Reading the Web:
Advanced Statistical Language Processing

www.cs.cmu.edu/~tom/rtw09/

Machine Learning 10-709

September 24, 2009

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

Today

- Coupled Semi-supervised training of multiple functions
  - Theory
  - Algorithms Co-training, CoEM, Co-regularization
- News:
  - class Wiki (courtesy Mehrbod Sharifi)
  - new software to access KBs
- HW for next week

When can Unlabeled Data help supervised learning?

Problem setting (the PAC learning setting):

- Set *X* of instances drawn from unknown distribution *P(X)*
- Wish to learn target function *f: X→ Y* (or, P(Y|X))
- Given a set *H* of possible hypotheses for *f*

Given:

- i.i.d. labeled examples $L = \{\langle x_1, y_1\rangle \ldots \langle x_m, y_m\rangle\}$
- i.i.d. unlabeled examples $U = \{x_{m+1}, \ldots x_{m+n}\}$

Wish to find hypothesis with lowest true error:

$$\hat{f} \leftarrow \arg\min_{h \in H} \Pr_{x \in P(X)}[h(x) \neq f(x)]$$

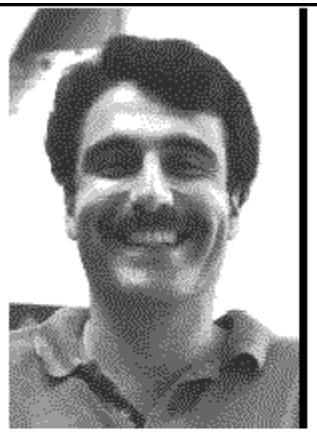
One Idea: Coupled Training

- In some settings, available data features are redundant and we can train two classifiers based on disjoint features
- In this case, the two classifiers should agree on the classification for each unlabeled example
- Therefore, we can use the unlabeled data to constrain joint training of both classifiers

Redundantly Sufficient Features

Professor Faloutsos

my advisor



U.S. mail address:
Department of Computer Science
University of Maryland
College Park, MD 20742
(97-99: on leave at CMU)
Office: 3227 A.V. Williams Bldg.
Phone: (301) 405-2695
Fax: (301) 405-6707
Email: christos@cs.umd.edu

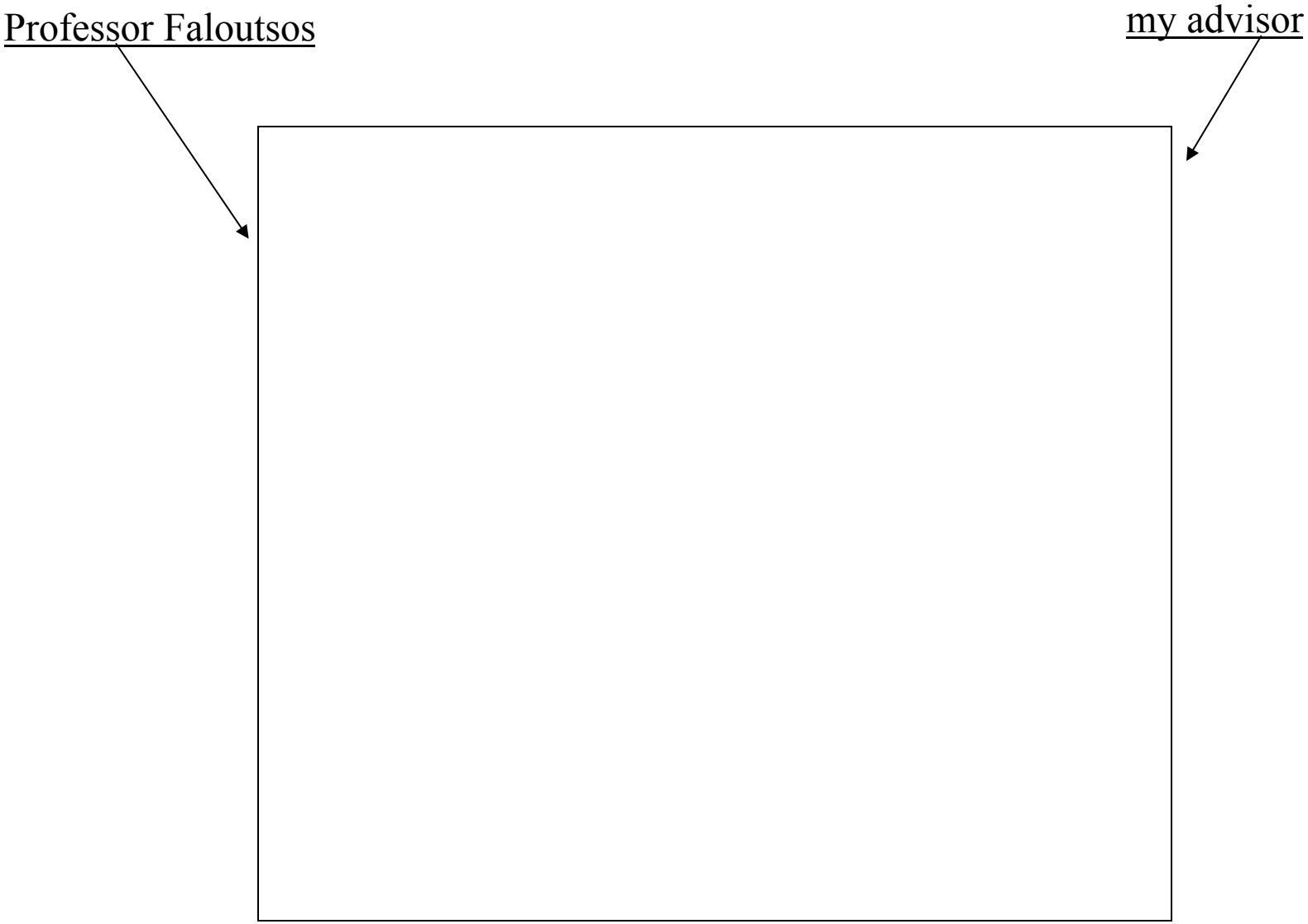
Christos Faloutsos

Current Position: Assoc. Professor of Computer Science. (97-98: on leave at CMU)
Join Appointment: Institute for Systems Research (ISR).
Academic Degrees: Ph.D. and M.Sc. (University of Toronto.); B.Sc. (Nat. Tech. U. Ath

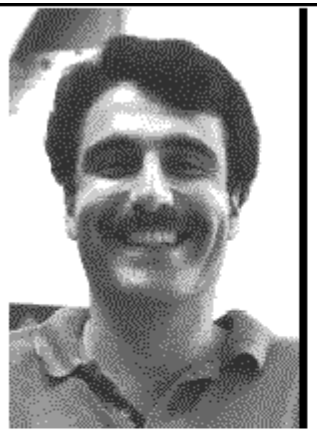
Research Interests:

- Query by content in multimedia databases;
- Fractals for clustering and spatial access methods;
- Data mining;

Redundantly Sufficient Features



Redundantly Sufficient Features



U.S. mail address:
Department of Computer Science
University of Maryland
College Park, MD 20742
(97-99: on leave at CMU)
Office: 3227 A.V. Williams Bldg.
Phone: (301) 405-2695
Fax: (301) 405-6707
Email: christos@cs.umd.edu

Christos Faloutsos

Current Position: Assoc. Professor of Computer Science. (97-98: on leave at CMU)
Join Appointment: Institute for Systems Research (ISR).
Academic Degrees: Ph.D. and M.Sc. (University of Toronto.); B.Sc. (Nat. Tech. U. Ath

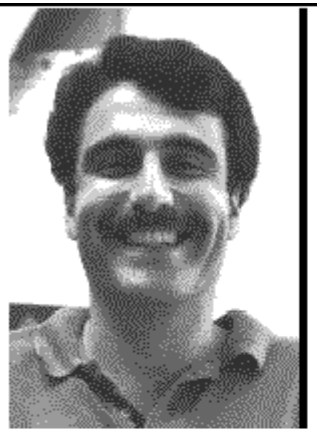
Research Interests:

- Query by content in multimedia databases;
- Fractals for clustering and spatial access methods;
- Data mining;

Redundantly Sufficient Features

Professor Faloutsos

my advisor



U.S. mail address:
Department of Computer Science
University of Maryland
College Park, MD 20742
(97-99: on leave at CMU)
Office: 3227 A.V. Williams Bldg.
Phone: (301) 405-2695
Fax: (301) 405-6707
Email: christos@cs.umd.edu

Christos Faloutsos

Current Position: Assoc. Professor of Computer Science. (97-98: on leave at CMU)
Join Appointment: Institute for Systems Research (ISR).
Academic Degrees: Ph.D. and M.Sc. (University of Toronto.); B.Sc. (Nat. Tech. U. Ath

Research Interests:

- Query by content in multimedia databases;
- Fractals for clustering and spatial access methods;
- Data mining;

CoTraining Algorithm #1

[Blum&Mitchell, 1998]

Given: labeled data L,

unlabeled data U

Loop:

Train g1 (hyperlink classifier) using L

Train g2 (page classifier) using L

Allow g1 to label *p* positive, *n* negative examps from U

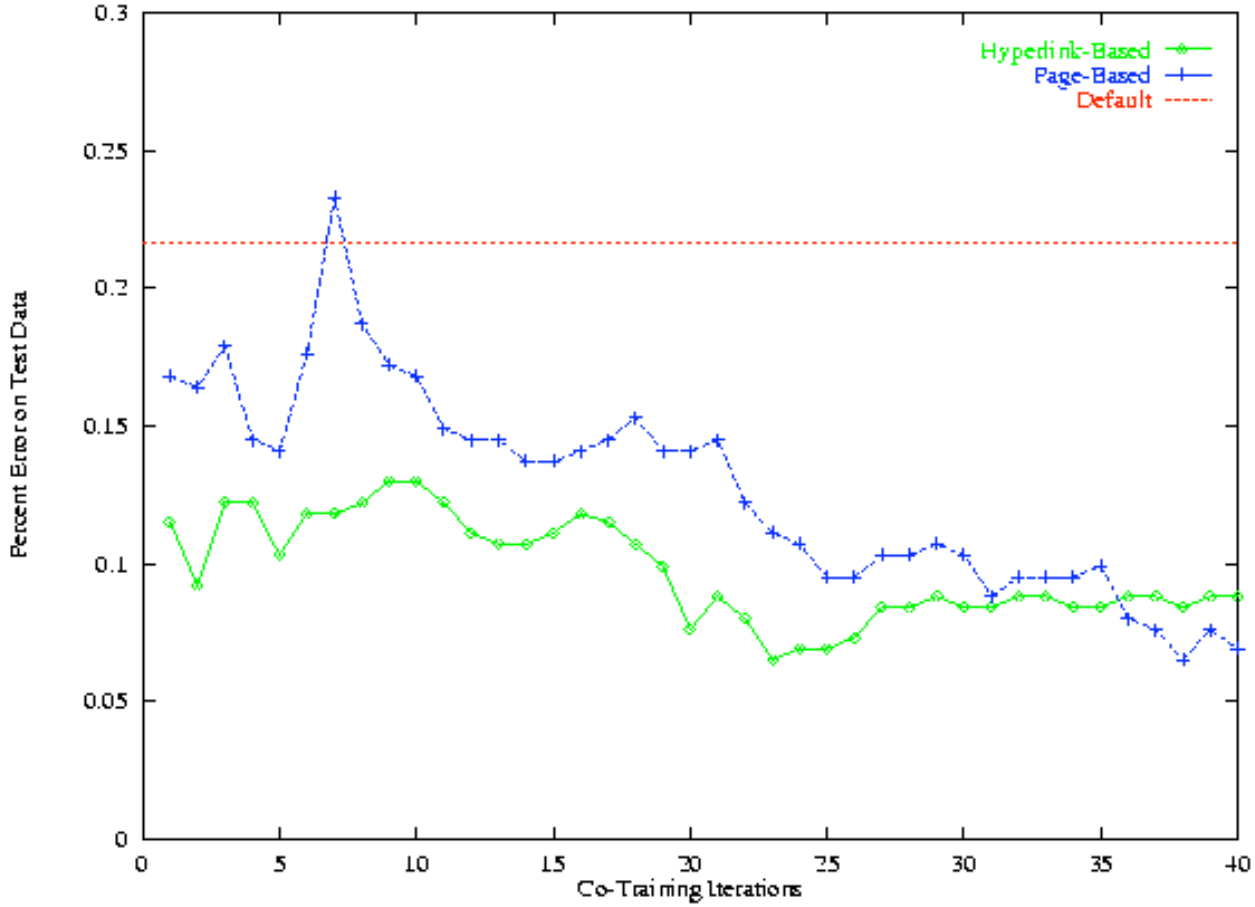
Allow g2 to label *p* positive, *n* negative examps from U

Add these self-labeled examples to L

CoTraining: Experimental Results

- begin with 12 labeled web pages (academic course)
- provide 1,000 additional unlabeled web pages
- average error: learning from labeled data 11.1%;
- average error: cotraining 5.0%

Typical run:



CoTraining setting:

- wish to learn f: X → Y, given L and U drawn from P(X)
- features describing X can be partitioned (X = X1 x X2)

  such that f can be computed from either X1 or X2

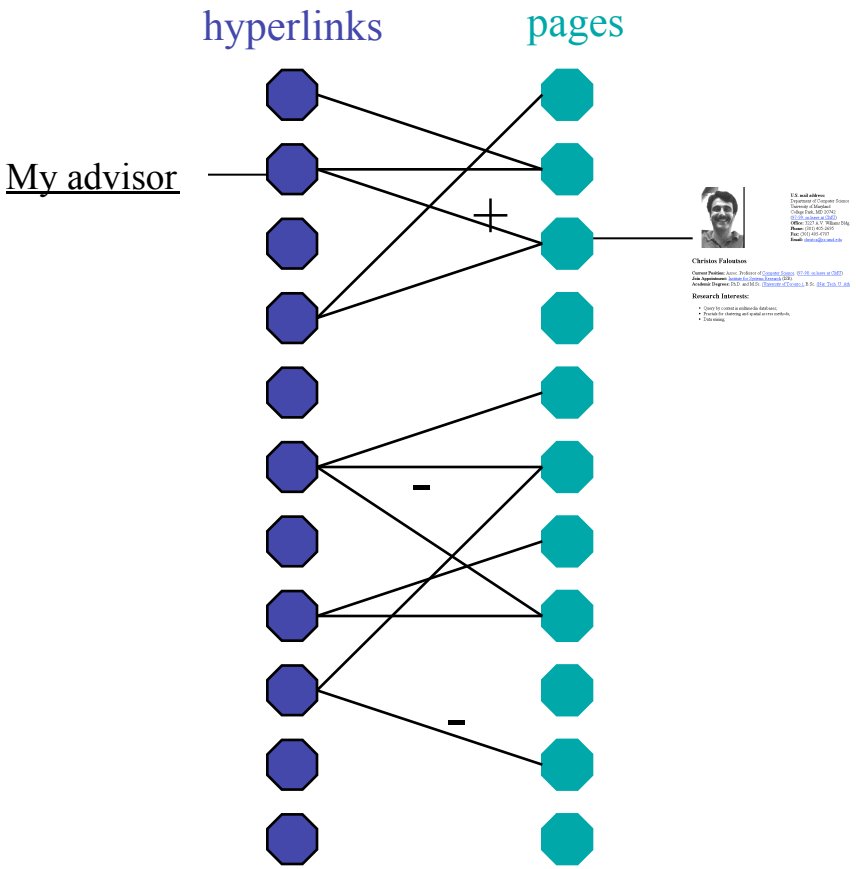
$$(\exists g_1, g_2)(\forall x \in X) \quad g_1(x_1) = f(x) = g_2(x_2)$$

One result [Blum&Mitchell 1998]:

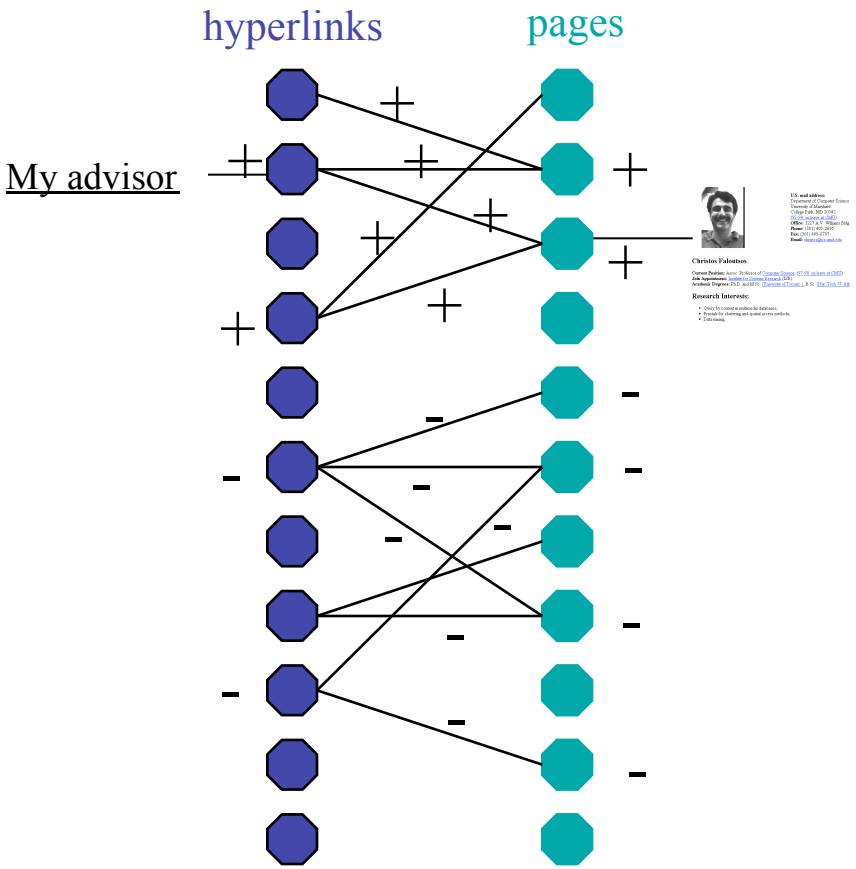
- If
  - X1 and X2 are conditionally independent given Y
  - f is PAC learnable from noisy *labeled* data
- Then
  - f is PAC learnable from weak initial classifier plus a polynomial number of *unlabeled* examples

Classifier with accuracy > 0.5

Example: Co-Training Rote Learners
f1:hyperlink → Y, f2: page → Y



Example: Co-Training Rote Learner



Questions

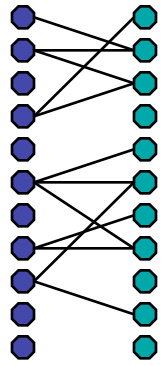
- Draw a best-case bipartite graph
- Draw a worst-case bipartite graph
  - consistent with co-training assumptions
  - inconsistent with co-training assumption

- How does classifier accuracy depend on
  - number of labeled examples?
  - number of unlabeled examples?

Expected Rote CoTraining error given *m* examples

$$CoTraining \quad setting:$$
$$learn \quad f: X \rightarrow Y$$
$$where \quad X = X_1 \times X_2$$
$$where \quad x \quad drawn \quad from \quad unknown \quad distribution$$
$$and \quad \exists g_1, g_2 \quad (\forall x) g_1(x_1) = g_2(x_2) = f(x)$$

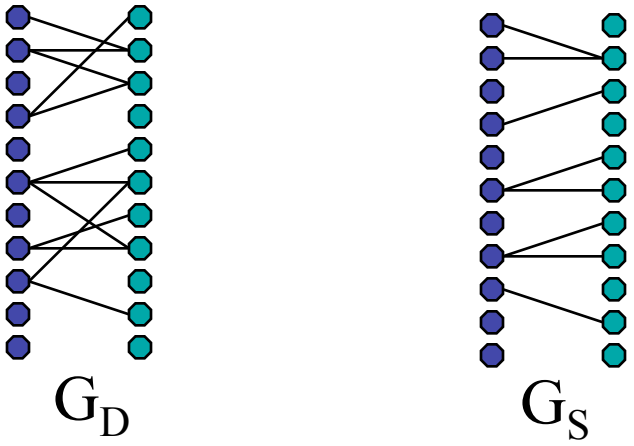
$$E[error] \leq \sum_j P(x \in g_j)(1 - P(x \in g_j))^m$$



Where g_j is the jth connected component of graph of L+U, m is number of labeled examples

How many *unlabeled* examples suffice?

Want to assure that connected components in the underlying distribution, G_D, are connected components in the observed sample, G_S



$O(\log(N)/\alpha)$ examples assure that with high probability, G_S has same connected components as G_D [Karger, 94]

N is size of G_D, α is min cut over all connected components of G_D

PAC Generalization Bounds on CoTraining

[Dasgupta et al., NIPS 2001]

This theorem assumes X1 and X2 are conditionally independent given Y

Theorem 1 *With probability at least* $1-\delta$ *over the choice of the sample* S*, we have that for all* h_1 *and* h_2*, if* $\gamma_i(h_1,h_2,\delta) > 0$ *for* $1 \leq i \leq k$ *then (a)* f *is a permutation and (b) for all* $1 \leq i \leq k$*,*

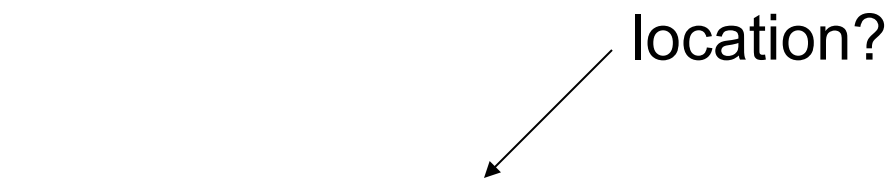
$$P(h_1 \neq i \mid f(y) = i, h_1 \neq \perp) \leq \frac{\widehat{P}(h_1 \neq i \mid h_2 = i, h_1 \neq \perp) + \epsilon_i(h_1,h_2,\delta)}{\gamma_i(h_1,h_2,\delta)}.$$

The theorem states, in essence, that if the sample size is large, and h_1 and h_2 largely agree on the unlabeled data, then $\widehat{P}(h_1 \neq i \mid h_2 = i, h_1 \neq \perp)$ is a good estimate of the error rate $P(h_1 \neq i \mid f(y) = i, h_1 \neq \perp)$.

$$\gamma_i(h_1,h_2,\delta) = \widehat{P}(h_1 = i \mid h_2 = i, h_1 \neq \perp) - \widehat{P}(h_1 \neq i \mid h_2 = i, h_1 \neq \perp) - 2\epsilon_i(h_1,h_2,\delta)$$

$$\epsilon_i(h_1,h_2,\delta) = \sqrt{\frac{(\ln 2)(|h_1|+|h_2|) + \ln\frac{2k}{\delta}}{2|S(h_2 = i,\ h_1 \neq \perp)|}}$$

Example 2: Learning to extract named entities

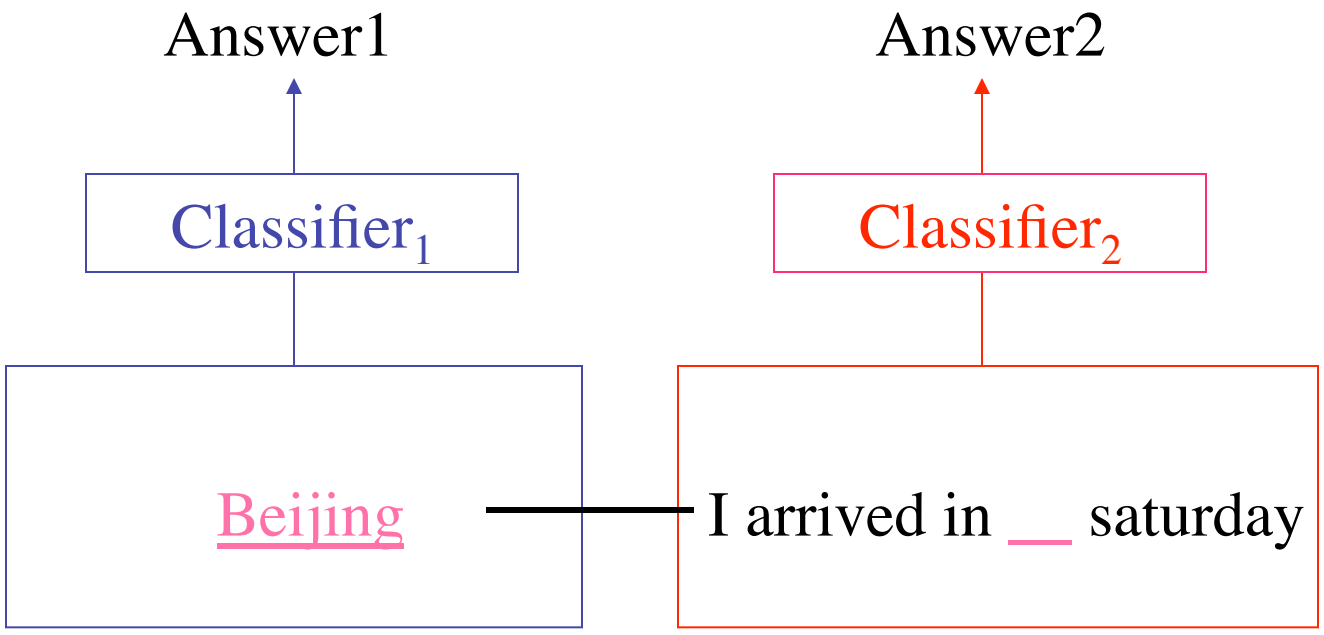


If: "I arrived in <X> on Saturday."

Then: Location(X)

Co-Training for Named Entity Extraction (i.e.,classifying which strings refer to people, places, dates, etc.)

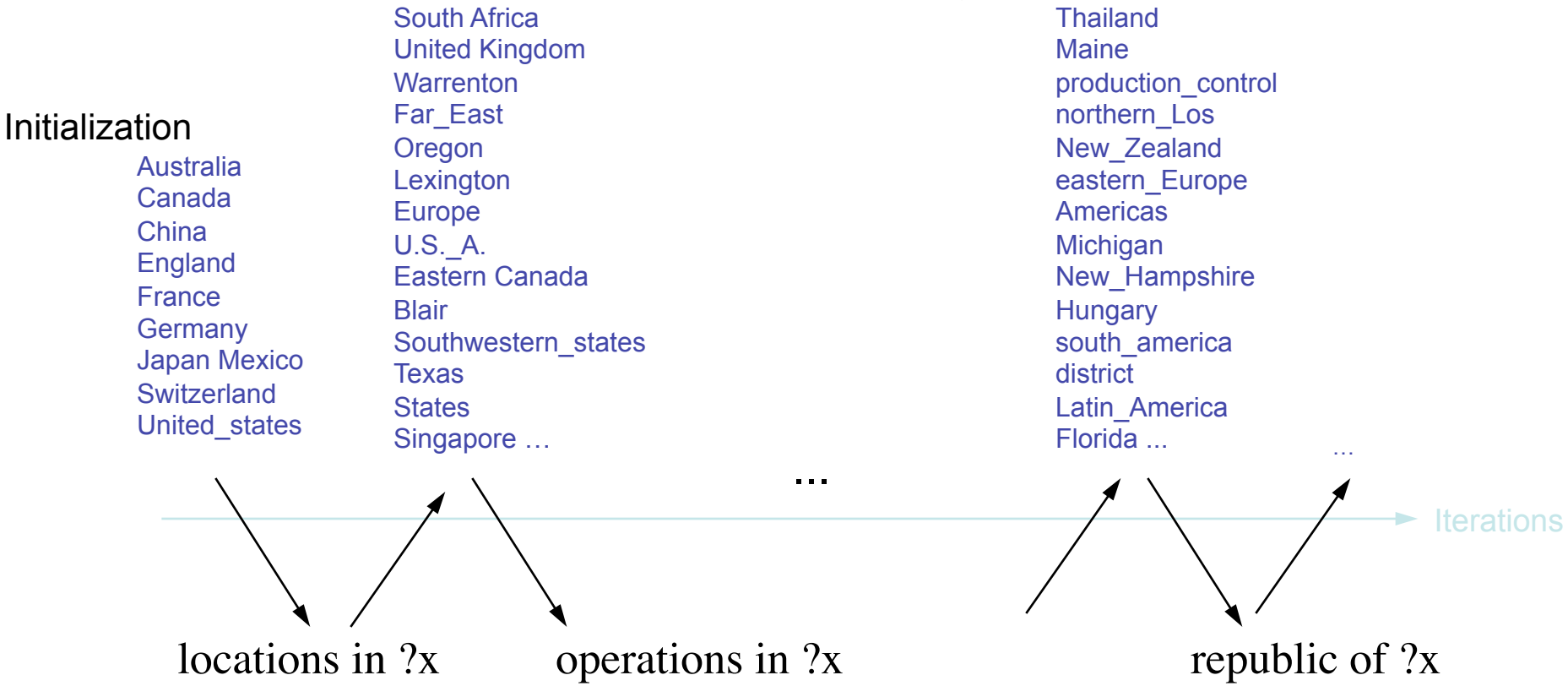
[Riloff&Jones 98; Collins et al., 98; Jones 05]



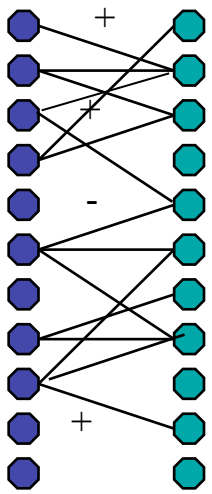
I arrived in **Beijing** saturday.

Bootstrap learning to extract named entities

[Riloff and Jones, 1999], [Collins and Singer, 1999], ...



What if CoTraining Assumption Not Perfectly Satisfied?



- Idea: Want classifiers that produce a *maximally consistent* labeling of the data
- If learning is an optimization problem, what function should we optimize?

Co-EM [Nigam & Ghani, 2000; Jones 2005]

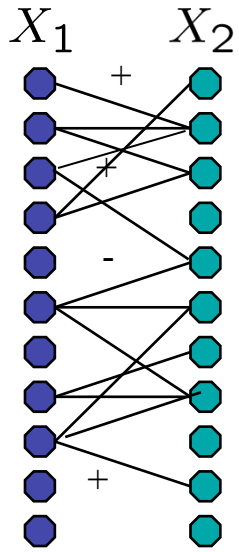
Idea:

- Like co-training, train two coupled functions
  - P(class | X1), P(class | X2)
- Like EM, iterative probabilistic algorithm
  - Assign probabilistic values to unobserved class labels
  - Updating model parameters (= labels of other feature set)

Goal to learn $X_1 \rightarrow Y,\ X_2 \rightarrow Y,\ X_1 \times X_2 \rightarrow Y$

$$P(Y|X_1 = k) = \sum_j P(Y|X_2 = j)P(X_2 = j|X_1 = k)$$

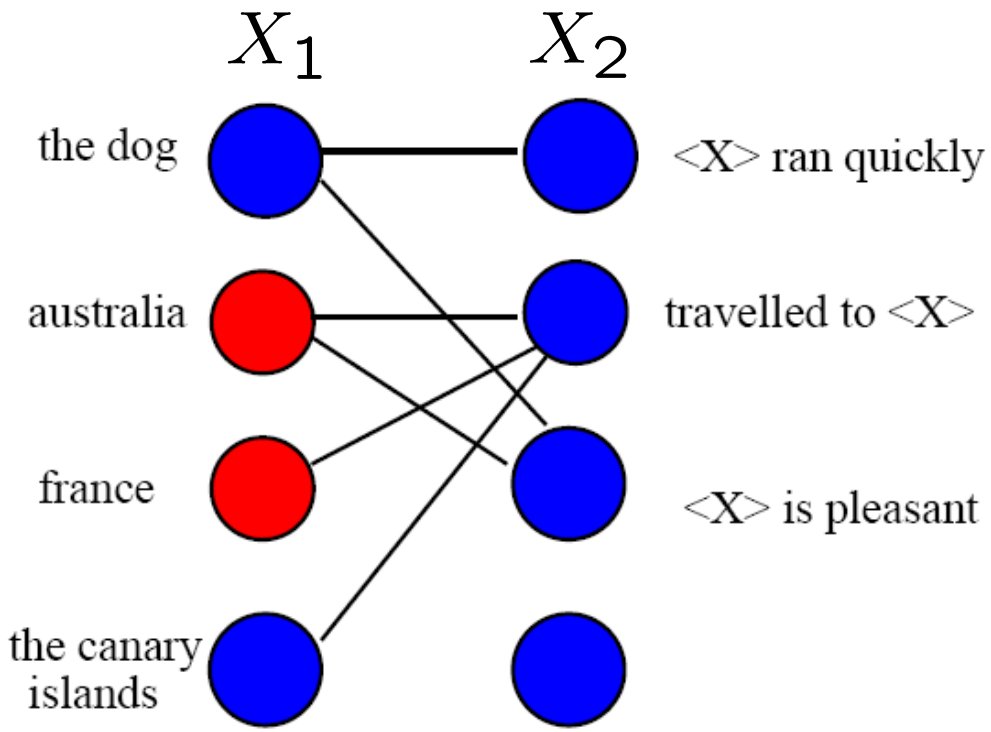
$$P(Y|X_2 = j) = \sum_k P(Y|X_1 = k)P(X_1 = k|X_2 = j)$$



CoEM applied to Named Entity Recognition

[Rosie Jones, 2005], [Ghani & Nigam, 2000]

$X_1 \times X_2 \rightarrow Y$



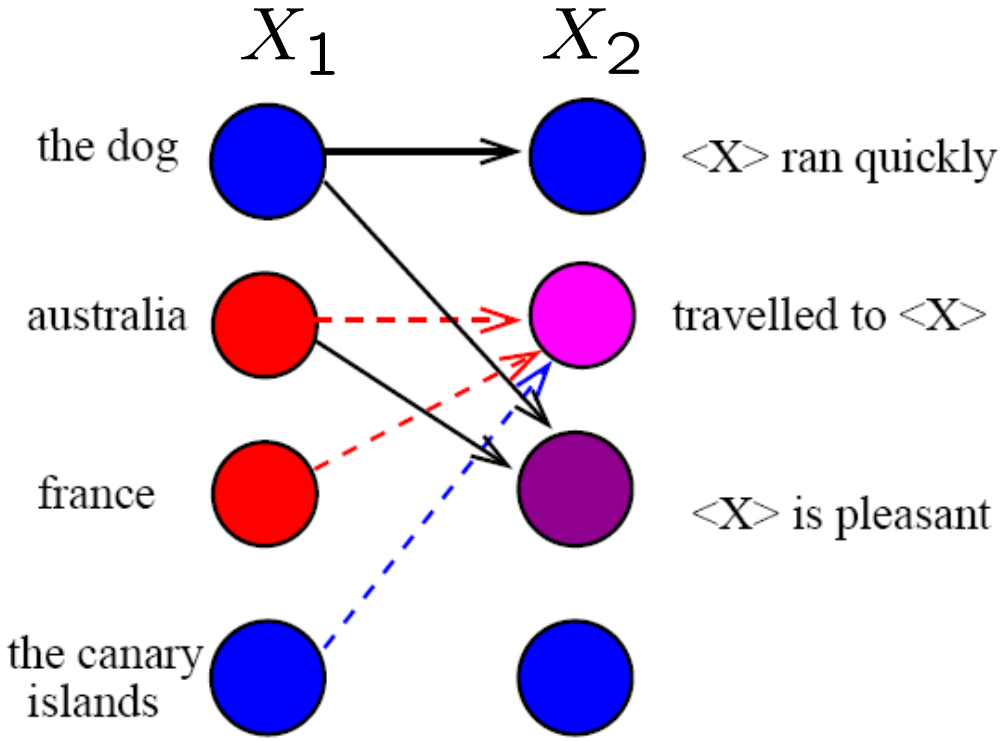
Update rules:

$$P(Y|X_1 = k) = \sum_j P(Y|X_2 = j)P(X_2 = j|X_1 = k)$$

$$P(Y|X_2 = j) = \sum_k P(Y|X_1 = k)P(X_1 = k|X_2 = j)$$

CoEM applied to Named Entity Recognition

[Rosie Jones, 2005], [Ghani & Nigam, 2000]



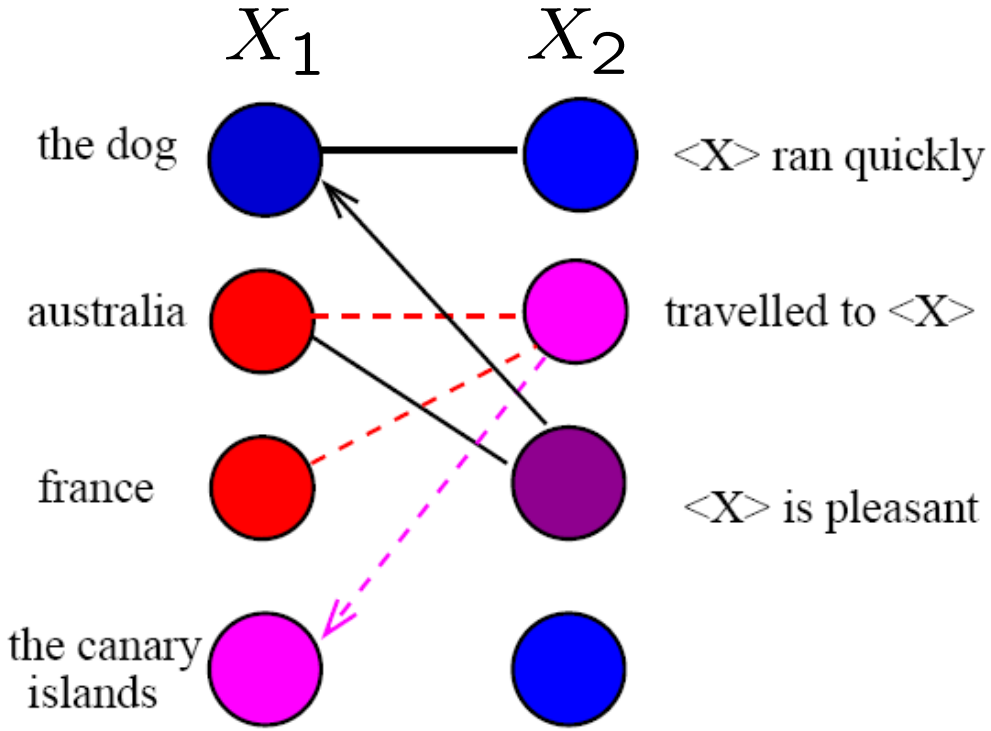
Update rules:

$$P(Y|X_1 = k) = \sum_j P(Y|X_2 = j)P(X_2 = j|X_1 = k)$$

$$P(Y|X_2 = j) = \sum_k P(Y|X_1 = k)P(X_1 = k|X_2 = j)$$

CoEM applied to Named Entity Recognition

[Rosie Jones, 2005], [Ghani & Nigam, 2000]



Update rules:

$$P(Y|X_1 = k) = \sum_j P(Y|X_2 = j)P(X_2 = j|X_1 = k)$$

$$P(Y|X_2 = j) = \sum_k P(Y|X_1 = k)P(X_1 = k|X_2 = j)$$

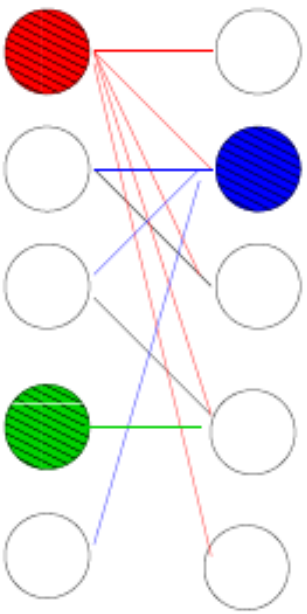
Co-EM Applied to our NPxContext data

[work by Weam AbuZaki]

- Train boolean classifiers for categories
  - organization ../CoEM/category_organization.coem.xls
  - company ../CoEM/category_company.coem.xls
  - person ../CoEM/category_person.coem.xls
  - …
- ~50 seed examples taken from RTW KB
- high accuracies for many categories

- macro-reading
- how would you do micro-reading?

Some nodes are more important than others [Jones, 2005]



Can use this for active learning...

Noun-phrase	Outdegree
you	1656
we	1479
it	1173
company	1043
this	635
all	520
they	500
information	448
us	367
any	339
products	332
i	319
site	314
one	311
1996	282
he	269
customers	269
these	263
them	263
time	234

Context	Outdegree
<x> including	683
including <x>	612
<x> provides	565
provides <x>	565
provide <x>	390
<x> include	389
include <x>	375
<x> provide	364
one of <x>	354
<x> made	345
<x> offers	338
offers <x>	320
<x> said	287
<x> used	283
includes <x>	279
to provide <x>	266
use <x>	263
like <x>	260
variety of <x>	252
<x> includes	250

CoTraining Summary

- Unlabeled data improves supervised learning when example features are redundantly sufficient
  - Family of algorithms that train multiple classifiers

- Theoretical results
  - Expected error for rote learning
  - If X1,X2 conditionally independent given Y, Then
    - PAC learnable from weak initial classifier plus unlabeled data
    - disagreement between g1(x1) and g2(x2) bounds final classifier error

- Many real-world problems of this type
  - Semantic lexicon generation [Riloff, Jones 99], [Collins, Singer 99]
  - Web page classification [Blum, Mitchell 98]
  - Word sense disambiguation [Yarowsky 95]
  - Speech recognition [de Sa, Ballard 98]
  - Visual classification of cars [Levin, Viola, Freund 03]

Coupled training type 2

Wish to learn f1: X → Y1, f2: X → Y2,

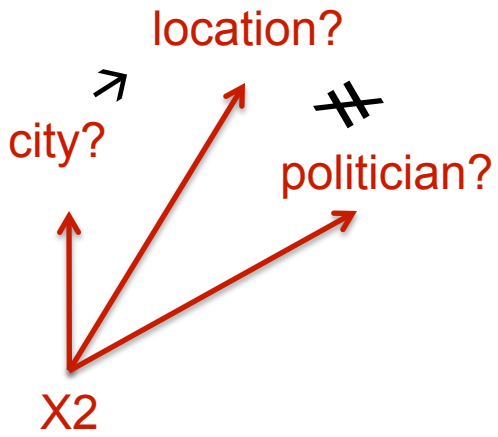
such that: (∀ x) g(f1(x), f2(x))

e.g.

location: NounPhraseInSentence → {0,1}

politician: NounPhraseInSentence → {0,1}

g(y1,y2) = not (and(y1,y2))



Luke is mayor of Pittsburgh.

Coupling functions with different outputs

[Daume, 2008]

Wish to learn f1: X → Y1, f2: X → Y2,

such that: (∀x) g(f1(x), f2(x))

Key theoretical question: what is sample complexity? How does it depend on g?

Key insight:

- g will be most useful if the probability that it is satisfied by random y1, y2 is low

Coupling functions with different outputs

[Daume, 2008]

Consider simpler one-sided learning of f2, given we know f1

1: Learn h_2 directly on D
2: For each example $(x, y_1) \in D^{\text{unlab}}$
3: Compute $y_2 = h_2(x)$
4: If $\chi(y_1, y_2)$, add (x, y_2) to D
5: Relearn h_2 on the (augmented) D
6: Go to (2) if desired

Definition 4. *We say the* discrimination *of* χ *for* h^0 *is* $\Pr_{\mathcal{D}}[\chi(f_1(x), h^0(x))]^{-1}$.

Coupling functions with different outputs

[Daume, 2008]

Theorem 1. *Suppose C_2 is PAC-learnable with noise in the structured setting, h_2^0 is a weakly useful predictor of f_2, and χ is correct with respect to $\mathcal{D}, f_1, f_2, h_2^0$, and has discrimination $\geq 2(|\mathcal{Y}| - 1)$. Then C_2 is also PAC-learnable with one-sided hints.*

(here |Y|= |Y1| x |Y2| is the number of values the two functions can take on)

Further Reading

- Semi-Supervised Learning, O. Chapelle, B. Sholkopf, and A. Zien (eds.), MIT Press, 2006. (excellent book)

- Semi-Supervised Learning for Computational Linguistics, S. Abney, Springer, 2007. (pretty good, pretty basic)

- EM for Naïve Bayes classifiers: K.Nigam, et al., 2000. "Text Classification from Labeled and Unlabeled Documents using EM", *Machine Learning,* 39, pp.103—134.

- CoTraining: A. Blum and T. Mitchell, 1998. “Combining Labeled and Unlabeled Data with Co-Training,” *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT-98).*

- S. Dasgupta, et al., “PAC Generalization Bounds for Co-training”, *NIPS 2001*

- Model selection: D. Schuurmans and F. Southey, 2002. “Metric-Based methods for Adaptive Model Selection and Regularization,” Machine Learning, 48, 51—84.