Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions

Xiaojin Zhu, Zoubin Ghahramani and John Lafferty

M. Perelló Nieto

Course:
Random Graphs and Complex Networks

Aalto, Nov 2013

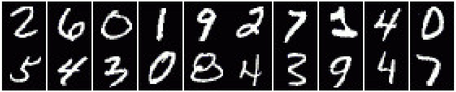
Aalto University
School of Science

Index

Aalto University
School of Science

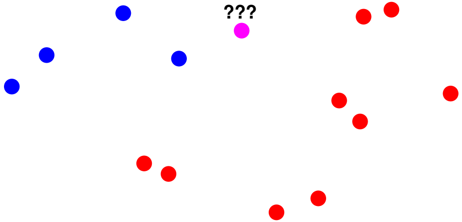
Index

Motivation



- Machine learning needs labeled data to train the models
- Labeled data is ussually *difficult* to collect
- Some times and *expert* has to annotate the labels
- We are surrounded of a *huge amount of unlabeled data*

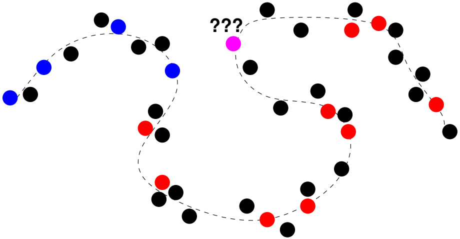
Semi-Supervised Learning



(KyungHyun Cho)

- Classification of a point *without* unlabeled data

Semi-Supervised Learning



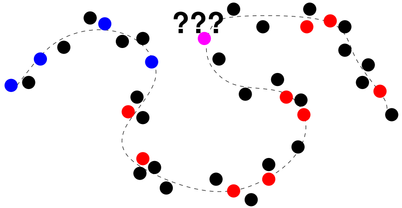
(KyungHyun Cho)

- Classification of the same point *with* unlabeled data

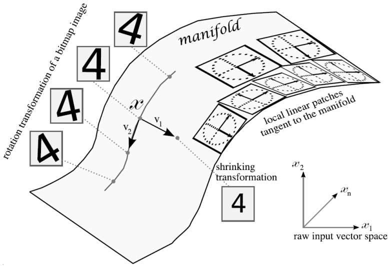
Index

Aalto University
School of Science

Data Manifold



(KyungHyun Cho)



(Bengio, 2009)

- Data live in a manifold

Index

Harmonic Functions

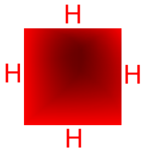
- Twice continuous differentiable function
- Satisfies Laplace's equations:

$$\frac{\partial^2 f}{\partial x_1^2} + \frac{\partial^2 f}{\partial x_2^2} + \cdots + \frac{\partial^2 f}{\partial x_n^2} = 0$$

$$\Delta f = 0$$

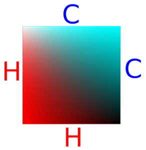
Aalto University
School of Science

Example of Harmonic Function

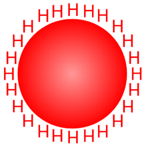


$$\mathtt{Temp}_c = \frac{1}{N}\sum_{i=1}^{N} w_{ci}\mathtt{Temp}_i$$

$$\mathtt{Temp}_c = \frac{H+H+H+H}{4}$$



$$\mathtt{Temp}_c = \frac{2H+2C}{4}$$



$$\mathtt{Temp}_c = \frac{1}{2\pi}\int_0^{2\pi} d\theta \mathtt{Temp}(\cos\theta, \sin\theta)$$

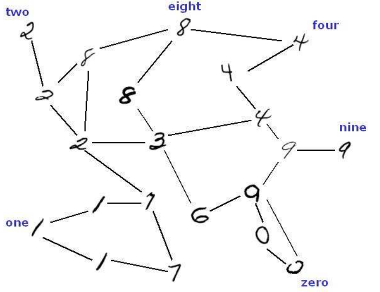
Aalto University
School of Science

Index

Aalto University
School of Science

Definitions

- *points* $(x_{1,1}, x_{1,2}, \ldots, x_{1,d}, y_1), \ldots, (x_{n,1}, x_{n,2}, \ldots, x_{n,d}, y_n)$
- l labeled points
- u unlabeled points
- n = l + u
- G = (V,E)
- V = {L, U}
- L = {1,. . . ,l}
- U = {l+1,. . . ,l+u}



Weights as a distance measure

$$w_{ij} = \exp\left(-\sum_{d=1}^{m} \frac{(x_{id} - x_{jd})^2}{\sigma_d^2}\right)$$

Definitions

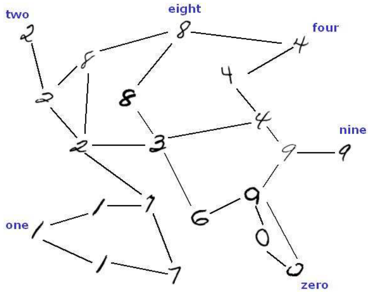
Energy of the system

$$E(f) = \frac{1}{2} \sum_{i,j} w_{ij}(f(i) - f(j))^2$$

where $f(i) = f_l(i) = y_i$

Gaussian field

$$p_\beta(f) = \frac{e^{-\beta E(f)}}{Z_\beta}$$



f in *unlabeled points* is the average of the neighbors

$$f(j) = \frac{1}{d_j} \sum_{i,j} w_{ij} f(i), \text{ for } j = l+1, \ldots, l+u$$

Harmonic solution

Minimum energy function is *harmonic*

$f = \arg\min_{f|L=f_l} E(f)$

it satisfies

$\Delta f = 0$

on unlabeled data points U
Where Δ is the *combinatorial Laplacian*

$\Delta = D - W$

Harmonic solution

$$\Delta = D - W$$

$$\begin{bmatrix} \Delta_{ll} & \Delta_{lu} \\ \Delta_{ul} & \Delta_{uu} \end{bmatrix} = \begin{bmatrix} D_{ll} & D_{lu} \\ D_{ul} & D_{uu} \end{bmatrix} - \begin{bmatrix} W_{ll} & W_{lu} \\ W_{ul} & W_{uu} \end{bmatrix}$$

And given the known and unknown labels

$$\begin{bmatrix} f_l \\ f_u \end{bmatrix}$$

Then the solution to $\Delta f = 0$ s.t. $f|L = f_l$ is given by

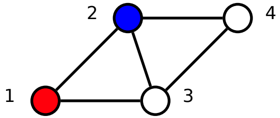
$$f_u = (D_{uu} - W_{uu})^{-1} W_{ul} f_l$$

Index

Aalto University
School of Science

Simple example

$$\begin{bmatrix} \Delta_{ll} & \Delta_{lu} \\ \Delta_{ul} & \Delta_{uu} \end{bmatrix} = \begin{bmatrix} D_{ll} & D_{lu} \\ D_{ul} & D_{uu} \end{bmatrix} - \begin{bmatrix} W_{ll} & W_{lu} \\ W_{ul} & W_{uu} \end{bmatrix}$$



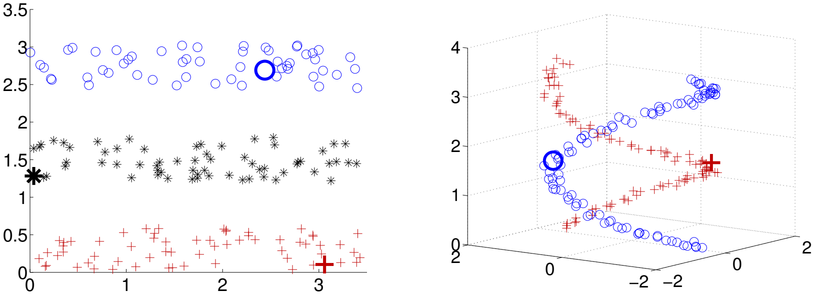
$$\begin{bmatrix} 2 & -1 & -1 & 0 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ 0 & -1 & -1 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix} - \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

$$f_u = (D_{uu} - W_{uu})^{-1} W_{ul} f_l$$

$$\begin{bmatrix} f_3 \\ f_4 \end{bmatrix} = \left(\begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix} - \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 1.6 \\ 1.8 \end{bmatrix}$$

Aalto University
School of Science

Harmonic energy minimization

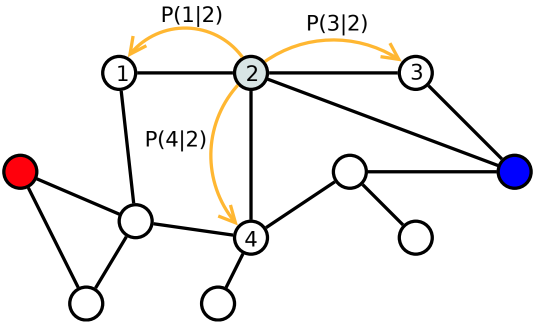


- Figure 1: $l = 3$, $u = 178$ and $\sigma = 0.22$
- Figure 2: $l = 2$, $u = 184$ and $\sigma = 0.43$
- *Large symbols* indicate *labeled* data, other points are unlabeled

Index

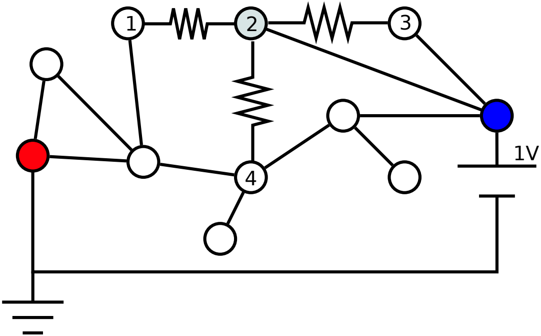
Aalto University
School of Science

Random Walks



- $f(i)$: probability that particle starting in node i hits node with label 1
- We fix the value of f in labeled points
- The walk depends on the time parameter t

Electric Networks



- *G* are resistors with conductance *W*
- Connect nodes labeled 1 to 1V and 0 to Ground
- f_u is the *voltage* in the resulting electric network
- Minimizes the energy dissipation
- Follows from *Kirchoff's and Ohm's* laws

Other interpretations

- Graph Kernels:
  - Using kernel $\hat{f}_t(j) = \sum_{i \in L} \alpha_i y_i K_t(i,j)$
  - Solution to heat equations with initial heat sources $\alpha_i y_i$
- Spectral clustering:
  - Objective function is minimization of the Raleigh quotient
  - $R(f) = \frac{f^T \Delta f}{f^T D f} = \frac{\sum_{ij} w_{ij}(f(i) - f(j))^2}{\sum_i d_i f(i)^2}$
  - Second smallest eigenvector $\Delta f = \lambda D f$
- Graph Mincuts:
  - Source (-1) and Sink (+1) nodes are labeled points

Aalto University
School of Science

Index

Aalto University
School of Science

Class Mass Normalization (CMN)

- Without prior knowledge assign *1* if $f(i) \geq \frac{1}{2}$, ow 0
  call this rule *"harmonic threshold"*
- If classes not separated produces *unbalanced classifications*
- If we know that *proportions* of class 1 and 0 are q and $(1 - q)$

$$\text{Mass of class 1} = \sum_i f_u(i)$$

$$\text{Mass of class 0} = \sum_i (1 - f_u(i))$$

- Then assign class 1 iff

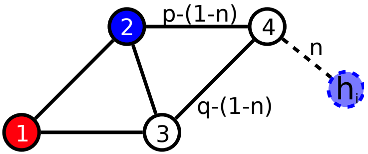
$$q \frac{f_u(i)}{\sum_i f_u(i)} \geq (1 - q) \frac{1 - f_u(i)}{\sum_i (1 - f_u(i))}$$

Aalto University
School of Science

Index

Aalto University
School of Science

External Classifier



- Use *external classifier* to produce h_u labels
- Attach a *"dongle" node* with this label
- *Assign a transition probability* n and substract to the rest
- Can be seen as *"Assignment costs"* to the energy function

If we doubt on the original labels:

- Attach "dongles" to labeled nodes and remove labels

Aalto University
School of Science

Index

Aalto University
School of Science

Learning weight functions

- Learn the σ_d's from labeled and unlabeled data

$$H(f) = \frac{1}{u} \sum_{i=l+1}^{n} H_i(f(i))$$

$$H_i(f(i)) = -f(i) \log f(i) - (1 - f(i)) \log(1 - f(i))$$

- *Small entropy* implies more “confidence” labeling
- H has a minimum at 0 as $\sigma_d \to 0$
- Replace P with a smoothed matrix $\tilde{P}$

$$\tilde{P} = \epsilon U + (1 - \epsilon) P$$

$$U_{ij} = \frac{1}{l + u}$$

Aalto University
School of Science

Gradient Descent

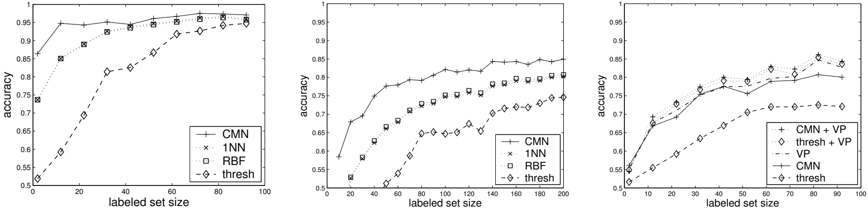
$$\frac{\partial H}{\partial \sigma_d} = \frac{1}{u} \sum_{i=l+1}^{l+u} \log \left(\frac{1 - f(i)}{f(i)} \right) \frac{\partial f(i)}{\partial \sigma_d}$$

$$\frac{\partial f_u}{\partial \sigma_d} = (I - \tilde{P}_{uu})^{-1} \left(\frac{\partial \tilde{P}_{uu}}{\partial \sigma_d} f_u + \frac{\partial \tilde{P}_{ul}}{\partial \sigma_d} f_l \right)$$

$$\frac{\partial p_{ij}}{\partial \sigma_d} = \frac{\frac{\partial w_{ij}}{\partial \sigma_d} - p_{ij} \sum_{n=1}^{l+u} \frac{\partial w_{in}}{\partial \sigma_d}}{\sum_{n=1}^{l+u} w_{in}}$$

$$\frac{\partial w_{ij}}{\partial \sigma_d} = \frac{2 w_{ij} (x_{di} - x_{dj})^2}{\sigma_d^3}$$

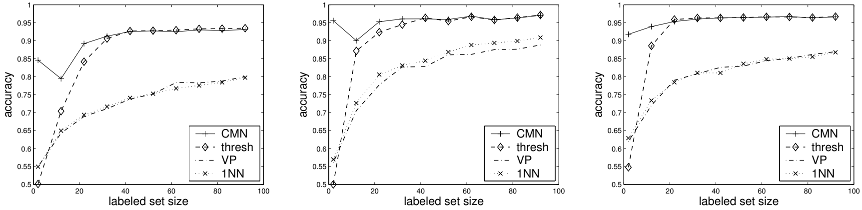
Handwritten digits dataset



(a) digits “1” vs. “2” *(b)* all 10 digits *(c)* odd vs. even

- $\sigma_d = 380$
- 10 random trials
- Unbalanced class sizes 455, 213, 129, . . . , 353
- *CMN improves* incorporating class priors

Document categorization

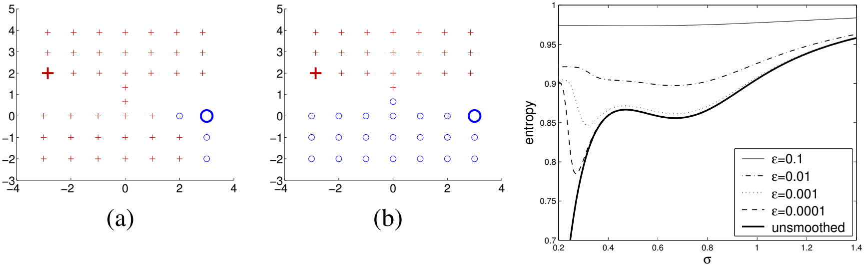


(a) PC vs. MAC *(b)* baseball vs. hockey *(c)* MS-Windows vs. MAC

- Processed into a “tf.idf” vector
- u, v are connected by edge if they are in 10 nearest neighbors

$$w_{uv} = \exp\left(-\frac{1}{0.03}\left(1 - \frac{u^T v}{|u||v|}\right)\right)$$

Learning the weight matrix



(a) (b) (c)

- *(a)* Unsmoothed, $H \to 0$ as $\sigma \to 0$
- *(b)* Optimal $\sigma_y = 0.67$, $\sigma_x \to \infty$, smoothed with $\epsilon = 0.01$
- *(c)* Smoothing helps to remove the entropy minimum

Aalto University
School of Science

Conclusion

1. Promising experimental results
2. Using random field gives *coherent probabilistic semantics*
3. Probabilistic framework suggests ways of incorporating *class priors* and learning *hyperparameters*

Bibliography I

Zhu, Xiaojin, Zoubin Ghahramani, and John Lafferty. "Semi-supervised learning using gaussian fields and harmonic functions." ICML. Vol. 3. 2003.

Aalto University
School of Science

Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions

Xiaojin Zhu, Zoubin Ghahramani and John Lafferty

M. Perelló Nieto

Course:
Random Graphs and Complex Networks

Aalto, Nov 2013

Aalto University
School of Science