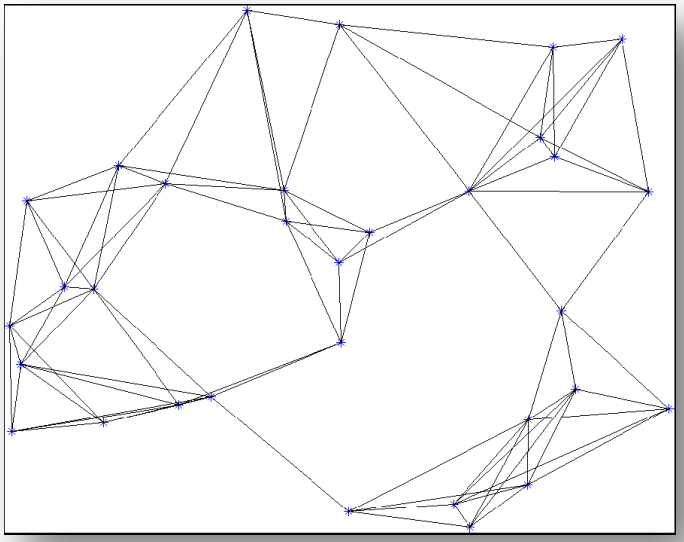


Kernels and Regularization on Graphs

Alexander J. Smola and Risi Kondor

Presenter: Zheng Zhao

Outline

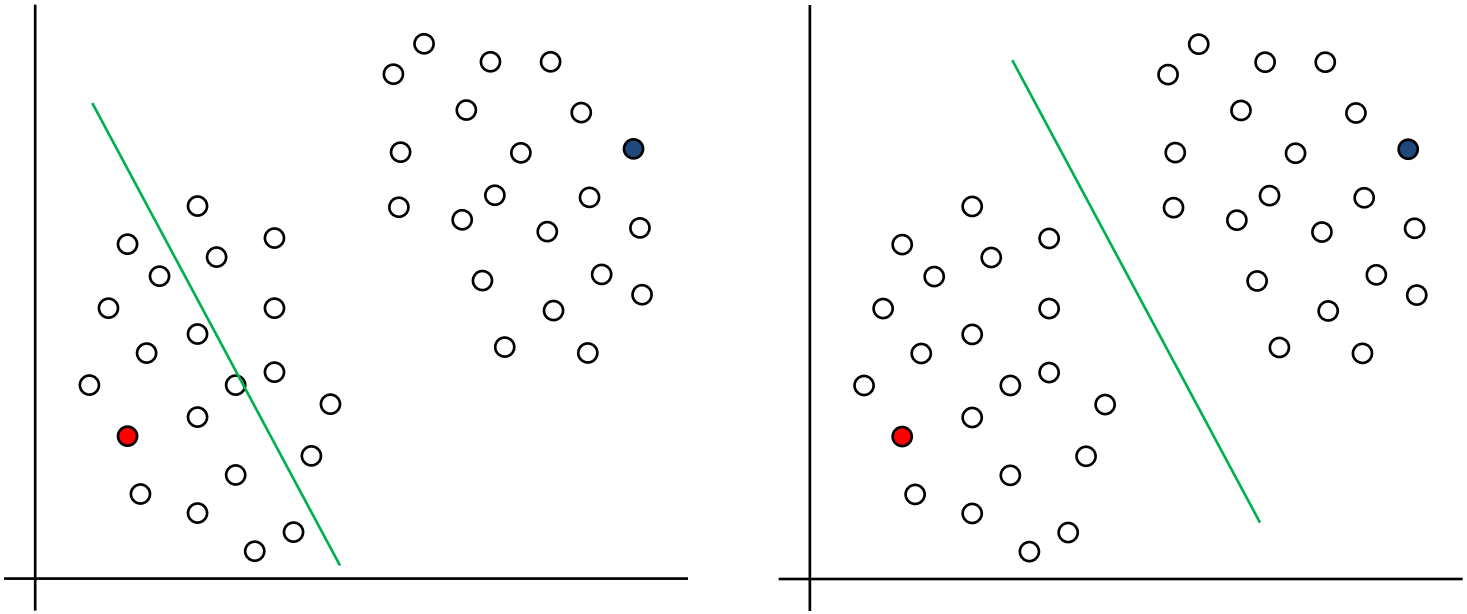
- Motivation of Regularization
- Regularization via Graph Laplacian
  - Examples
  - Graph Laplacian
  - Spectrum Design
- Kernel Connection
  - kernels in RKHS
- Misc.

Regularization = Robustness [1]

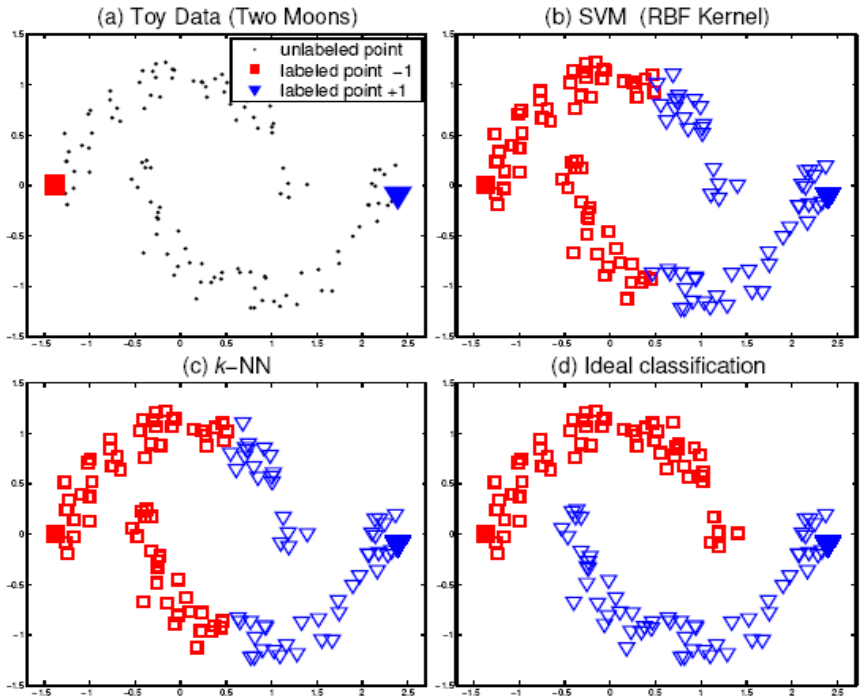
- Apply bias to shrink hypothesis space and penalizing dispreferable learning models. (*in addition to the loss function*)
  - Prefer simpler (sparse) learning model.
  - Prefer learning model having estimation with small variance.
  - Prefer learning model being consistent with the structure of the data.
- Examples:
  - Ridge Regression and SVM
  - Regularized LDA
  - The class of graph based semi-supervised learning models.

Regularization for Consistency

- Enforcing learning model to be consist with the structure of the data.
- An Example:



Another Example (Two Moons Data) [2]



$$\mathcal{Q}(F) = \frac{1}{2}\left(\sum_{i,j=1}^{n} W_{ij} \left\| \frac{1}{\sqrt{D_{ii}}} F_i - \frac{1}{\sqrt{D_{jj}}} F_j \right\|^2 + \mu \sum_{i=1}^{n} \left\| F_i - Y_i \right\|^2 \right)$$

Regularization via Graph Laplacian

- Basic idea:
  - Cluster assumption: Instances near to each other should share similar label.
  - Formulate pair wise similarity relationship among instances using graph.
  - Penalize functions that change abruptly on the graph.
- Tools:
  - The family of functions based on graph laplacian.

Graph Laplacian

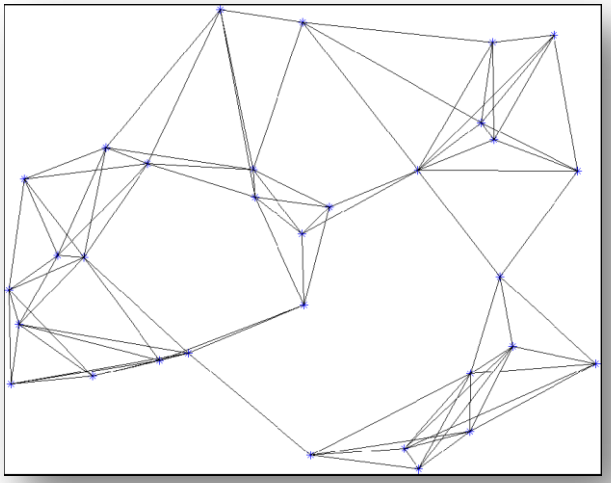
- W: similarity matrix. D: degree matrix (diagonal matrix- $D_{ii} = \sum_{j=1}^{n} w_{ij}$)
- Graph Laplacian:
  - Graph laplacian: L = D-W
  - Normalized graph laplacian: $\tilde{L} = D^{-\frac{1}{2}}(D-W)D^{-\frac{1}{2}}$
- Some properties
  - L is symmetric and positive semi-definite
  - The smallest Eigen pair of L is (0, **1**).
  - $f^T Lf = -\frac{1}{2}\sum_{i\sim j} w_{ij}(f_i - f_i)^2$

Using Graph Laplacian for Regularization

- Denote f the estimation function, we measure the consistency of the function by:

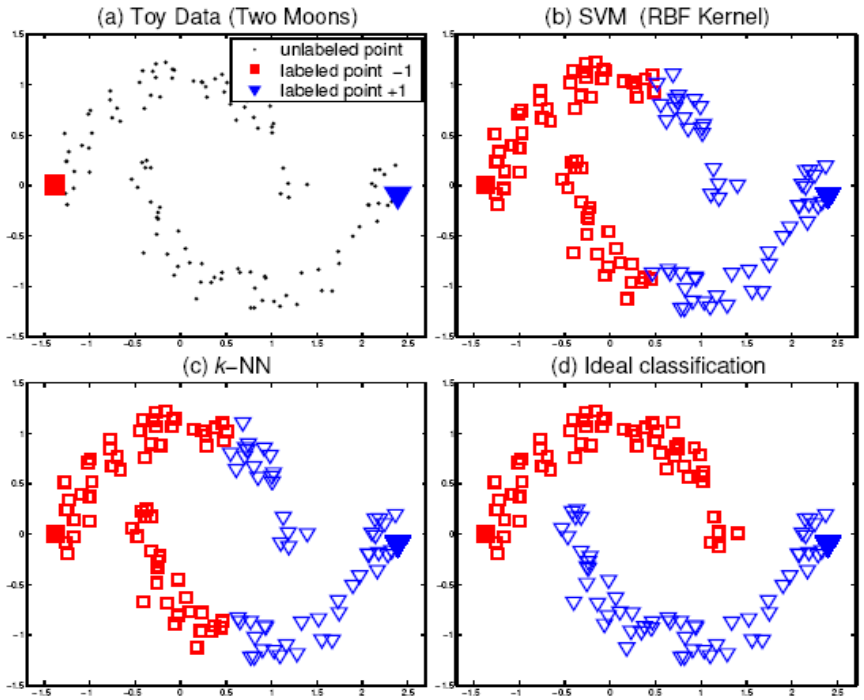
$$f^T L f = -\frac{1}{2}\sum_{i\sim j} w_{ij}(f_i - f_j)^2$$

$$f^T \tilde{L} f = -\frac{1}{2}\sum_{i\sim j} w_{ij}\left(\frac{f_i}{D_{ii}} - \frac{f_j}{D_{jj}}\right)^2$$



$$\mathcal{Q}(F) = \frac{1}{2}\left(\sum_{i,j=1}^{n} W_{ij}\left\|\frac{1}{\sqrt{D_{ii}}}F_i - \frac{1}{\sqrt{D_{jj}}}F_j\right\|^2 + \mu\sum_{i=1}^{n}\|F_i - Y_i\|^2\right)$$

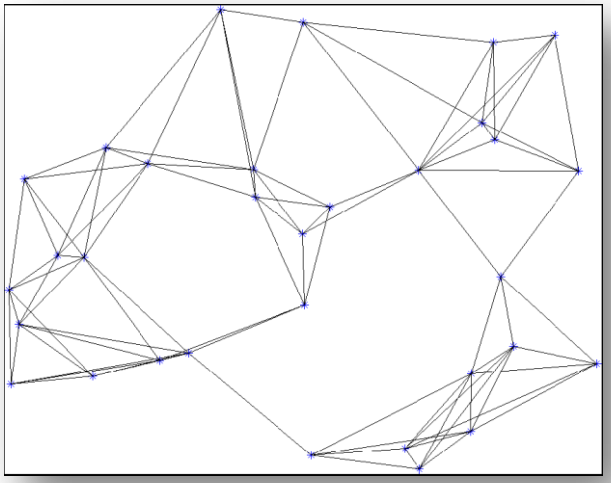
Another Example (Two Moons Data) [2]

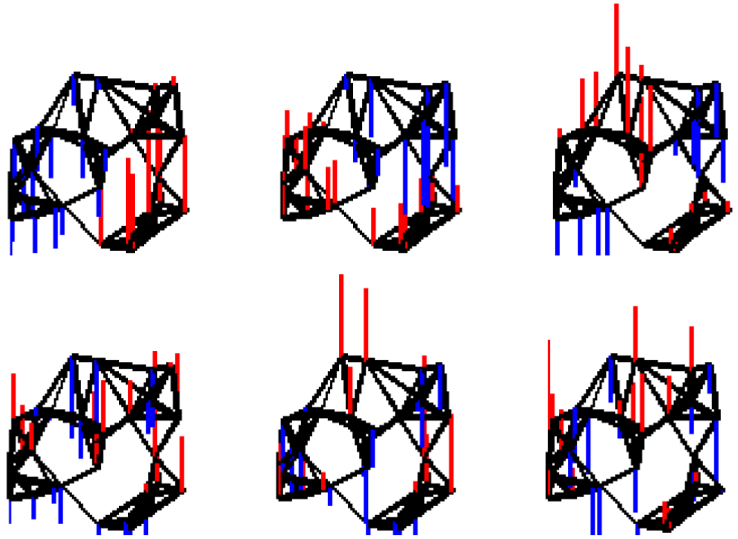


$$\mathcal{Q}(F) = \frac{1}{2}\left(\sum_{i,j=1}^{n} W_{ij} \left\| \frac{1}{\sqrt{D_{ii}}} F_i - \frac{1}{\sqrt{D_{jj}}} F_j \right\|^2 + \mu \sum_{i=1}^{n} \left\| F_i - Y_i \right\|^2 \right)$$

Spectrum of Graph Laplacian

- The spectrums of graph laplacian is closely related to spectral clustering.
  - The Eigen vector corresponding to the second smallest Eigen value is the optimal continuous cluster indicator.





Spectrum Design

- Let $L = \sum \lambda_i u_i u_i^T$ be the spectral decomposition of L

$$f^T L f = f^T \sum \lambda_i u_i u_i^T f = \sum \lambda_i f^T u_i u_i^T f = \sum \lambda_i (f^T u_i)^2$$

- We penalize f harder, if f closely correlated to eigenvectors which has large eigenvalue.
- Moreover, we can penalize even harder by spectram designing.

$$\sum r(\lambda_i)(f^T u_i)^2$$

- We require $r()$ to be an increasing function.

Functions of Particular Interest

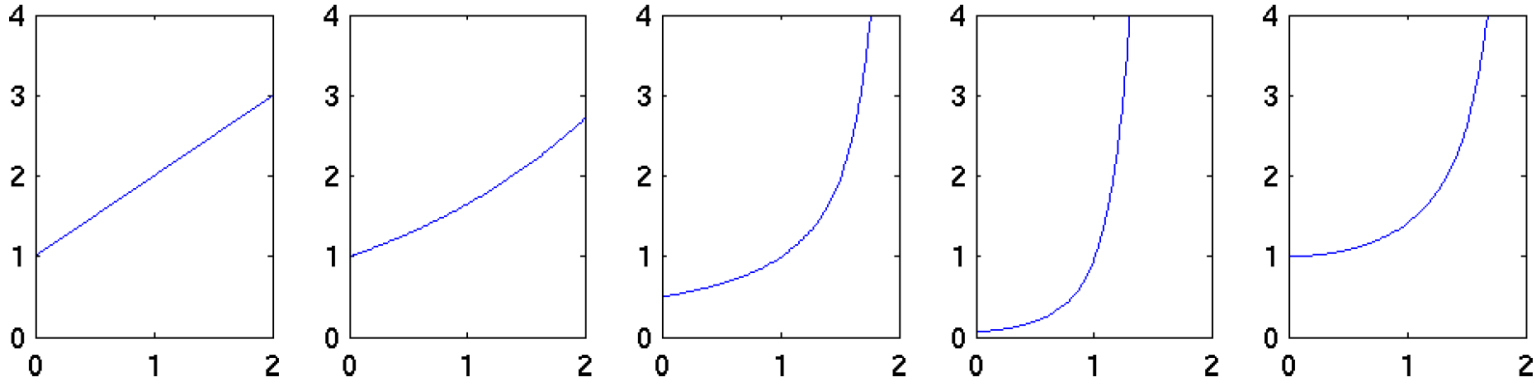
$$r(\lambda) = 1 + \sigma^2\lambda \qquad \text{(Regularized Laplacian)}$$

$$r(\lambda) = \exp\left(\sigma^2/2\lambda\right) \qquad \text{(Diffusion Process)}$$

$$r(\lambda) = (aI - \lambda)^{-1} \text{ with } a \geq 2 \qquad \text{(One-Step Random Walk)}$$

$$r(\lambda) = (aI - \lambda)^{-p} \text{ with } a \geq 2 \qquad (p\text{-Step Random Walk})$$

$$r(\lambda) = (\cos \lambda\pi/4)^{-1} \qquad \text{(Inverse Cosine)}$$



Kernel Connection (1)

- Regularization in feature space [3]:

$$c\left(\left(x_1, y_1, f(x_1)\right), \ldots, \left(x_m, y_m, f(x_m)\right)\right) + \Omega\left(\|f\|_{\mathcal{H}}\right)$$

- We want to find the reproducing Hilbert space *H*, such that:

$$\|f\|_H^2 = \langle f, f \rangle_H = f^T L f$$

- Using reproducing property we have:

$$\boldsymbol{f}^\top = \boldsymbol{f}^\top P K$$

Kernel Connection (2)

- Therefore for L the corresponding kernel is:

$$K = L^{-1}$$

- For $P = \sum r(\lambda_i)(f^T u_i)^2$ the corresponding kernel is:

$$K = \sum_{i=1}^{m} r^{-1}(\lambda_i)\, \boldsymbol{v}_i \boldsymbol{v}_i^\top$$

$$K = (I + \sigma^2 \tilde{L})^{-1} \quad \text{(Regularized Laplacian)}$$

$$K = \exp(-\sigma^2/2\tilde{L}) \quad \text{(Diffusion Process)}$$

$$K = (aI - \tilde{L})^p \text{ with } a \geq 2 \quad (p\text{-Step Random Walk})$$

$$K = \cos \tilde{L}\pi/4 \quad \text{(Inverse Cosine)}$$

- Since r() is an increasing function, the inverse of r() is an decreasing function.

Misc.

- The graph laplacian of Regular Graph
- Permutation Invariant Linear Function on Graphs
- Laplace Operator(Δ) and Fourier Transformation.
- Approximate Computation
- Products of Graph
- Link Analysis

Reference

- [1] Machine Learning (Theory): http://hunch.net/?p=197
- [2] Learning with Local and Global Consistency, NIPS04
- [3] Learning With Kernels, B. Scholkopf, A. J. Smola

Thank you and any questions