# Co-Training

# When can Unlabeled Data Help Learn f: X→Y ?

Consider problem 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:

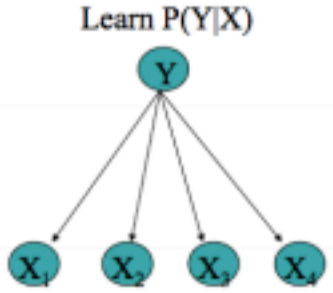
- i.i.d. labeled examples $L = \{\langle x_1, y_1\rangle \ldots \langle x_m, y_m\rangle\}$
- i.i.d. unlabeled examples $U = \{x_{m+1}, \ldots x_{m+n}\}$

Wish to find hypothesis with lowest true error:

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

# When can Unlabeled Data Help Learn f: X→Y ?

- EM




| Y | X1 | X2 | X3 | X4 |
|---|---|---|---|---|
| 1 | 0 | 0 | 1 | 1 |
| 0 | 0 | 1 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 |
| ? | 0 | 1 | 1 | 0 |
| ? | 0 | 1 | 0 | 1 |

- Metric regularization
  - [Schuurmans & Southey, MLJ 2002]
  - use unlabeled data to detect (and avoid) overfitting

- CoTraining, Multiview learning, CoRegularization

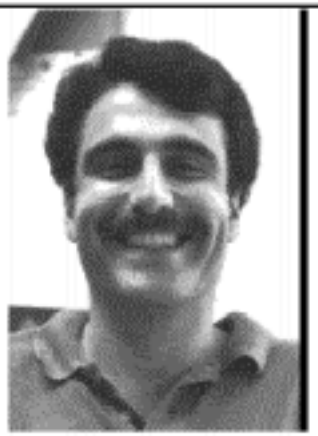
# CoTraining

- 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

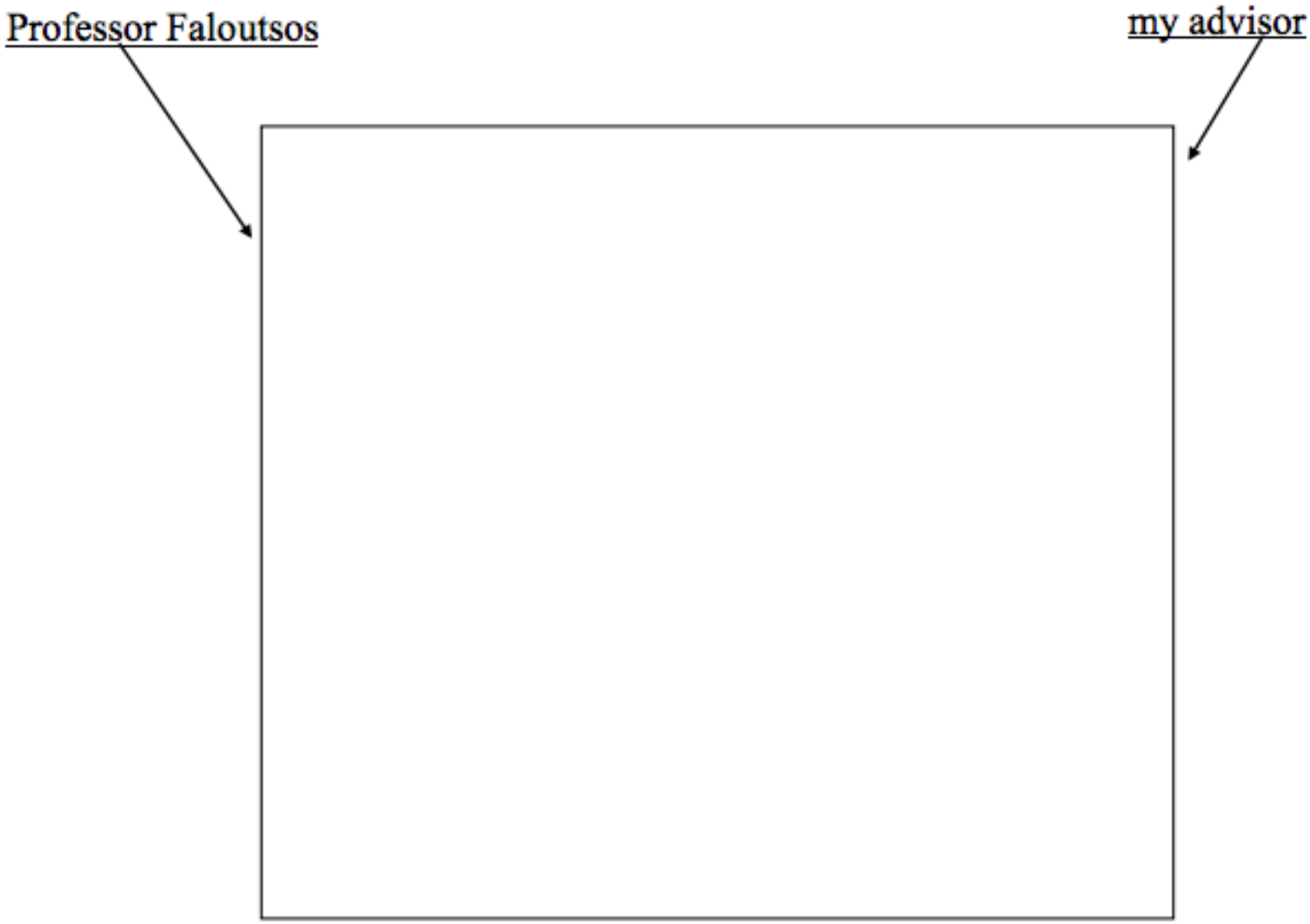
## Christos Faloutsos

**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

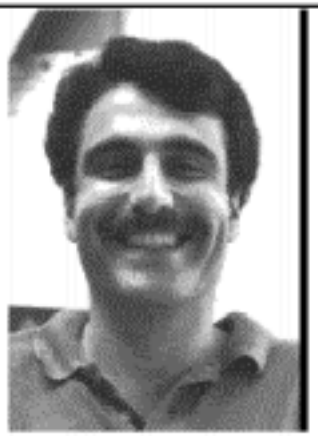
## Research Interests:

- 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

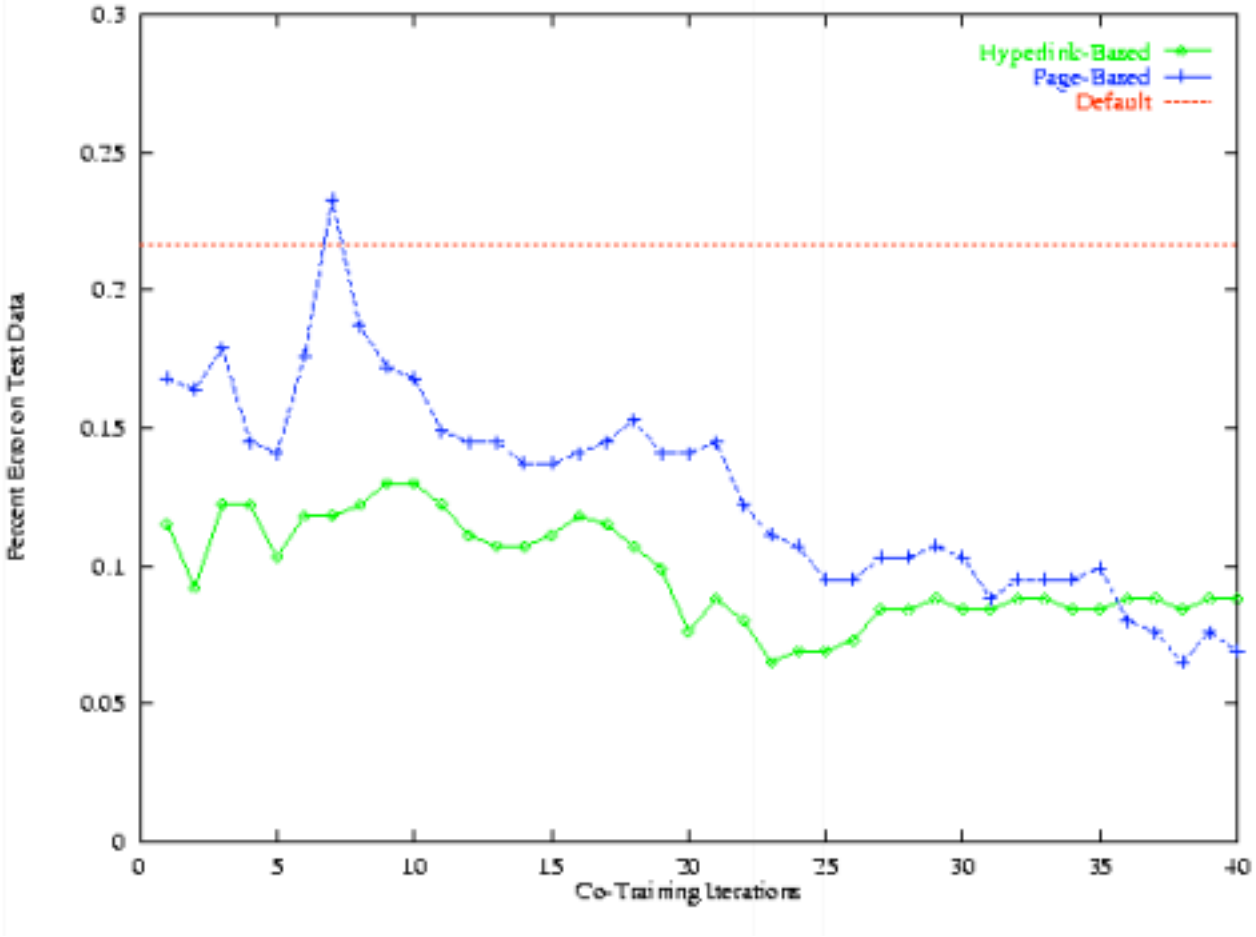
Allow g2 to label *p* positive, *n* negative examps from U

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 polynomial number of *unlabeled* examples

Classifier with accuracy > 0.5

# Can Unlabeled Data Help Estimate True Error?

Consider two functions making *independent errors*

P(disagree)=P(g1 right, g2 wrong) + P(g2 right, g1 wrong)

e.g., If true error of g1 is 0.1, true error of g2 is 0.1, what is P(disagree?)

# 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 $P(h_1 \neq i \mid f(y) = i, h_1 \neq \perp)$.

# 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 $P(h_1 \neq i \mid f(y) = i, h_1 \neq \perp)$.

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

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

# Co Regularization

- Let's build our assumption that g1 and g2 must agree directly into the objective we're optimizing

- e.g.,

$$\langle \theta_1, \theta_2 \rangle \leftarrow \arg \min_{\langle \theta_1, \theta_2 \rangle} \sum_{x^l \in L} (y^l - g_1(x^l; \theta_1))^2$$

$$+ \sum_{x^l \in L} (y^l - g_2(x^l; \theta_2))^2$$

$$+ \sum_{x^u \in U} (g_1(x^u; \theta_1) - g2(x^u; \theta_2))^2$$

# CoTraining Summary

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

- Theoretical results
  - 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]

# Further Reading

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

- 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.