, use weighted Laplacian
  - Need proper boundary conditions when domains/manifolds have boundaries

Experiments

- A Toy Example
  - Mixture of two unit variance Gaussians in R^{20} at $(\pm 1.5, 0, 0, \cdots, 0)$, one labeled point for each Gaussian

$$\min_{f \in \mathbb{R}^n} \sum_{x_i \in X_L} (f(x_i) - y_i)^2 + \mu f^T L^m f$$



$$x_1 \leq 0 \ (y = +1)$$
$$x_1 > 0 \ (y = -1)$$

Experiments

- More Examples: Regular vs Iterated Laplacian Regularization $\min_{f \in \mathbb{R}^n} \sum_{x_i \in X_L} (f(x_i) - y_i)^2 + \mu f^T L^m f$

Classification errors % with std for $m = 1$ and $m = 4$.

DATA SET	$m = 1$	$m = 4$
MNIST 3VS8	7.5 ± 1.5	5.6 ± 1.3
MNIST 4VS9	12.2± 2.9	8.0 ± 2.6
PCMAC	16.6± 2.4	11.5± 1.4
AUT-AVN	13.7± 2.6	10.1± 1.3
REAL-SIM	9.4 ± 3.3	5.8 ± 0.8
CCAT	24.0± 2.8	21.5± 2.8
GCAT	13.1± 1.5	12.0± 1.3
GENE-P	39.3± 9.1	29.0± 8.9
GENE-B	45.3± 7.3	41.9± 9.2

Noniterated Laplacian

Iterated Laplacian

Experiments

- More Examples:

Classification errors % for SSL Benchmark.

	QC+CMN	LapRLS	Best	IterLap
		$\|X_L\| = 10$		
g241c	39.96	43.95	22.76	18.01
g241d	46.55	45.68	18.64	20.99
Digit1	9.80	5.44	5.44	6.54
USPS	13.61	18.99	13.61	13.10
BCI	50.36	48.97	46.90	46.71
Text	40.79	33.68	27.15	38.84
		$\|X_L\| = 100$		
g241c	22.05	24.36	13.49	14.82
g241d	28.20	26.46	4.95	10.55
Digit1	3.15	2.92	2.44	2.22
USPS	6.36	4.68	4.68	3.96
BCI	46.22	31.36	31.36	43.78
Text	25.71	23.57	23.09	25.77

Comparison to the results from "O. Chapelle, B. Schölkopf, and A. Zien, editors. Semi-Supervised Learning. MIT Press, Cambridge, MA, 2006."

$$\min_{f \in \mathbb{R}^n} \sum_{x_i \in X_L} (f(x_i) - y_i)^2 + \mu f^T L^m f + \epsilon \|f\|_{L_2}^2$$

QC+CMN: Chapter 11, *m=1*.
LapRLS : Chapter 12.
Best :The best from 13 different algorithms.
IterLap : iterated Laplacian, *m>1*

$$\|f\|_{L_2}^2 \leftrightarrow L_2(\Omega) = H^0(\Omega)$$

Summary

- Use regularizer $f^T L^m f$ instead of $f^T L f$ in most applications
  - Sound theory support + improvement in practice

- Intuition is *IMPORTANT*, but we need *MORE*: careful mathematical analysis
  - $\int_\Omega \|\nabla f(x)\|^2 dx$ is not "good" to describe smoothness in $d \geq 2$

Summary

- Use regularizer $f^T L^m f$ instead of $f^T L f$ in most applications
  - Sound theory support + improvement in practice

- Intuition is *IMPORTANT*, but we need *MORE*: careful mathematical analysis
  - $\int_\Omega \|\nabla f(x)\|^2 dx$ is not "good" to describe smoothness in $d \geq 2$

Thanks to Todd Dupont, Nathan Srebro, Boaz Nadler for helpful discussions.

Summary

- Use regularizer $f^T L^m f$ instead of $f^T L f$ in most applications
  - Sound theory support + improvement in practice

- Intuition is *IMPORTANT*, but we need *MORE*: careful mathematical analysis
  - $\int_\Omega \|\nabla f(x)\|^2 dx$ is not "good" to describe smoothness in $d \geq 2$

Thank you!

Thanks to Todd Dupont, Nathan Srebro, Boaz Nadler for helpful discussions.