Graph-based Semi-supervised Learning

Zoubin Ghahramani

Department of Engineering
University of Cambridge, UK

zoubin@eng.cam.ac.uk
http://learning.eng.cam.ac.uk/zoubin/

MLSS 2012
La Palma

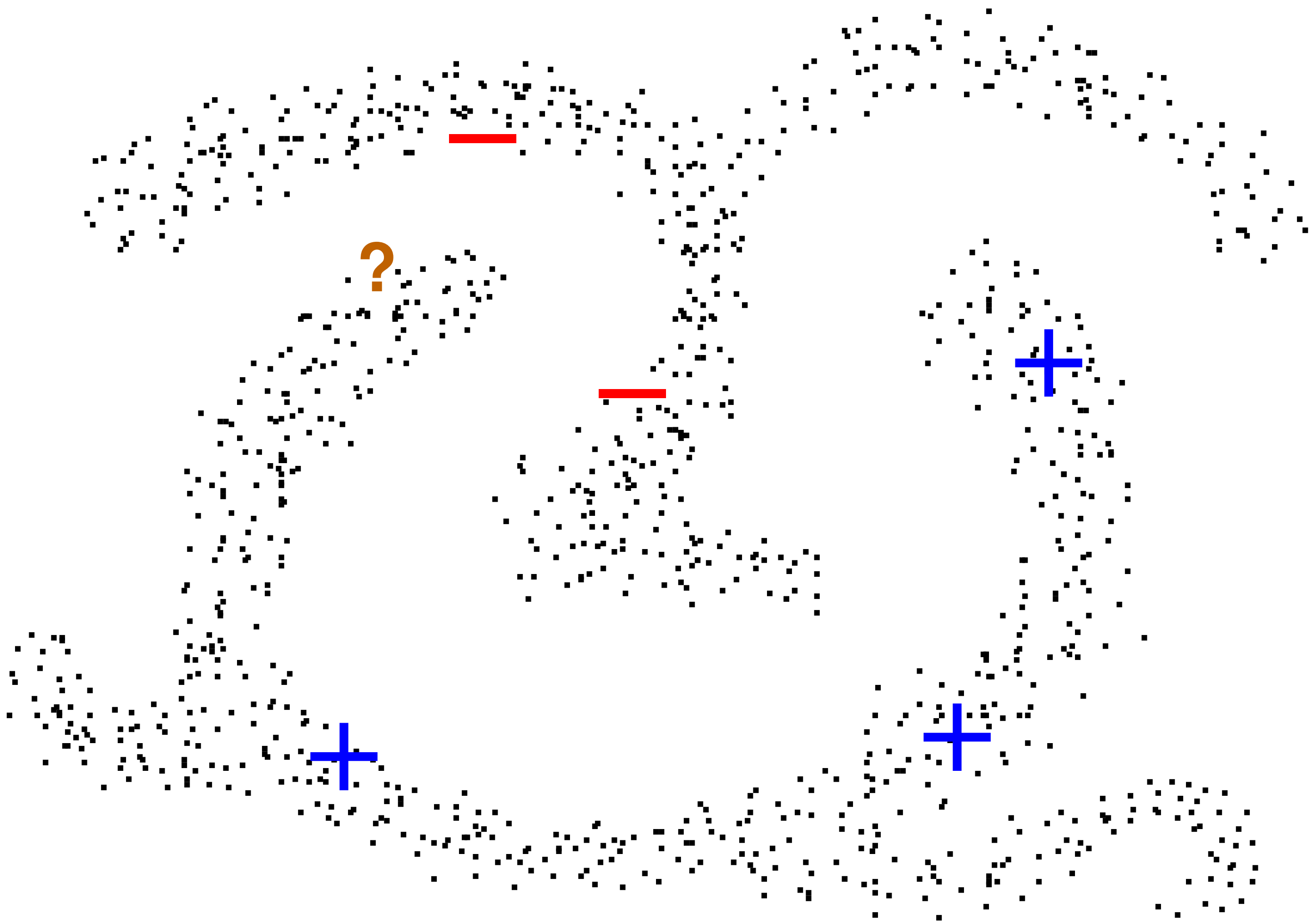
Motivation

- Large amounts of unlabelled data, small amounts of labelled data
- Labelling/annotating data is expensive
- We want supervised learning methods that can use information in the input distribution

Example: Images

Classification using Unlabelled Data

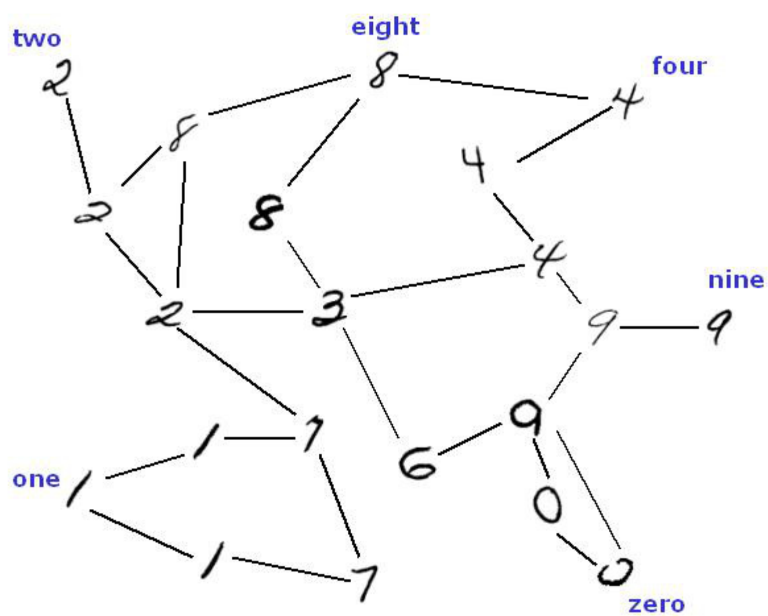
Assumption: there is information in the data distribution



Outline

- Graph-based semi-supervised learning
- Active graph-based semi-supervised learning
- Some thoughts on Bayesian semi-supervised learning

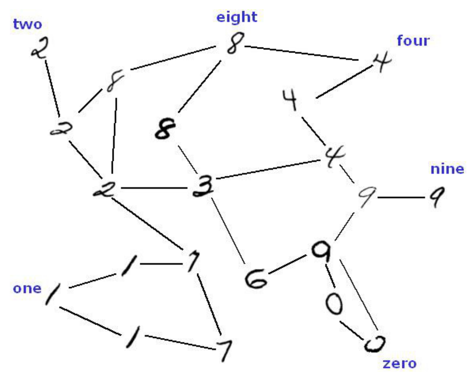
Graph-based Semi-supervised Learning
Labeled and Unlabeled Data as a Graph



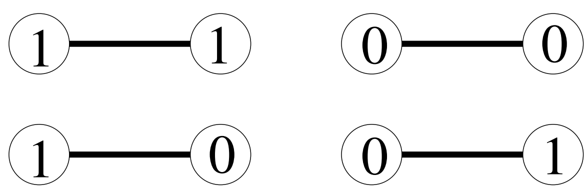
- Idea: Construct a graph connecting similar data points
- Let the hidden/observed labels be random variables on the nodes of this graph (i.e. the graph is an MRF)
- Intuition: Similar data points have similar labels
- Information "propagates" from labeled data points
- Graph encodes intuition

Work with Xiaojin Zhu (U Wisconsin) and John Lafferty (CMU)

The Graph



- nodes: instances in $L \cup U$. Binary labels $\mathbf{y} \in \{0,1\}^n$
- edges: local similarity. $n \times n$ symmetric weight matrix W assumed given.
- energy: $E(\mathbf{y}) = \frac{1}{2} \sum_{i,j} w_{ij} \left(y_i - y_j\right)^2$



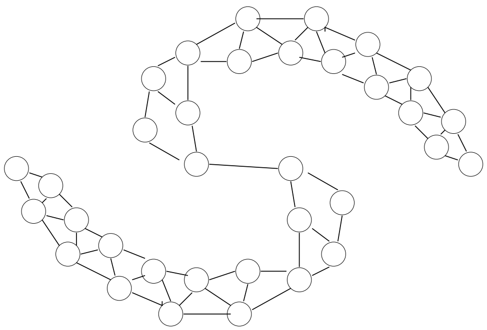
happy, low energy

unhappy, high energy

Low energy $\rightarrow$ Label Propagation

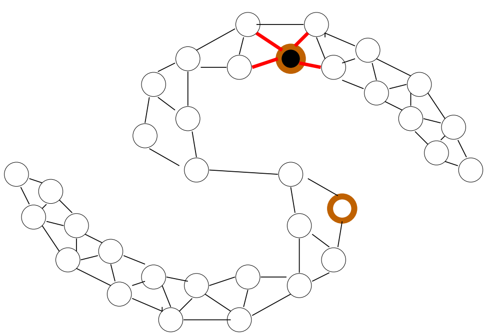
energy: $E(\mathbf{y}) = \frac{1}{2} \sum_{i,j} w_{ij} \left(y_i - y_j\right)^2$

With no labelled data, then $\mathbf{y} = \mathbf{1}$ or $\mathbf{y} = \mathbf{0}$ is a min energy configuration:

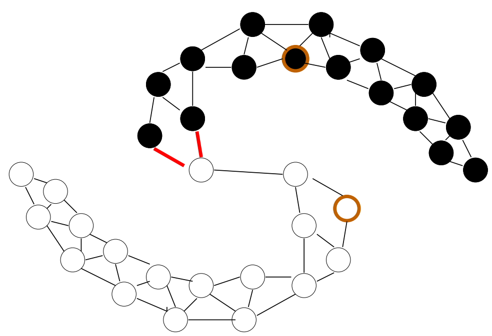


energy=0

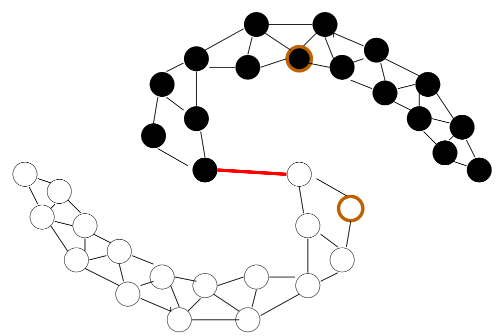
Conditioned on labeled data:



energy=4

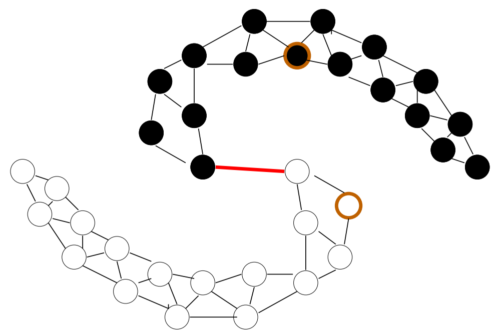


energy=2



energy=1

Discrete Markov Random Fields



$$E(\mathbf{y}) = \frac{1}{2} \sum_{i,j} w_{ij} \left(y_i - y_j\right)^2$$

$$p(\mathbf{y}) \propto \exp(-E(\mathbf{y})) \mid_{\mathbf{y}_L = L}$$

$$y_i \in \{0, 1\}$$

Graph mincut can find the min energy (MAP) configuration.

Problems: computing the probabilities is expensive, multi-class case is also harder to compute, and learning W is very hard.

[Zhu & Ghahramani 02] see also [Blum and Chawla 01]

We relaxed this to a
Gaussian random fields

Discrete Markov Random Fields, revisited

$$p(\mathbf{y}) \propto \exp(-E(\mathbf{y})) \mid_{\mathbf{y}_L = L}$$

$$y_i \in \{0, 1\}$$

Gaussian Random Fields

$$p(\mathbf{y}) \propto \exp(-E(\mathbf{y})) \mid_{\mathbf{y}_L = L}$$

$$y_i \in \mathbb{R}$$

Gaussian Random Fields

$$
\begin{aligned}
p(\mathbf{y}) &\propto \exp(-E(\mathbf{y}))\ |_{\mathbf{y}_L = L} \\
&= \exp\left(-\frac{1}{2}\sum_{i,j} w_{ij}\,(y_i - y_j)^2\right)\ |_{\mathbf{y}_L = L} \\
&= \exp\left(-\mathbf{y}^\top \Delta \mathbf{y}\right)\ |_{\mathbf{y}_L = L}
\end{aligned}
$$

$$
W = \begin{bmatrix} w_{11} & \dots & w_{1n} \\ & \dots & \\ w_{n1} & \dots & w_{nn} \end{bmatrix} \quad D = \begin{bmatrix} \sum w_{1\cdot} & & \mathbf{0} \\ & \dots & \\ \mathbf{0} & & \sum w_{n\cdot} \end{bmatrix}
$$

The Laplacian $\Delta = D - W$

$$
\Delta = \left[\begin{array}{c|c} \Delta_{LL} & \Delta_{LU} \\ \hline \Delta_{UL} & \Delta_{UU} \end{array}\right]
$$

The Laplacian

$$W = \begin{bmatrix} w_{11} & \dots & w_{1n} \\ & \dots & \\ w_{n1} & \dots & w_{nn} \end{bmatrix} \quad D = \begin{bmatrix} \sum w_{1\cdot} & & \mathbf{0} \\ & \dots & \\ \mathbf{0} & & \sum w_{n\cdot} \end{bmatrix}$$

This is the combinatorial or graph Laplacian $\Delta = D - W$

$$\Delta = \left[\begin{array}{c|c} \Delta_{LL} & \Delta_{LU} \\ \hline \Delta_{UL} & \Delta_{UU} \end{array}\right]$$

The graph Laplacian plays the same role on graphs as the Laplace operator in other spaces.

For example, in a Cartesian coordinate system, the Laplacian is given by sum of second partial derivatives of the function

$$\Delta f = \nabla \cdot \nabla f = \sum_i \frac{\partial^2 f}{\partial x_i^2}$$

Gaussian Random Fields

$$
\begin{aligned}
p(\mathbf{y}) &\propto \exp(-E(\mathbf{y})) \,|_{\mathbf{y}_L=L} \\
&= \exp\left(-\frac{1}{2}\sum_{i,j} w_{ij}\,(y_i - y_j)^2\right) \,|_{\mathbf{y}_L=L} \\
&= \exp\left(-\mathbf{y}^\top \Delta \mathbf{y}\right) \,|_{\mathbf{y}_L=L}
\end{aligned}
$$

The distribution of $\mathbf{y}_U$ given $\mathbf{y}_L$ is Gaussian: $\mathbf{y}_U \sim \mathcal{N}\left(f_U, \frac{1}{2}(\Delta_{UU})^{-1}\right)$

The mean is $f_U = -(\Delta_{UU})^{-1}\Delta_{UL}\mathbf{y}_L$

The Mean f_U

$$
\begin{aligned}
\text{The mean } f_U &\equiv \text{mode of Gaussian Random Field} \\
&\equiv \text{min energy state}
\end{aligned}
$$

- "soft labels", unique
- harmonic

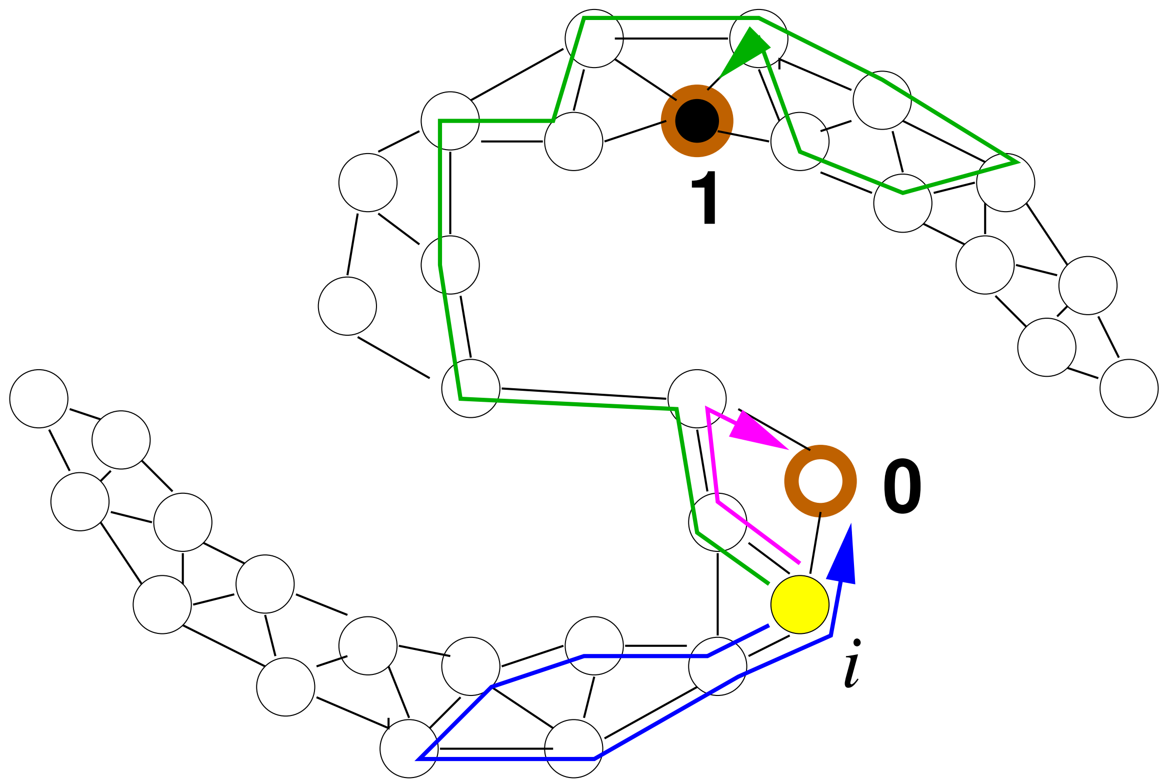
$$\Delta \mathbf{f} = 0 \;\text{ or }\; f_i = \frac{\sum_{j \sim i} w_{ij} f_j}{\sum_{j \sim i} w_{ij}}, \; i \in U$$

$$0 < f_i < 1$$

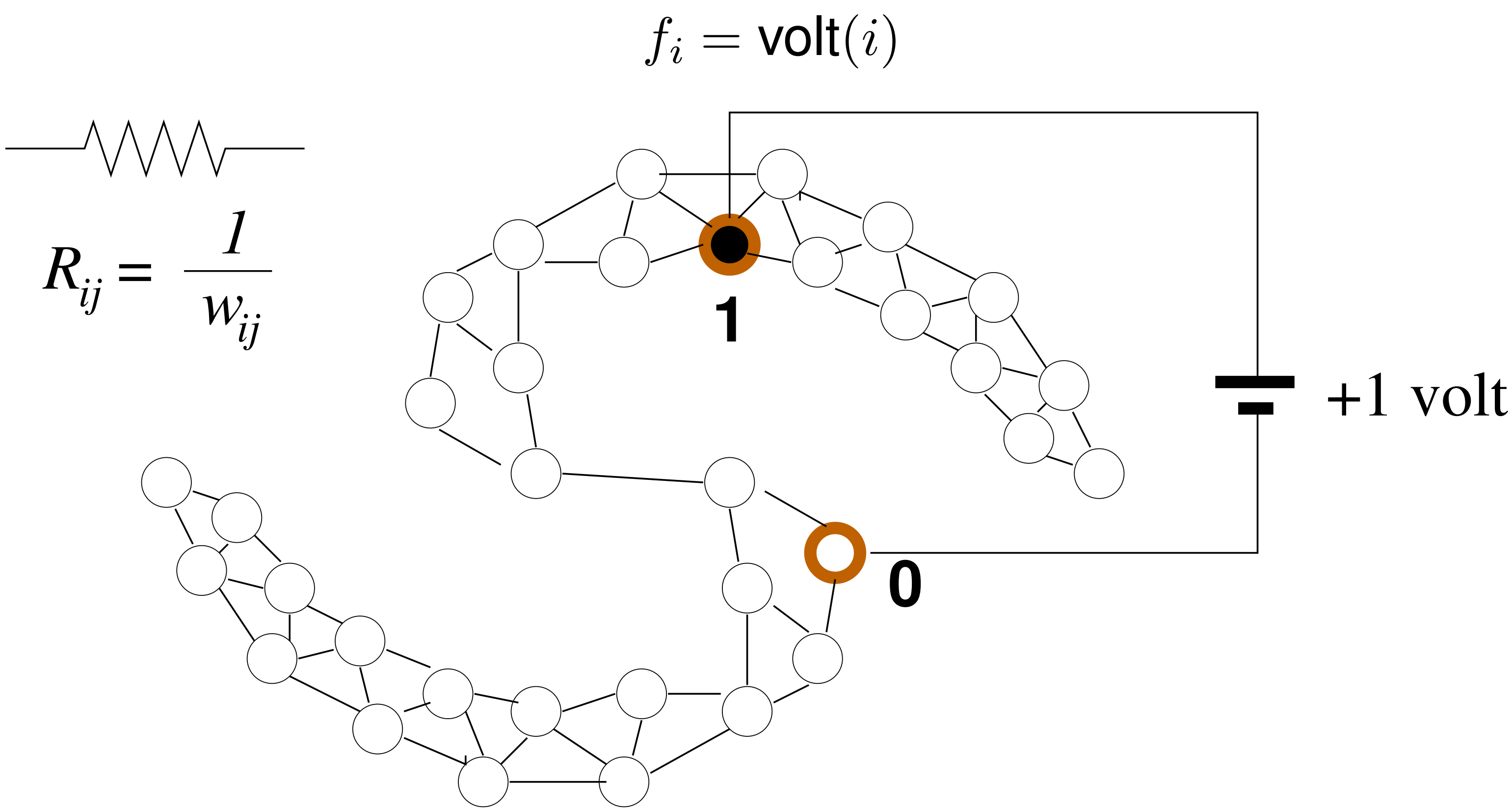
- Related to heat kernels etc. in spectral graph theory.

f_U Interpretation: Random Walks

$P(j|i) = \frac{w_{ij}}{\sum_k w_{ik}}$ $\qquad$ $f_i = P(\text{reach label } \mathbf{1} | \text{from } i)$



f_U Interpretation: Electric Networks



a

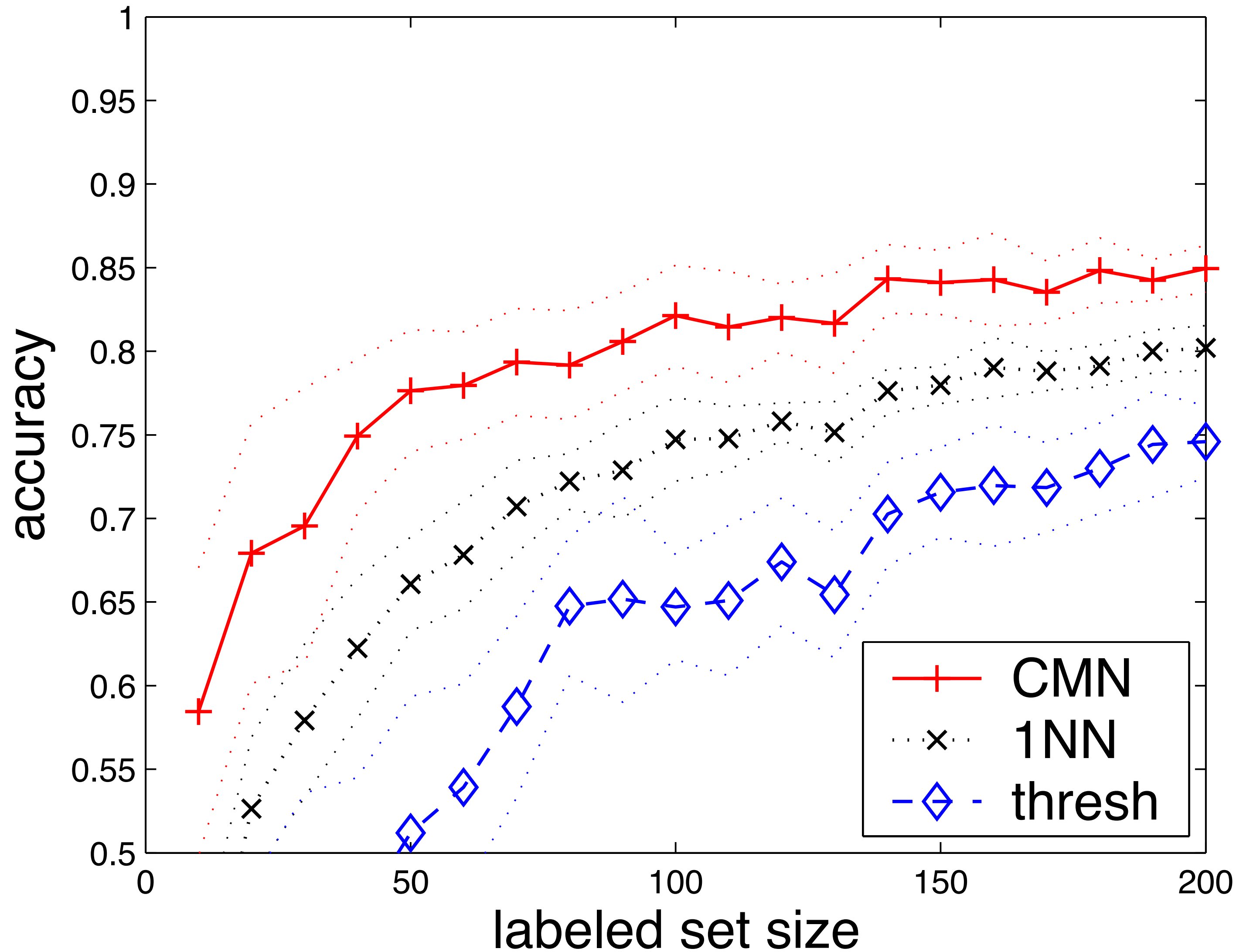
Classification

- naive: threshold f_U at 0.5. Classification often unbalanced.
- incorporating Class Priors (heuristic)

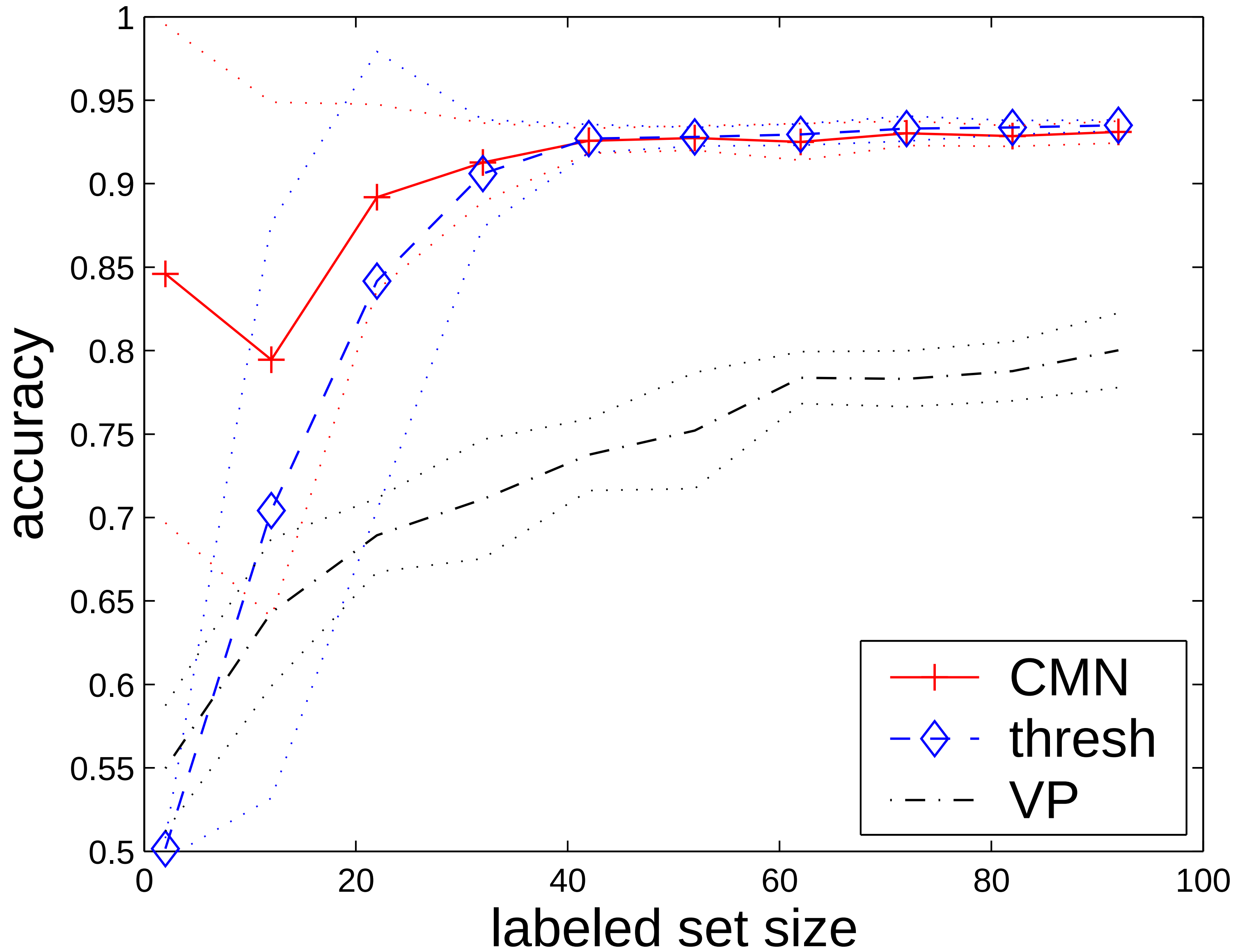
  e.g. prior: 90% class $\mathbf{1}$

$$\begin{array}{rl} \text{minimize} & E(\mathbf{y}) = \mathbf{y}^\top \Delta \mathbf{y} \\ \text{subject to} & y_L = L \\ & \text{and } \frac{\sum f_U}{|U|} = 0.9 \end{array}$$

OCR Ten Digits ($|L \cup U| = 4000$)



20-Newsgroups (PC vs. MAC, $|L \cup U| = 1943$)



Threads?

Hyperparameter Learning

Learn the graph weights (or hyperparameters):

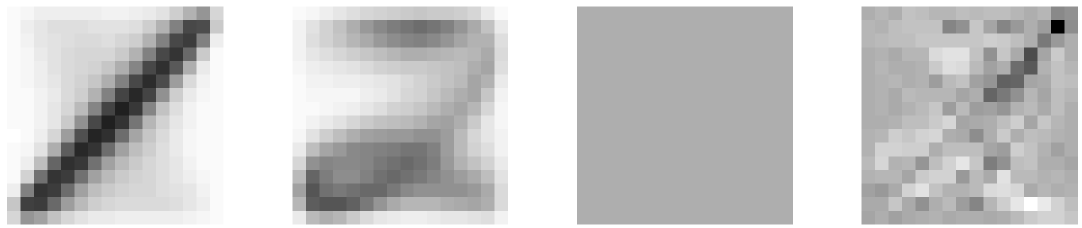
- $w_{ij} = \exp\left(-\sum_{d=1}^{m} \frac{(x_{id}-x_{jd})^2}{\sigma_d^2}\right)$, length scales;
- kNN unweighted graph, k;
- ϵNN unweighted graph, ϵ, etc.;

Hyperparameter Learning

- Minimize entropy on U (maximize label confidence);
- Evidence maximization with Gaussian process classifiers [tech report CMU-CS-03-175].

Hyperparameter Learning

OCR Digits "1" vs. "2", $|L| = 92$, $|U| = 2108$.



	H (bits)	**GF acc**
start	0.6931	94.70 $\pm$ 1.19 %
end	0.6542	98.02 $\pm$ 0.39 %

An Example Application of Graph-based SSL

Person Identification in Webcam Images: An Application of Semi-Supervised Learning

Maria-Florina Balcan NINAMF@CS.CMU.EDU
Avrim Blum AVRIM@CS.CMU.EDU
Patrick Pakyan Choi PAKYAN@CS.CMU.EDU
John Lafferty LAFFERTY@CS.CMU.EDU
Brian Pantano BPANTANO@ANDREW.CMU.EDU
Mugizi Robert Rwebangira RWEBA@CS.CMU.EDU
Xiaojin Zhu ZHUXJ@CS.CMU.EDU
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA

The FreeFoodCam

Figure 1. Four typical FreeFoodCam images.

Background Extraction

date	10/24	11/13	1/6	1/14	1/20	1/21	1/27	
1	128			193			153	474
2	256				193			448
3	288			305				593
4	204					190		394
5	266	41		189		19		515
6	195	34	179				104	512
7	126	163	200	180	70	22	28	789
8	189	66	172	117		15		559
9	189	94	215	69		30	43	640
10			65	143	122			330
total	1841	398	831	1196	384	276	328	5254

Figure 2. Left: mean background image used for background subtraction. Right: breakdown of the 10 subjects by date.

Foreground Extraction and Face Detection

Figure 3. Examples of foregrounds extracted by background subtraction and morphological transforms.

Figure 4. Examples of face images detected by the face detector.

A node and its neighbours

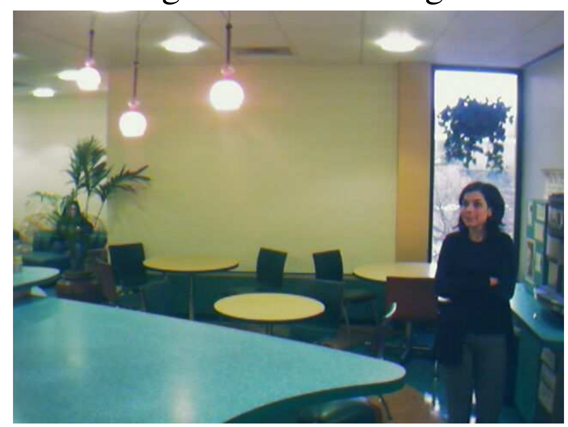
image 2910

neighbor 1: time edge

neighbor 2: color edge

neighbor 3: color edge

neighbor 4: color edge



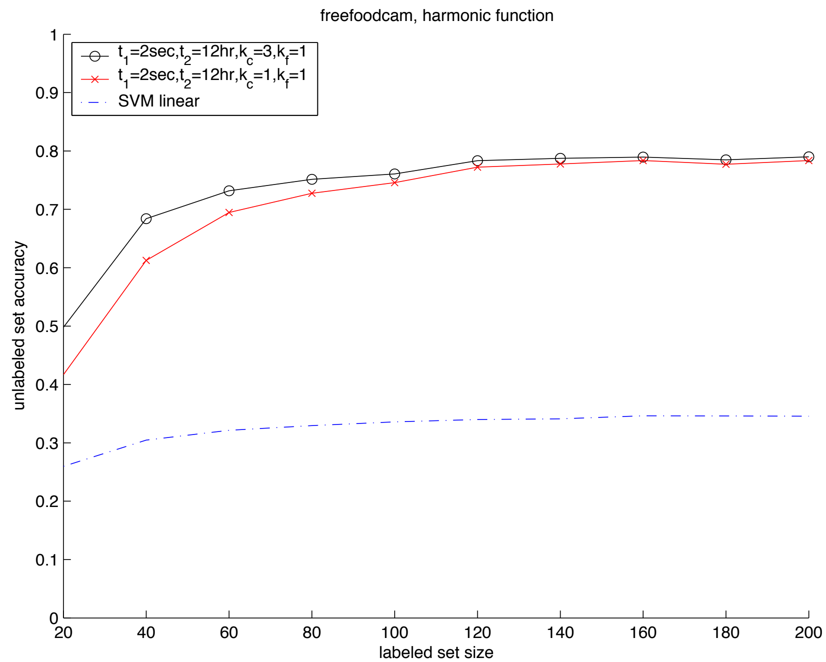
neighbor 5: face edge

Figure 5. A random image and its neighbors in the graph.

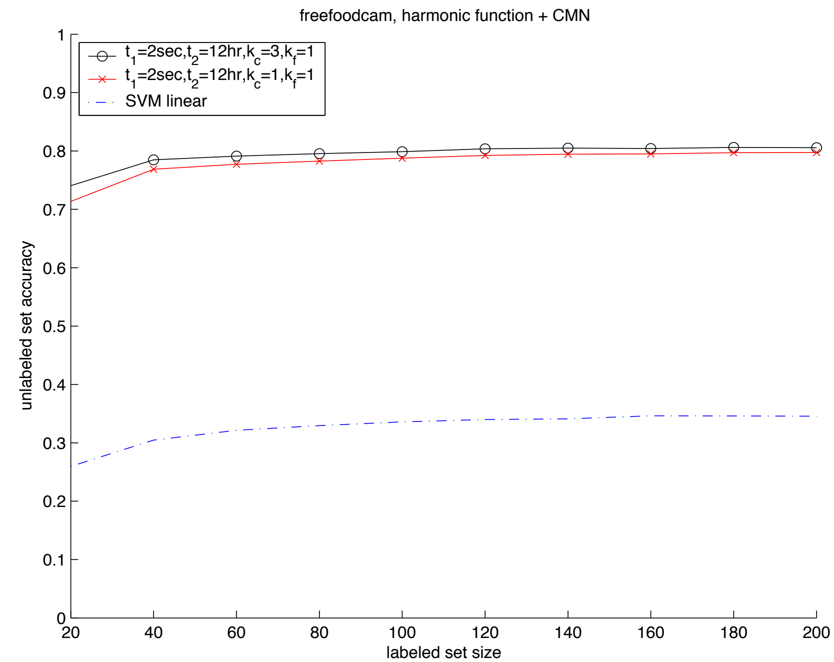
A walk on the graph

Figure 7. An example "gradient walk" on the graph. The walk starts from an unlabeled image, through assorted edges, and ends at a labeled image.

Some results



(a) harmonic function accuracy



(b) harmonic function + CMN accuracy

Figure 8. Harmonic function and CMN accuracy on two graphs. Also shown is the SVM linear kernel baseline. (a) The harmonic function algorithm significantly outperforms the linear kernel SVM, demonstrating that the semi-supervised learning algorithm successfully utilizes the unlabeled data to associate people in images with their identities. (b) The semi-supervised learning algorithm classifies even more accurately by incorporating class proportion knowledge through the CMN heuristic.

Computation

The basic computation involves solving a **sparse linear system of equations**.

$$f_U = -(\Delta_{UU})^{-1}\Delta_{UL}\mathbf{y}_L$$

Some ways of solving this for large systems:

- Conjugate gradients
- Belief propagation
- Convert the original graph into a much smaller backbone graph (Zhu and Lafferty 2005)

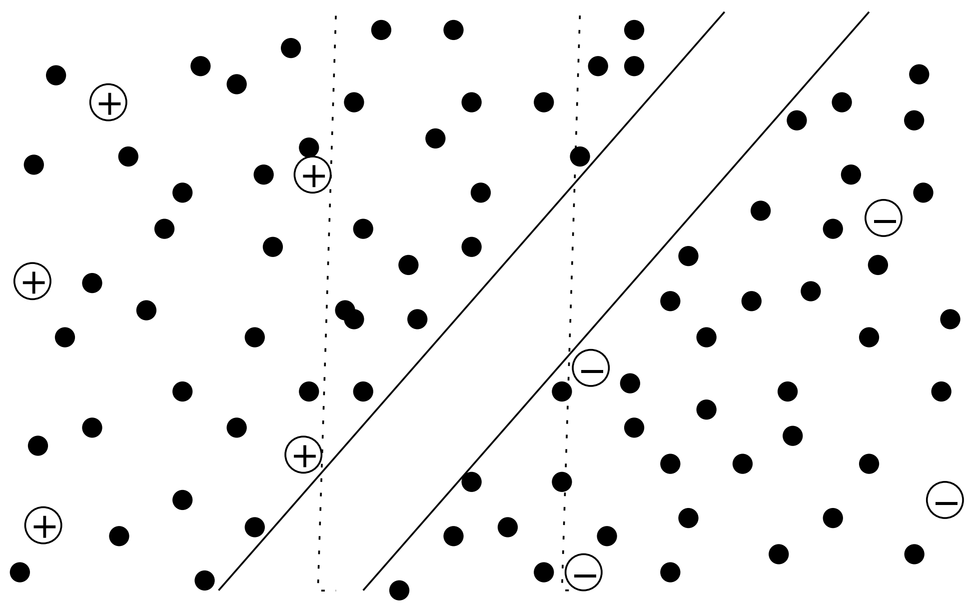
Other Approaches to Semi-supervised Learning

Caveat: This is a very big field, a lot has happened since 2003!

- Nigam et al. (2000): An EM algorithm for SSL applied to text.
- Szummer and Jaakkola (2001): SSL using Markov random walks on graphs.
- Belkin and Niyogi (2002): regularize f by using the top few eigenvectors of the Laplacian Δ
- Lawrence and Jordan (2005): a Gaussian process approach similar to TSVM using a null category noise model.
- Zhou et al (2004) use the loss function $\sum_i (f_i - y_i)^2$ and the normalised graph Laplacian $D^{-1/2} \Delta D^{-1/2}$ as a regulariser.
- Transductive SVMs (also called Semi-Supervised Support Vector Machines (S3VM)).

Transductive Support Vector Machines

Instead of finding maximum margin between labelled points, optimize over both margin and labels of unlabelled points.



Active Semi-Supervised Learning

[Zhu, Lafferty, Ghahramani, 2003]

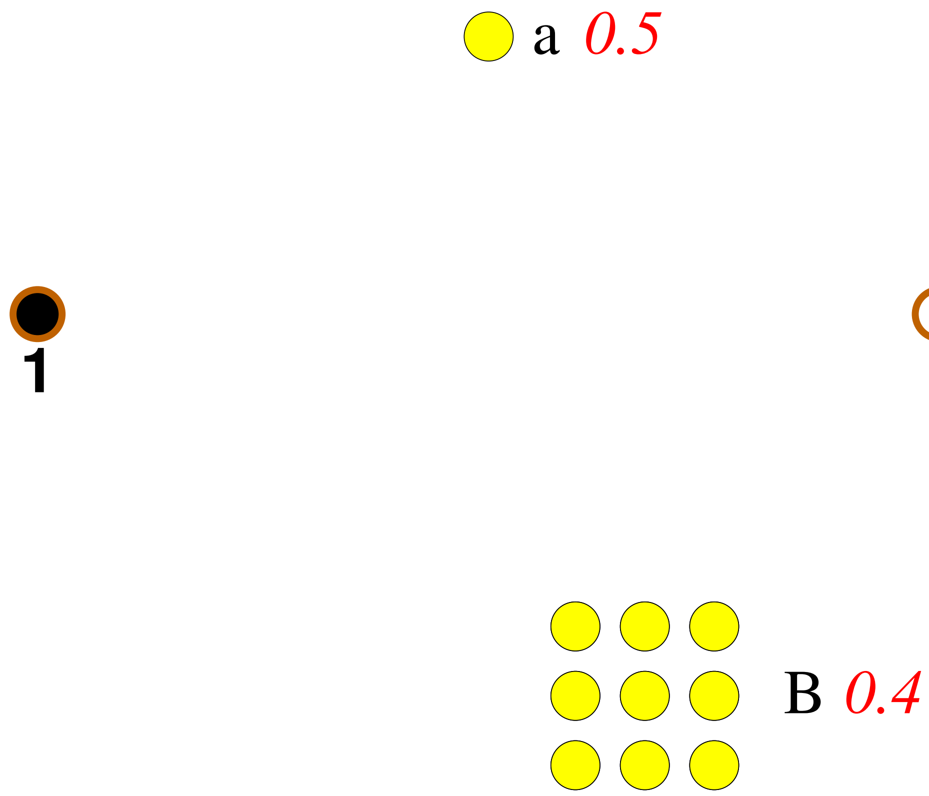
Semi-supervised learning uses U to help classification.

Active learning (pool based) selects queries in U to ask for labels.

Put it together, we have a better query selection criterion than naively selecting the point with maximum label ambiguity.

Active Learning

Select a query to minimize the estimated generalization error, not by maximum ambiguity.



Active Learning

generalization error

$$\text{err} = \sum_{i \in U} \sum_{y_i = 0,1} (\text{sgn}(f_i) \neq y_i)\, P_{\text{true}}(y_i)$$

approximation

$$P_{\text{true}}(y_i = 1) \leftarrow f_i$$

estimated generalization error

$$\hat{\text{err}} = \sum_{i \in U} \min\left(f_i, 1 - f_i\right)$$

Active Learning

estimated generalization error after querying x_k and receiving label y_k

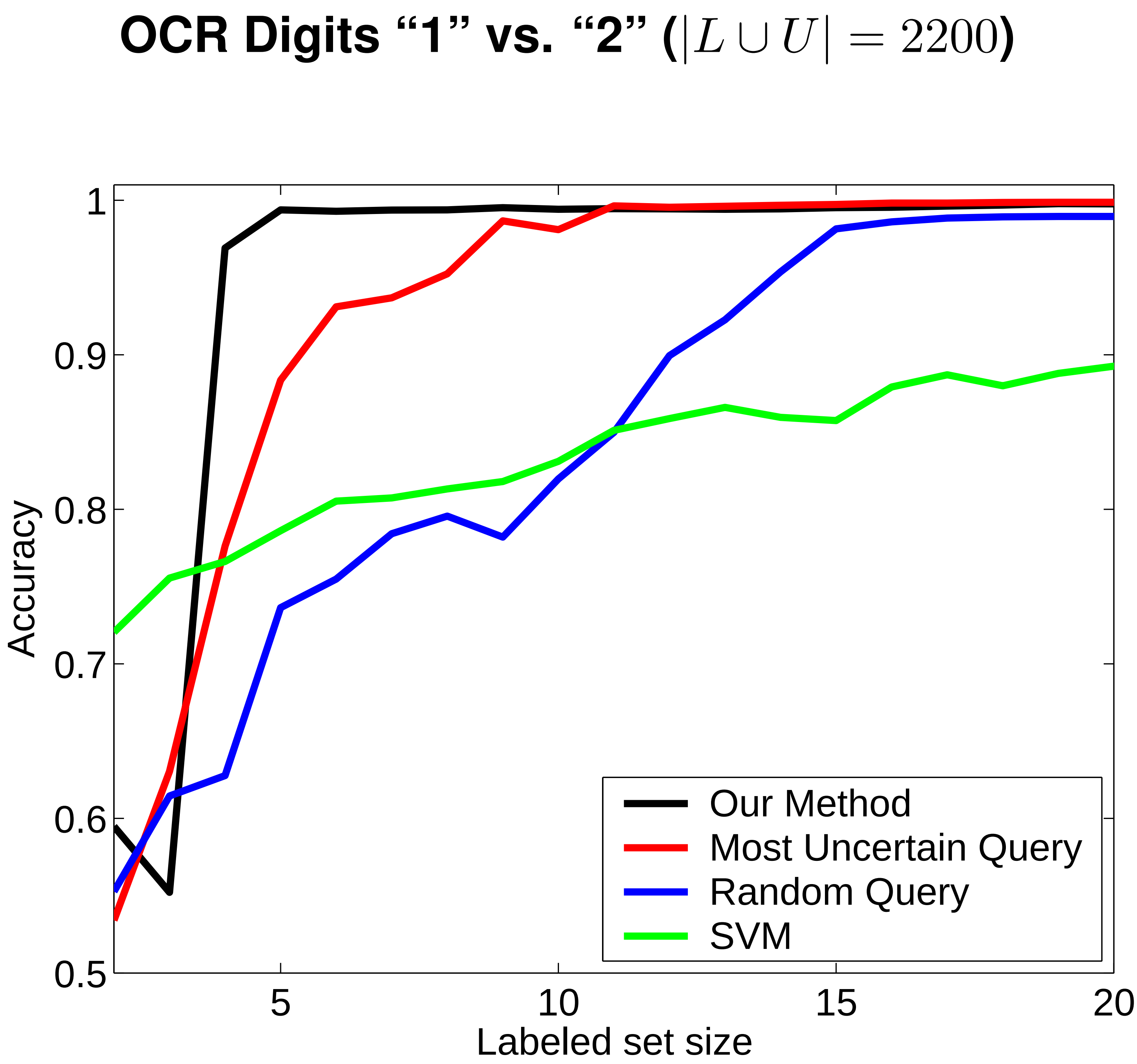
$$\hat{\text{err}}^{+(x_k,y_k)} = \sum_{i \in U} \min\left(f_i^{+(x_k,y_k)}, 1 - f_i^{+(x_k,y_k)}\right)$$

're-train' is fast for the harmonic function

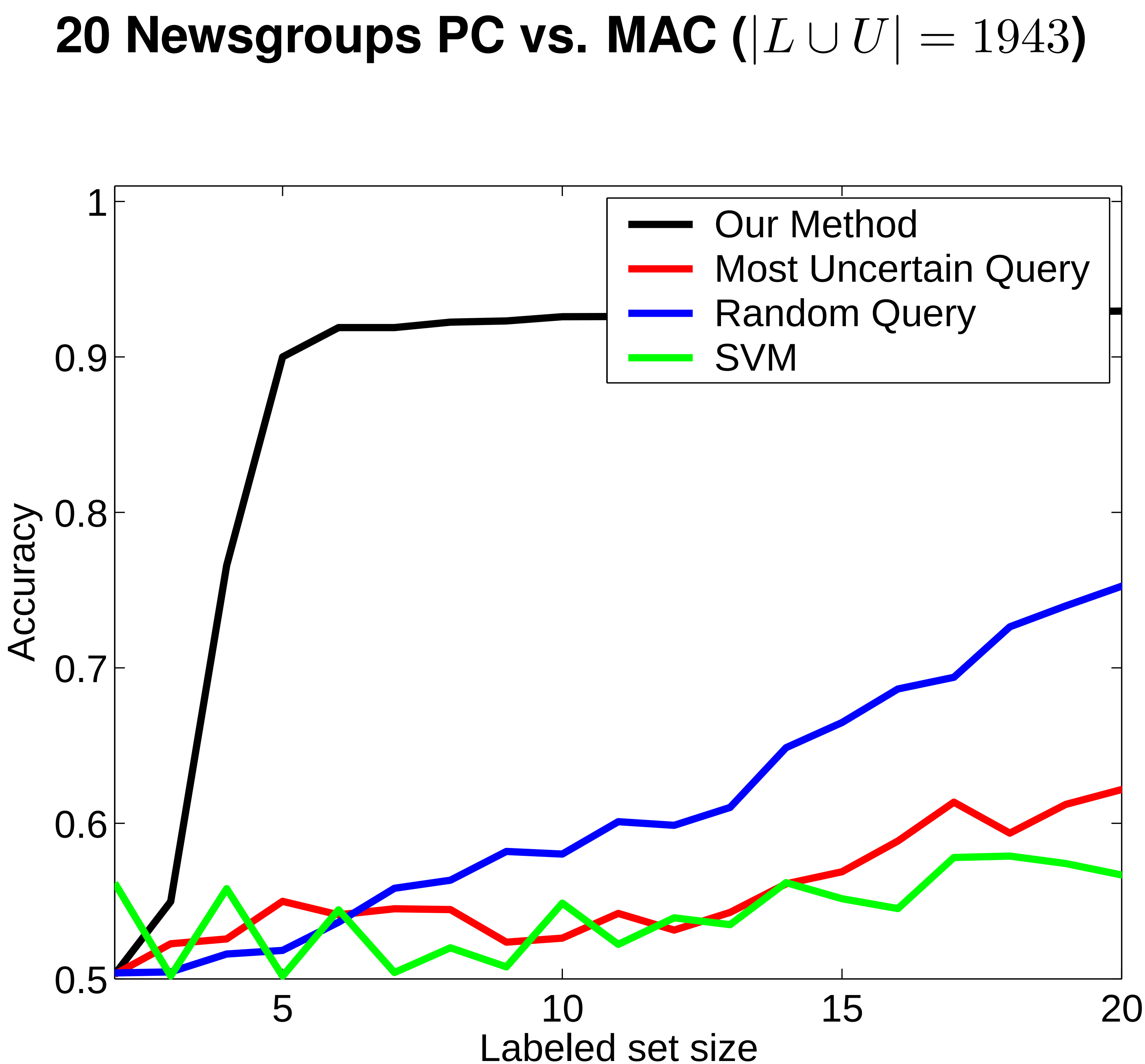
$$f_U^{+(x_k,y_k)} = f_U + (y_k - f_k)\frac{(\Delta_{UU})^{-1}_{\cdot k}}{(\Delta_{UU})^{-1}_{kk}}$$

select query k^* s.t.

$$k^* = \arg\min_k \ (1 - f_k)\hat{\text{err}}^{+(x_k,0)} + f_k \hat{\text{err}}^{+(x_k,1)}$$



OCR Digits "1" vs. "2" ($|L \cup U| = 2200$)
Accuracy
Labeled set size
1
0.9
0.8
0.7
0.6
0.5
5
10
15
20
Our Method
Most Uncertain Query
Random Query
SVM


20 Newsgroups PC vs. MAC ($|L \cup U| = 1943$)
Our Method
Most Uncertain Query
Random Query
SVM
Accuracy
Labeled set size
1
0.9
0.8
0.7
0.6
0.5
5
10
15
20

Part II: Some thoughts on Bayesian semi-supervised learning

Moving forward...

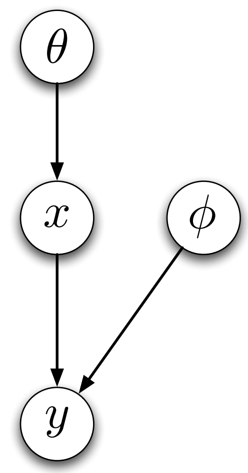
- We have good methods for transduction.
- But we don't seem to have a single unified Bayesian framework for inductive SSL.
- How would we view this problem from a fully Bayesian framework?

Bayesian Semi-Supervised Learning

x inputs, y labels:

$$p(x, y) = p(x)p(y|x) = p(y)p(x|y)$$

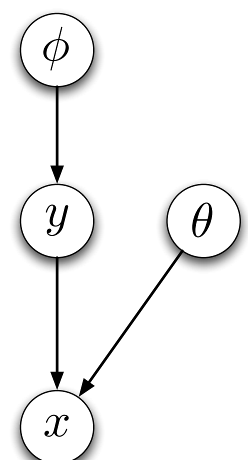
Usually we assume some model with parameters:



- Discriminative:

$$p(x, y|\theta, \phi) = p(x|\theta)p(y|x, \phi)$$

SSL possible if θ is somehow related to ϕ, works well when $p(y|x, \phi)$ is very flexible (e.g. non-parametric, kernel-based).



- Generative:

$$p(x, y|\theta, \phi) = p(y|\phi)p(x|y, \theta)$$

SSL possible but these methods are not currently widely used.

Bayesian Semi-Supervised Learning

Generative:

$$p(x, y|\theta, \phi) = p(y|\phi)p(x|y, \theta)$$

Limitations of the Generative approach:

- Often we don't *want* to model the full x.
  (Solution: maybe we can model some features of x?)

- Our models of $p(x|y, \theta)$ are usually too inflexible.
  (Solution: use non-parametric methods?)

Some examples:

- Kemp et al (2003) Semi-supervised learning with trees.

- Radford Neal's entry using Dirichlet Diffusion trees into the NIPS feature selection competition.

From a Bayesian perspective, semi-supervised learning is just another missing data problem!

Summary

- Semi-supervised learning with harmonic functions
- Active semi-supervised learning using harmonic functions by minimizing expected generalization error
- Much research in this area but still some open questions...

References

- Balcan, M.-F., Blum, A., Choi, P. P., Lafferty, J., Pantano, B., Rwebangira, M. R., & Zhu, X. (2005a). Person identification in webcam images: An application of semi-supervised learning. ICML 2005 Workshop on Learning with Partially Classified Training Data.
- Blum, A., & Chawla, S. (2001). Learning from labeled and unlabeled data using graph mincuts. ICML 18.
- Joachims, T. (1999). Transductive inference for text classification using support vector machines. ICML 16: 200-209.
- Lawrence, N. D., & Jordan, M. I. (2005). Semi-supervised learning via Gaussian processes. NIPS 17.
- Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39, 103-134.
- Seeger, M. (2001). Learning with labeled and unlabeled data (Technical Report). University of Edinburgh.
- Szummer, M., & Jaakkola, T. (2001). Partially labeled classification with Markov random walks. NIPS 14.
- Zhou, D., Bousquet, O., Lal, T., Weston, J., & Schölkopf, B. (2004). Learning with local and global consistency. NIPS 16.
- Zhu, X., Ghahramani, Z. and Lafferty, J. (2003) Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions. ICML 20: 912–919.
- Zhu, X., & Lafferty, J. (2005). Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning. ICML 22.
- Zhu, X., Lafferty, J. and Ghahramani, Z. (2003) Combining Active Learning and Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions. In *ICML 2003 Workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*. pp 58–65.
- Zhu, X. and Goldberg, A.B. (2009) *Introduction to Semi-Supervised Learning*. Morgan-Claypool.