Semi-Supervised Learning

Modified from Avrim Blum's slides

~~Semi-~~Supervised Learning

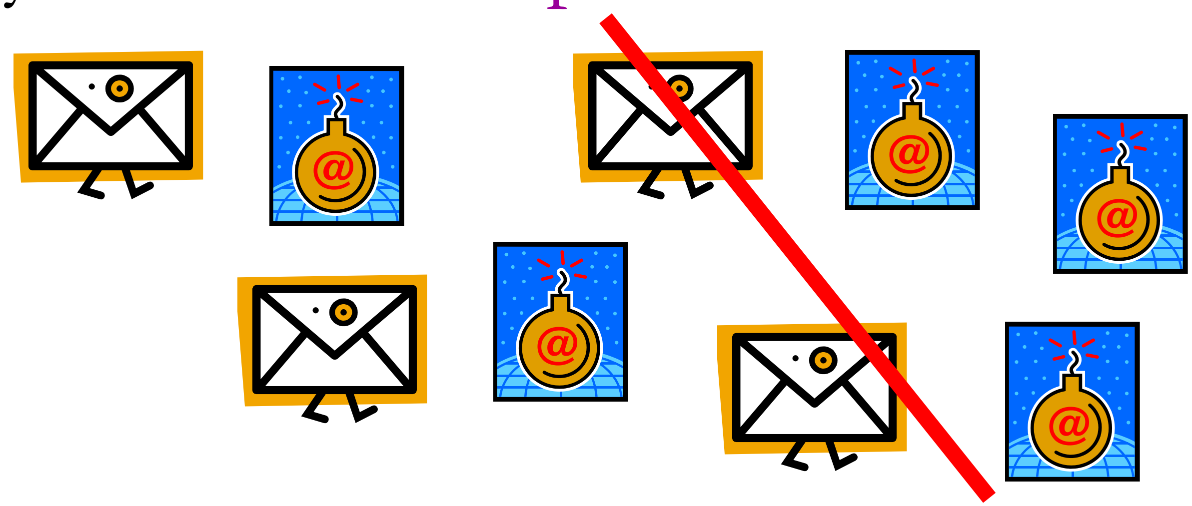
Supervised Learning = learning from labeled data. Dominant paradigm in Machine Learning.

- For example, you want to train an email classifier to distinguish spam from important messages

~~Semi-~~Supervised Learning

Supervised Learning = learning from labeled data. Dominant paradigm in Machine Learning.

- E.g, say you want to train an email classifier to distinguish spam from important messages
- Take sample S of data, labeled according to whether they were/weren't spam.



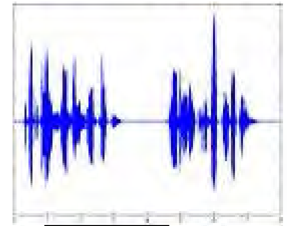
~~Semi-~~Supervised Learning

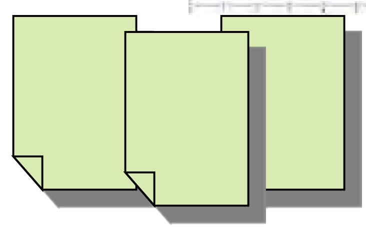
Supervised Learning = learning from labeled data. Dominant paradigm in Machine Learning.

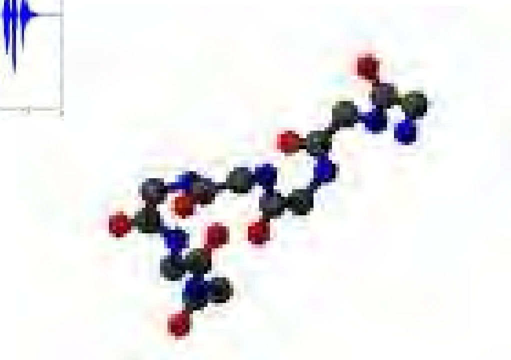
- E.g, say you want to train an email classifier to distinguish spam from important messages
- Take sample S of data, labeled according to whether they were/weren't spam.
- Train a classifier (like SVM, decision tree, etc) on S. Make sure it's not overfitting.
- Use to classify new emails.

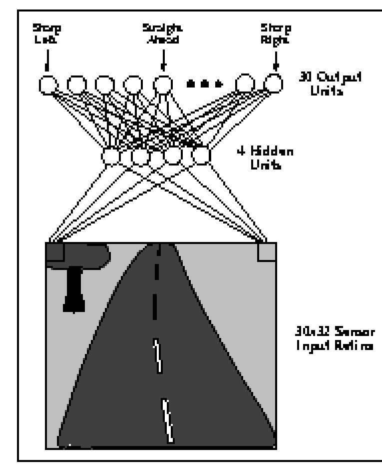
Basic paradigm has many successes

- recognise speech,
- steer a car,
- classify documents
- classify proteins
- recognising faces, objects in images
- ...









However, for many problems, labeled data can be rare or expensive.

Need to pay someone to do it, requires special testing,…

Unlabeled data is much cheaper.

However, for many problems, labeled data can be rare or expensive.

Need to pay someone to do it, requires special testing,…

Unlabeled data is much cheaper.

Speech

Customer modeling

Images

Protein sequences

Medical outcomes

Web pages

However, for many problems, labeled data can be rare or expensive.

Need to pay someone to do it, requires special testing,…

Unlabeled data is much cheaper.

[From Jerry Zhu]

Task: speech analysis

- Switchboard dataset
- telephone conversation transcription
- 400 hours annotation time for each hour of speech

film $\Rightarrow$ `f ih_n uh_gl_n m`

be all $\Rightarrow$ `bcl b iy iy_tr ao_tr ao l_dl`

However, for many problems, labeled data can be rare or expensive.

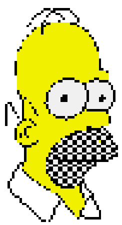
Need to pay someone to do it, requires special testing,…

Unlabeled data is much cheaper.

Can we make use of cheap unlabeled data?

Semi-Supervised Learning

Can we use unlabeled data to augment a small labeled sample to improve learning?



But unlabeled data is missing the most important info!!

But maybe still has useful regularities that we can use.

Bu Bu But…

Semi-Supervised Learning

Substantial recent work in ML. A number of interesting methods have been developed.

This lecture:

- Discuss several diverse methods for taking advantage of unlabeled data.

Method 1:

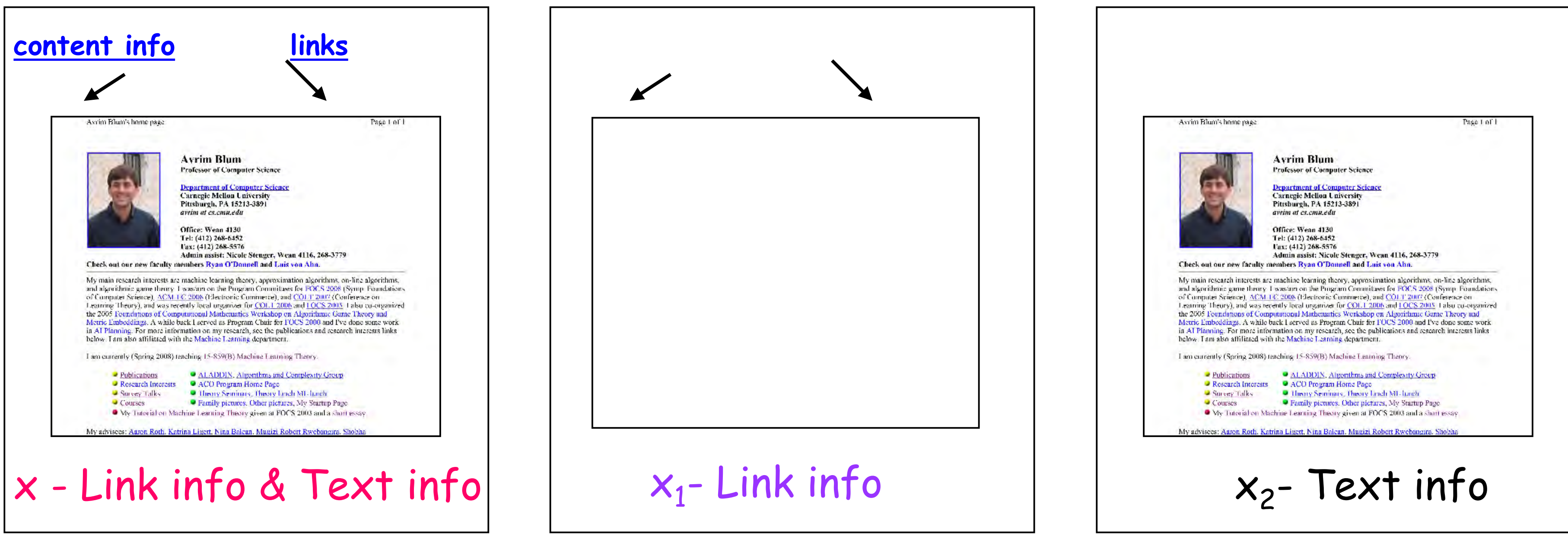
Co-Training

Co-training

[Blum&Mitchell'98]

Many problems have two different sources of info you can use to determine label.

E.g., classifying webpages: can use words on page or words on links pointing to the page.



Co-training

Idea: Use small labeled sample to learn initial rules.

– e.g., “colleagues” pointing to a page is a good indicator it is a faculty home page.

– e.g., “I am teaching ML course” on a page is a good indicator it is a faculty home page.

Links to Co-workers

Avrim Blum's home page — Page 1 of 1

Avrim Blum
Professor of Computer Science

Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3891
avrim at cs.cmu.edu

Office: Wean 4130
Tel: (412) 268-6452
Fax: (412) 268-5576
Admin assist: Nicole Stenger, Wean 4116, 268-3779

Check out our new faculty members Ryan O'Donnell and Luis von Ahn.

My main research interests are machine learning theory, approximation algorithms, on-line algorithms, and algorithmic game theory. I was/am on the Program Committees for FOCS 2008 (Symp. Foundations of Computer Science), ACM-EC 2008 (Electronic Commerce), and COLT 2007 (Conference on Learning Theory), and was recently local organizer for COLT 2006 and FOCS 2005. I also co-organized the 2005 Foundations of Computational Mathematics Workshop on Algorithmic Game Theory and Metric Embeddings. A while back I served as Program Chair for FOCS 2000 and I've done some work in AI Planning. For more information on my research, see the publications and research interests links below. I am also affiliated with the Machine Learning department.

I am currently (Spring 2008) teaching 15-859(B) Machine Learning Theory.

- Publications
- Research Interests
- Survey Talks
- Courses
- ALADDIN, Algorithms and Complexity Group
- ACO Program Home Page
- Theory Seminars, Theory lunch ML lunch
- Family pictures, Other pictures, My Startup Page
- My Tutorial on Machine Learning Theory given at FOCS 2003 and a short essay.

My advisees: Aaron Roth, Katrina Ligett, Nina Balcan, Mugizi Robert Rwebangira, Shobha Venkataraman.

Past advisees: Prasad Chalasani, Santosh Vempala, Carl Burch, Adam Kalai, John Langford, Nikhil Bansal, Martin Zinkevich, Shuchi Chawla, Brendan McMahan.

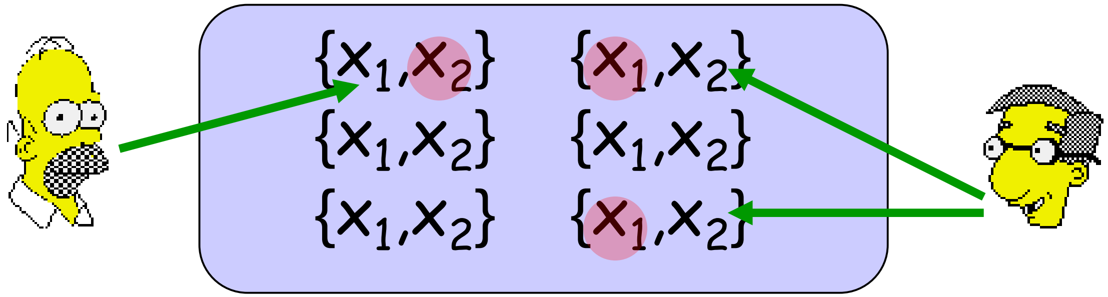
Google

Co-training

Idea: Use small labeled sample to learn initial rules.

Then look for unlabeled examples where one rule is confident and the other is not. Have it label the example for the other.

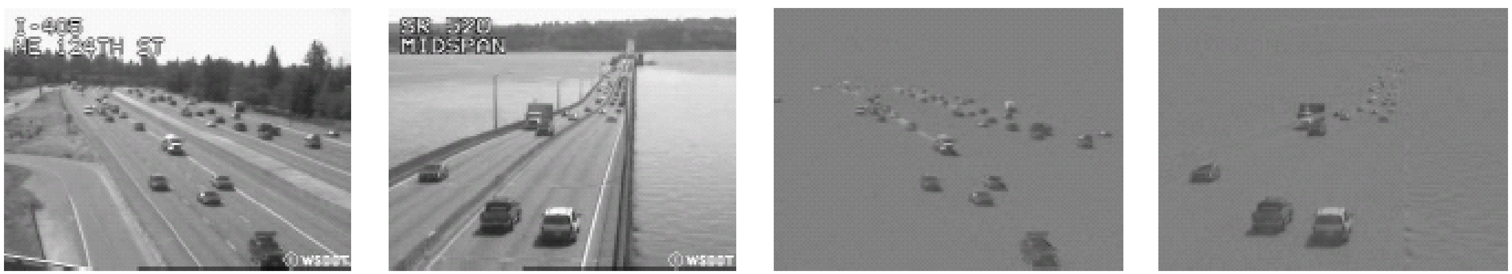
Training 2 classifiers, one on each type of info. Using each to help train the other.



Co-training

Turns out a number of problems can be set up this way.

E.g., [Levin-Viola-Freund03] identifying objects in images. Two different kinds of preprocessing.



E.g., [Collins&Singer99] named-entity extraction.

- "I arrived in London yesterday"

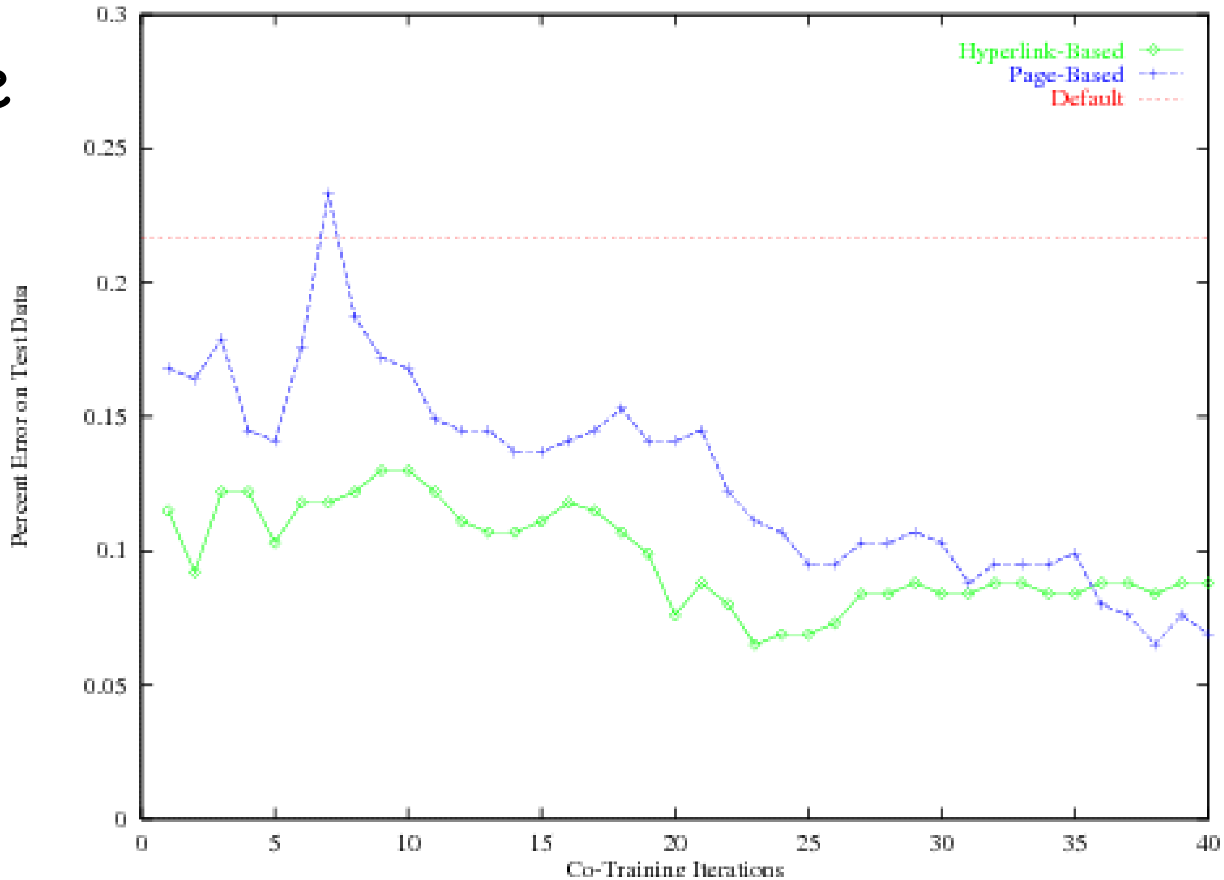
- …

Results: webpages

12 labeled examples, 1000 unlabeled

	Page-based	Hyperlink-based	Combined
Std. Supervised	12.9	12.4	11.1
Co-training	6.2	11.6	5.0
Just say neg	22	22	22

(sample run)



Results: images [Levin-Viola-Freund '03]:

- Visual detectors with different kinds of processing

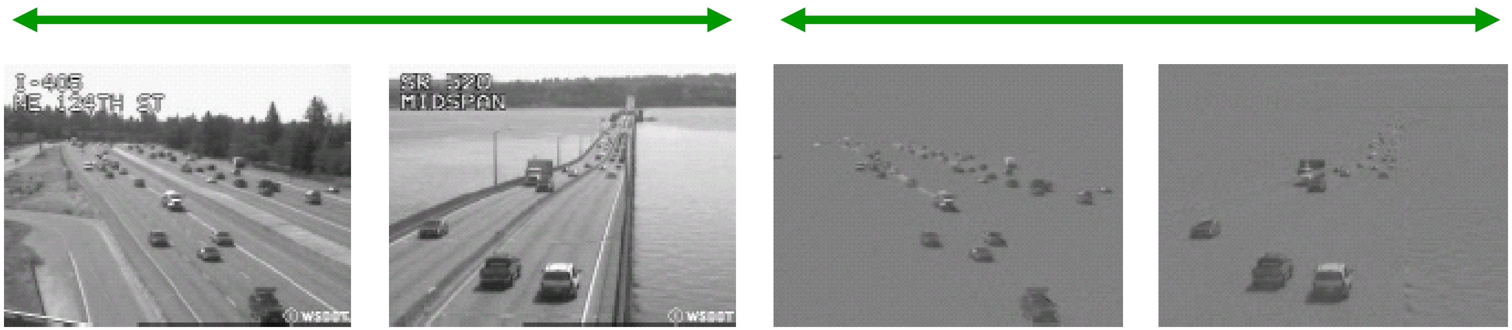
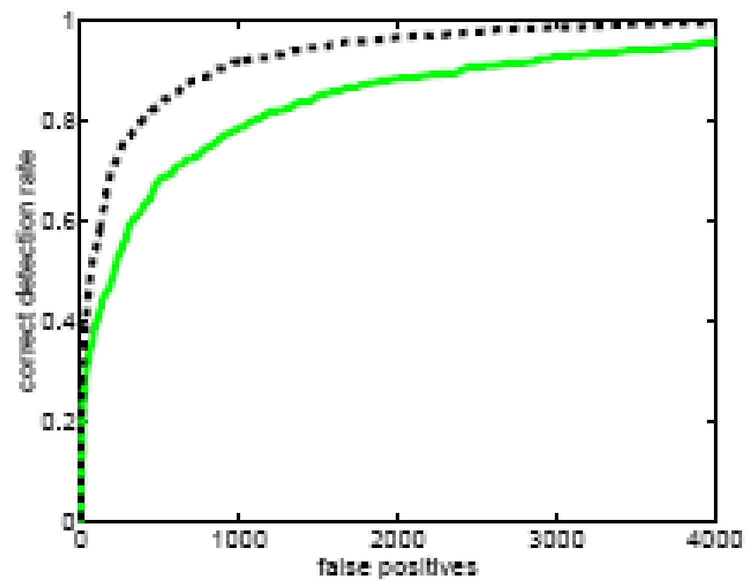


Figure 1: Example images used to test and train the car detection system. On the left are the original images. On the right are background subtracted images.

- Images with 50 labeled cars. 22,000 unlabeled images.
- Factor 2-3+ improvement.

From [LVF03]

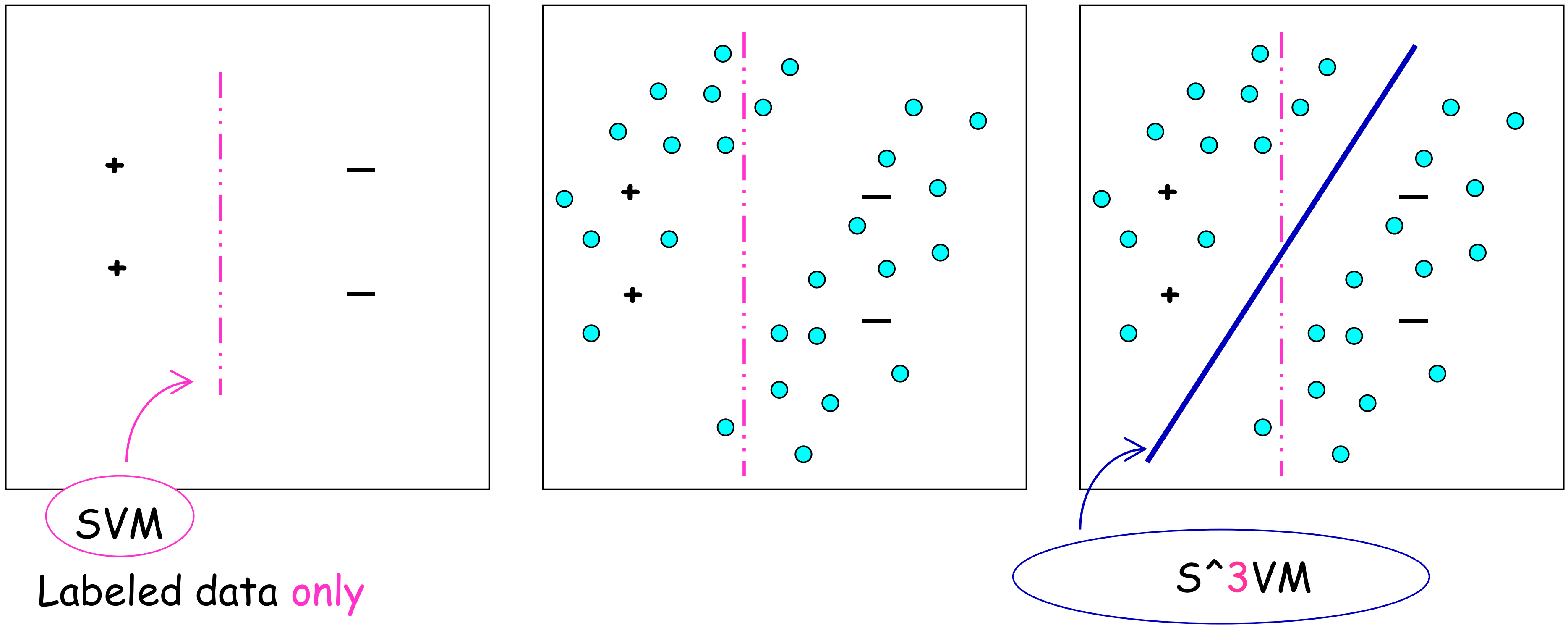


Method 2:

Semi-Supervised (Transductive) SVM

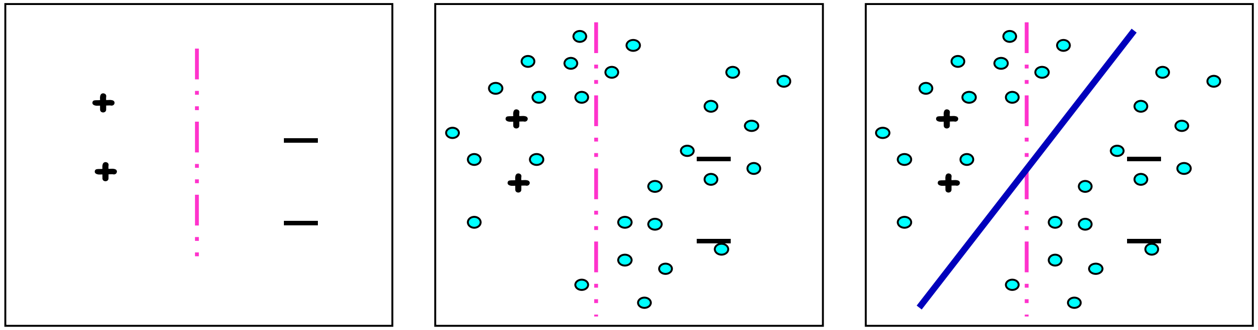
S^{3}VM [Joachims98]

- Suppose we believe target separator goes through low density regions of the space/large margin.
- Aim for separator with large margin wrt labeled and unlabeled data. (L+U)



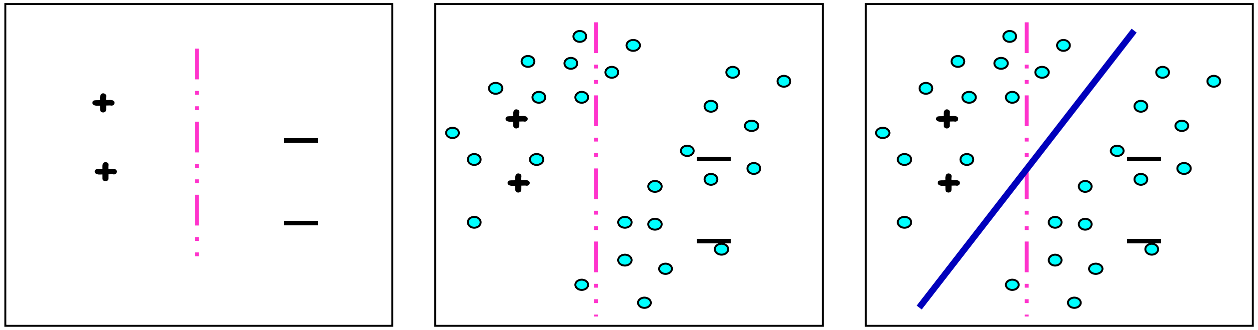
S^{3}VM [Joachims98]

- Suppose we believe target separator goes through low density regions of the space/large margin.
- Aim for separator with large margin wrt labeled and unlabeled data. (L+U)
- Unfortunately, optimization problem is now NP-hard. Algorithm instead does local optimization.
  - Start with large margin over labeled data. Induces labels on U.
  - Then try flipping labels in greedy fashion.

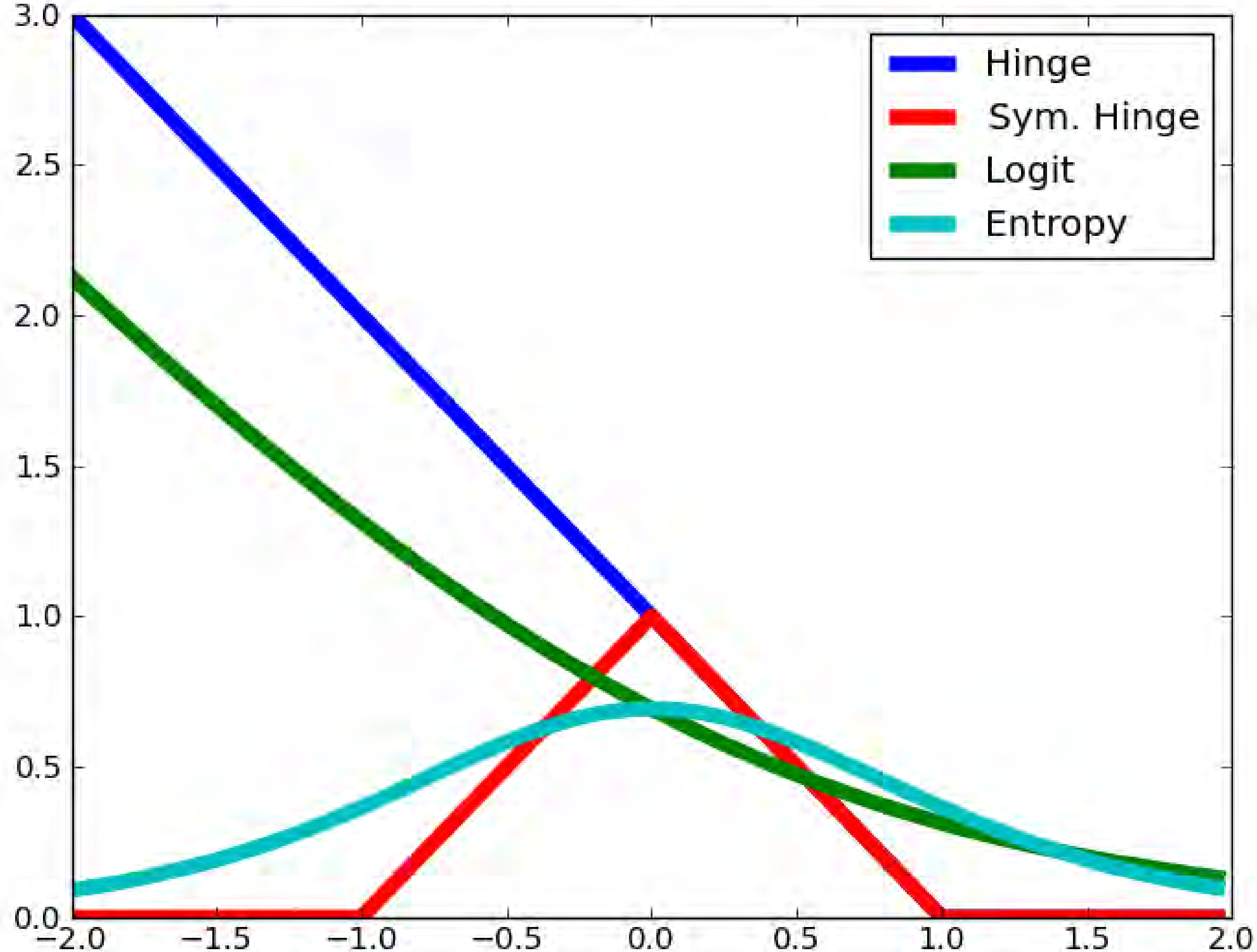


S^3VM [Joachims98]

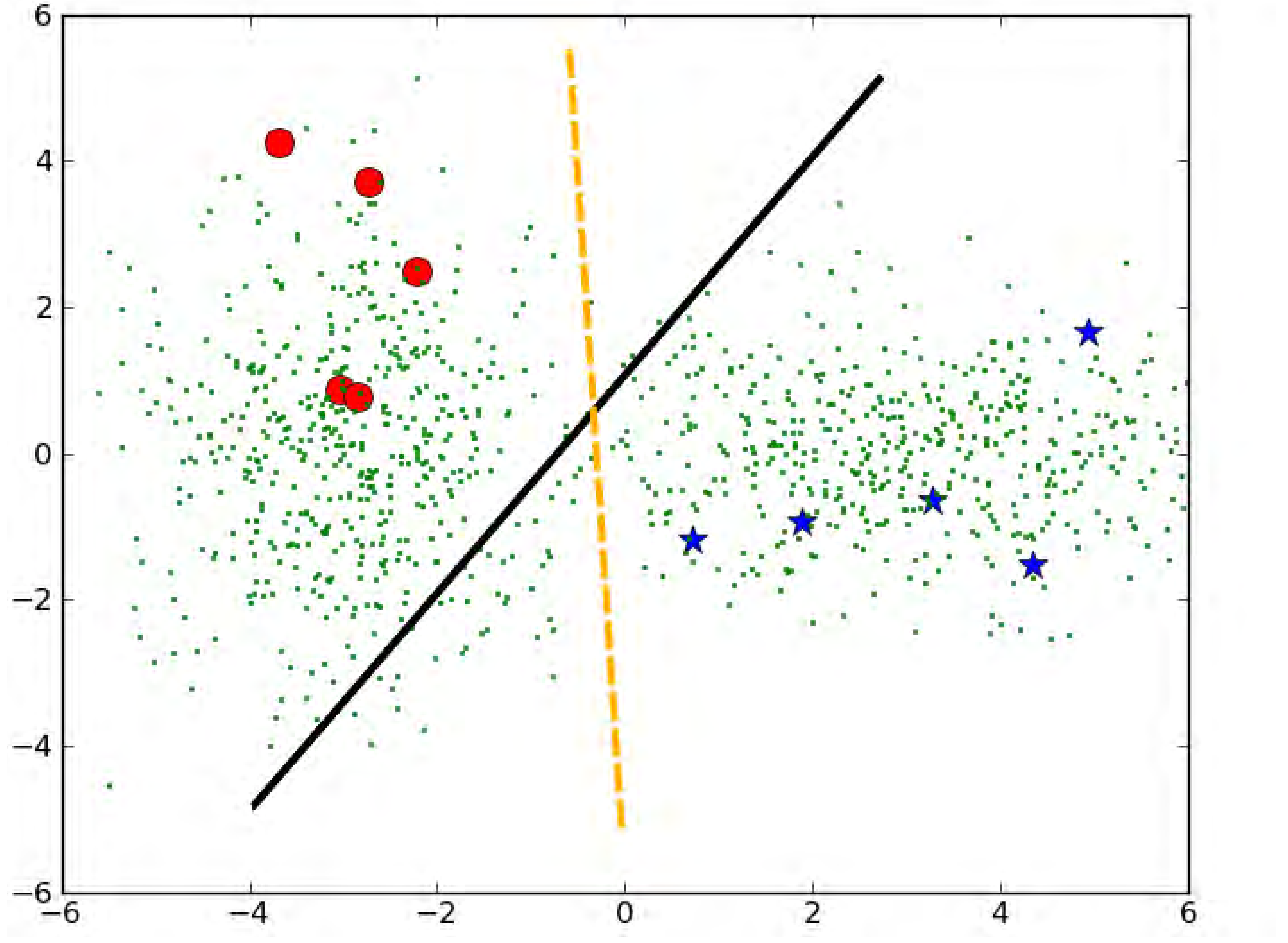
- Suppose we believe target separator goes through low density regions of the space/large margin.
- Aim for separator with large margin wrt labeled and unlabeled data. (L+U)
- Unfortunately, optimization problem is now NP-hard. Algorithm instead does local optimization.
  - Or, branch-and-bound, other methods (Chapelle etal06)
- Quite successful on text data.



Loss of S3VM



Hinge
Sym. Hinge
Logit
Entropy
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0



6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6

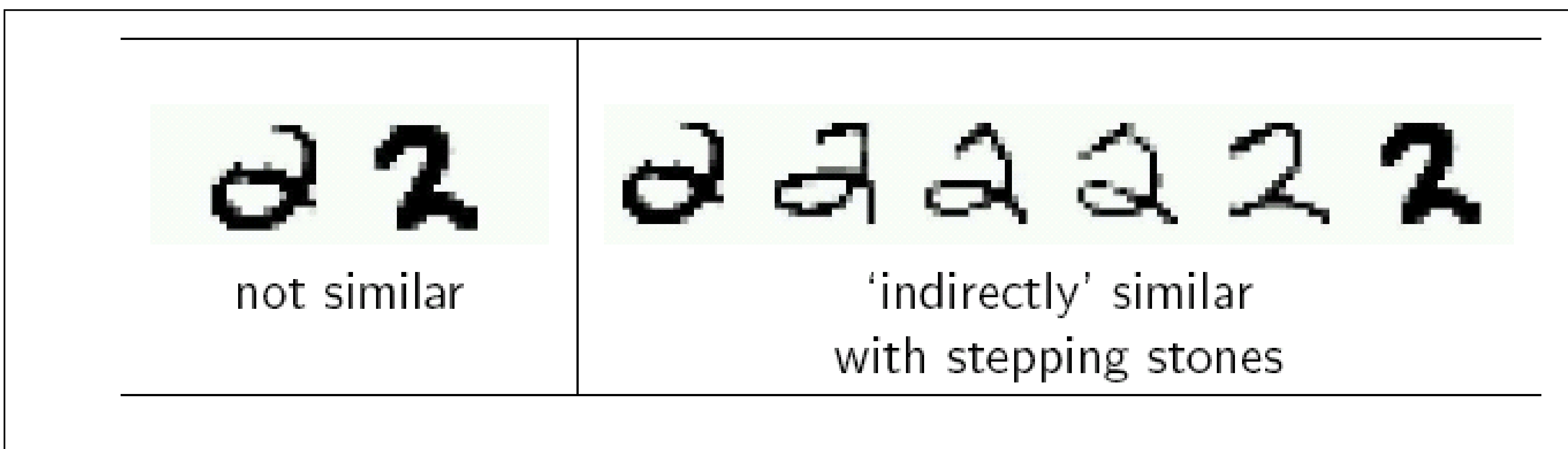
Method 3:

Graph-based methods

Graph-based methods

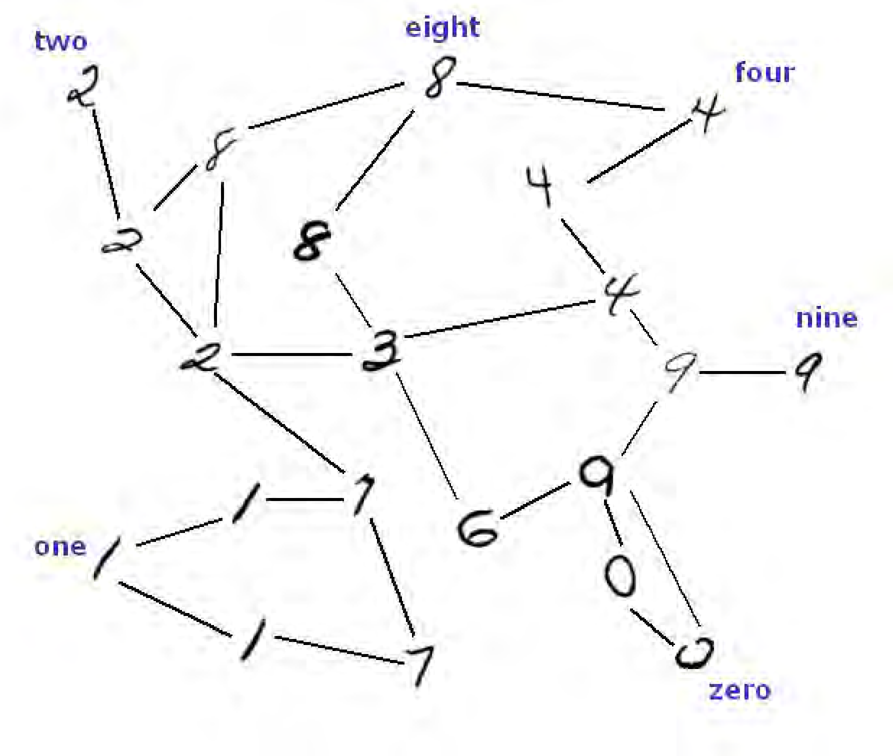
- Suppose we believe that very similar examples probably have the same label.
- If you have a lot of labeled data, this suggests a Nearest-Neighbor type of alg.
- If you have a lot of unlabeled data, perhaps can use them as "stepping stones"

E.g., handwritten digits [Zhu07]:



Graph-based methods

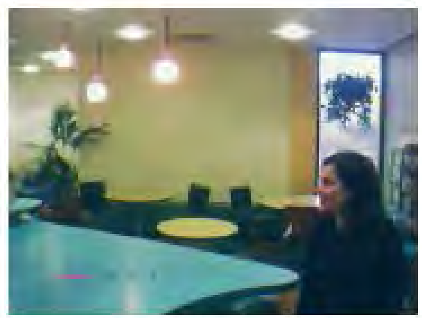
- Idea: construct a graph with edges between very similar examples.
- Unlabeled data can help "glue" the objects of the same class together.



Graph-based methods

- Idea: construct a graph with edges between very similar examples.
- Unlabeled data can help "glue" the objects of the same class together.

image 4005



neighbor 1: time edge

neighbor 2: color edge

neighbor 3: color edge

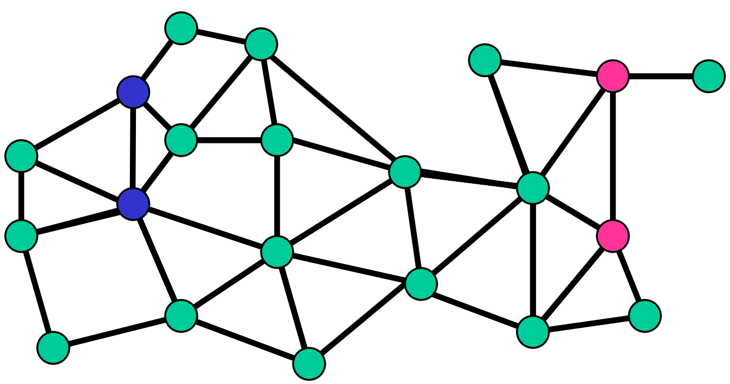
neighbor 4: color edge

neighbor 5: face edge

Graph-based methods

- Idea: construct a graph with edges between very similar examples.
- Unlabeled data can help "glue" the objects of the same class together.
- Solve for:
  - Minimum cut [BC,BLRR]
  - Minimum "soft-cut" [ZGL]

    $\sum_{e=(u,v)}(f(u)-f(v))^2$
  - Spectral partitioning [J]
  - ...



Laplacian SVM

LapSVM

$$\min_{w} \frac{\lambda}{2}\|w\|_2^2 + \frac{1}{|\mathcal{D}_l|} \sum_{(\mathbf{x},y)\in\mathcal{D}_l} \max(0, 1 - y\langle w, \mathbf{x}\rangle) +$$

$$+ \frac{\gamma}{|\mathcal{D}_u|^2} \sum_{\mathbf{x}\in\mathcal{D}_u} \sum_{\mathbf{x}'\in\mathcal{D}_l\cup\mathcal{D}_u} s(\mathbf{x},\mathbf{x}')(\langle w,\mathbf{x}\rangle - \langle w,\mathbf{x}'\rangle)^2$$

Graph Laplacian

$$\min_{w} \frac{\lambda}{2}\|w\|_2^2 + \frac{1}{|\mathcal{D}_l|} \sum_{(\mathbf{x},y)\in\mathcal{D}_l} \max(0, 1 - y\langle w, \mathbf{x}\rangle) +$$

$$+ \frac{\gamma}{|\mathcal{D}_u|^2} \langle \mathbf{f}, L\mathbf{f} \rangle$$

Conclusions

- Semi-supervised learning is an area of increasing importance in Machine Learning.
- Automatic methods of collecting data make it more important than ever to develop methods to make use of unlabeled data.
- Several promising algorithms (only discussed a few).