Semi-supervised Learning with Gaussian Random Fields

Jerry Xiaojin Zhu

Joint work with:
Zoubin Ghahramani
John Lafferty

February 16, 2004

Traditional Types of Learning

Supervised (classification, regression, etc.)

Unsupervised (clustering etc.)

usage	supervised learning	unsupervised learning
$\{(x, y)\}$ labeled data	yes	no
$\{x\}$ unlabeled data	no	yes

For Classification

Labeled data are often hard, expensive, slow to obtain.

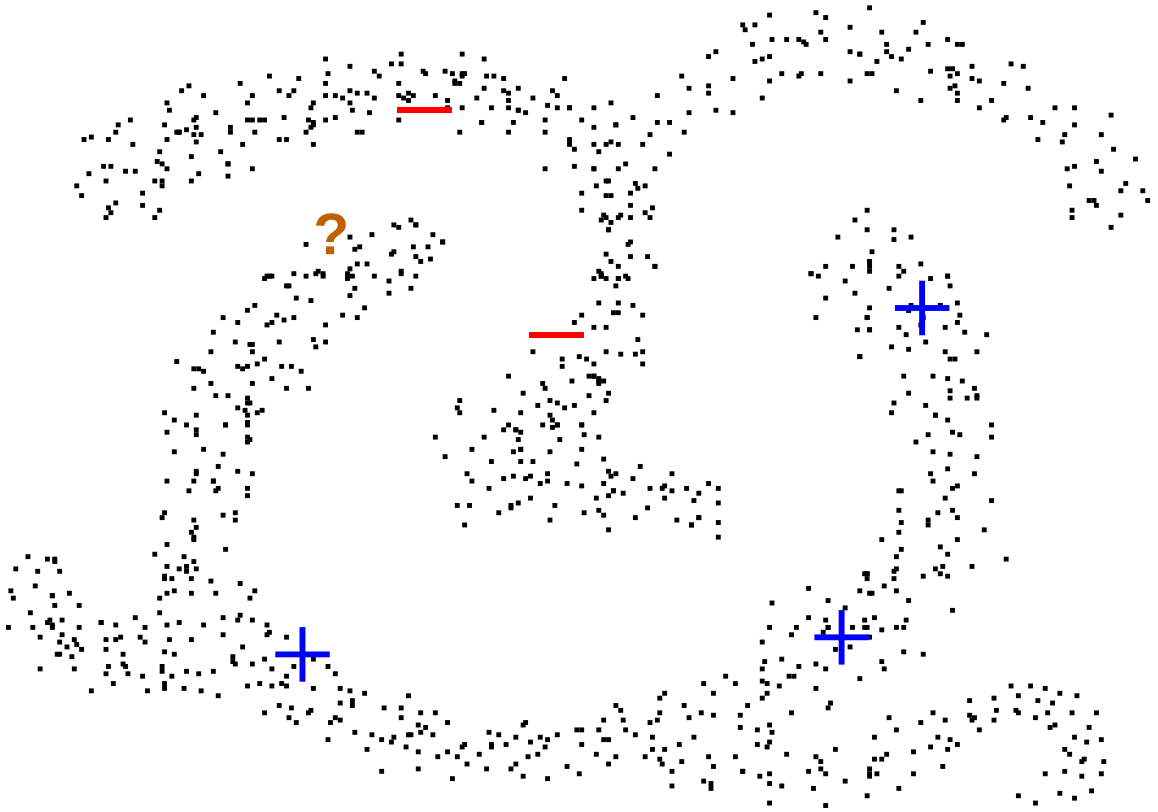
Unlabeled data are often easy to obtain, a lot.

- Speech recognition: slow to transcribe
- Text categorization: limited user patience
- Protein structure: months of X-ray crystallography

Semi-supervised Learning

usage	supervised learning	semi-supervised learning	unsupervised learning
$\{(x, y)\}$ labeled data	yes	yes	no
$\{x\}$ unlabeled data	no	yes	yes

How is it possible? Intuition

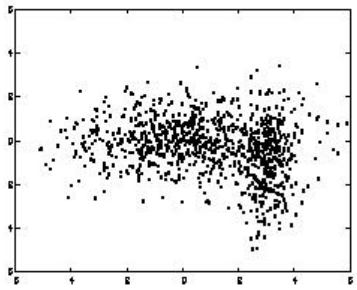


Some Semi-supervised Learning Methods

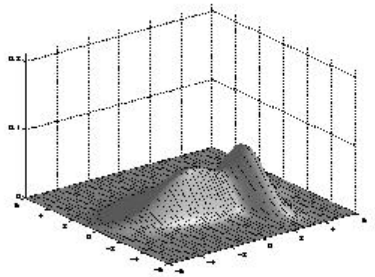
- mixture models / EM
- transductive SVM
- co-training
- graph methods $\leftarrow$ this talk

Mixture Models. Cluster then label

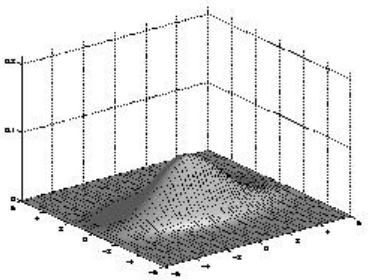
Method
mixture
trans. SVM
co-training
graph



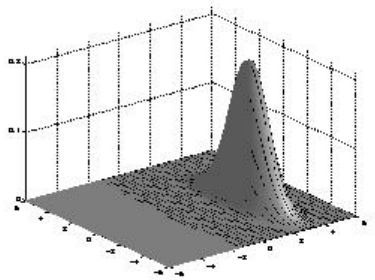
(a) Unlabeled data U



(b) Density $p(x)$ from infinite U



(c) One Gaussian component



(d) The other Gaussian component

[Castelli & Cover 95], [Ratsaby & Venkatesh 95], [Nigam et al. 00]

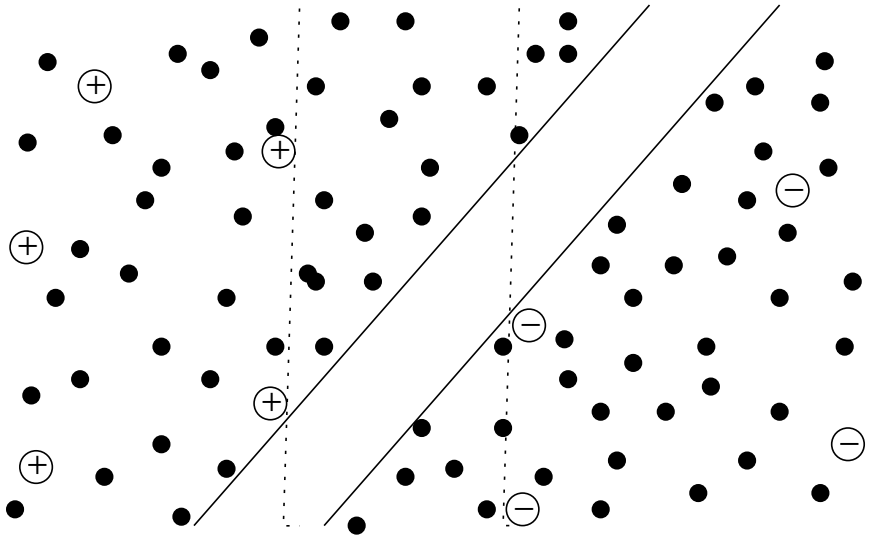
Mixture Models. Cluster then label

Method
mixture
trans. SVM
co-training
graph

- Unlabeled data may hurt, when...
  - mixture assumption wrong
  - EM stuck in bad local minima
- Cluster first, label next
  - same flavor
  - harder to analyze

Transductive SVM

Method
mixture
trans. SVM
co-training
graph

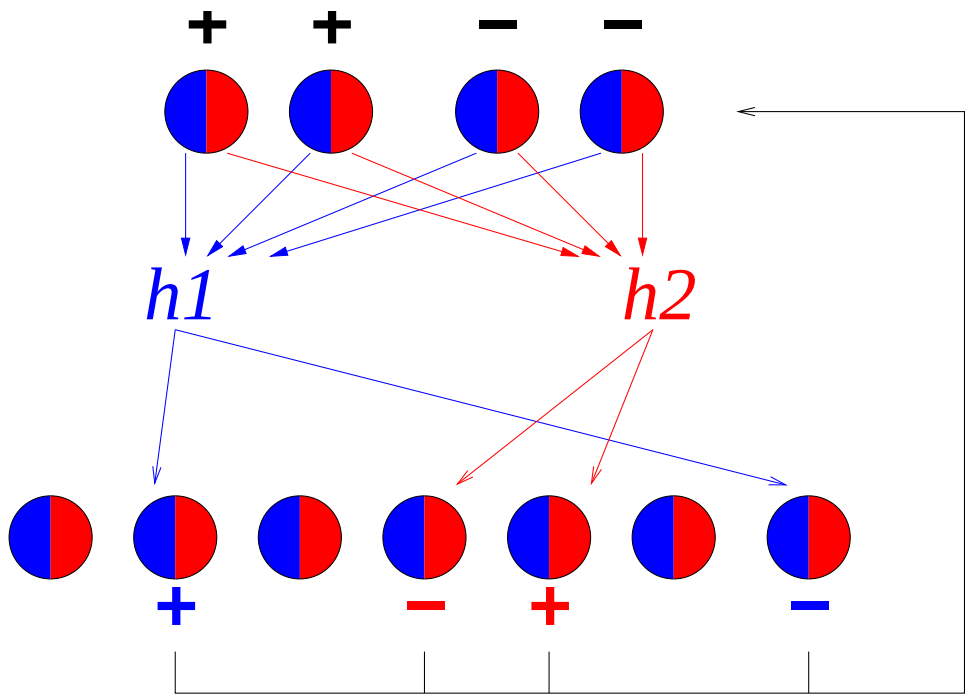


Unlabeled data steer linear boundary away from dense regions.

[Vapnik 98], [Joachims 99]

Co-training

Method
mixture
trans. SVM
co-training
graph



Unlabeled data reduce version space size by forcing h_1, h_2 to agree.

[Blum & Mitchell 98]

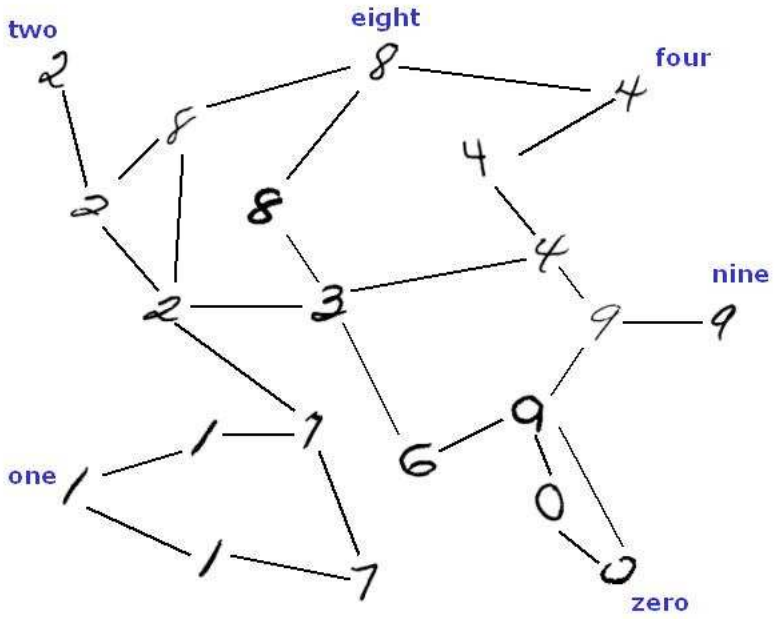
Graph Methods

Nodes: labeled and unlabeled data

Edges: "local similarity"

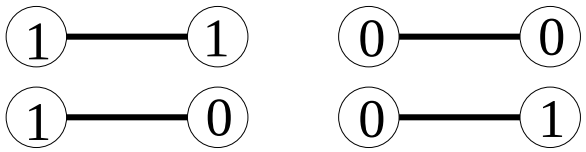
Labels 'propagate'

Method
mixture
trans. SVM
co-training
graph

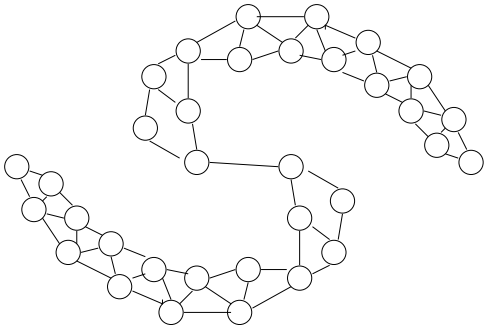


Energy

- labels: Assume binary $\mathbf{y} \in \{0,1\}^n$
- edges: weight matrix W. (Very, very important)
- energy: $E(\mathbf{y}) = \frac{1}{2}\sum_{i,j} w_{ij}\left(y_i - y_j\right)^2$

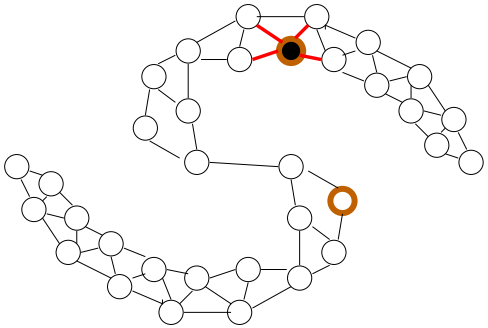


Low energy $\rightarrow$ Label Propagation

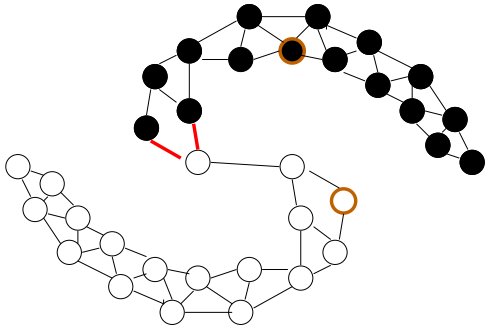


energy=0

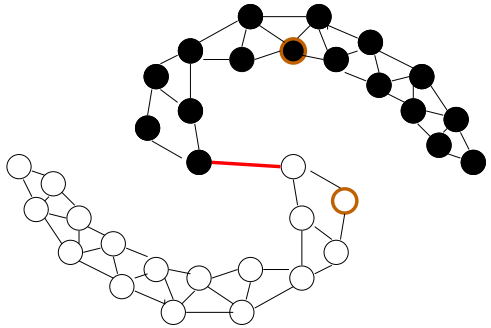
Given labeled data:



energy=4



energy=2



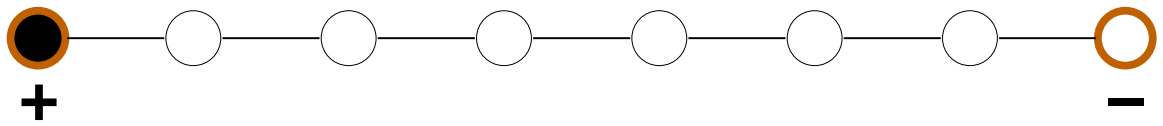
energy=1

Markov Random Fields

$$p(\mathbf{y}) \propto \exp(-E(\mathbf{y}))\ |_{\mathbf{y}_L = f_L}$$

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

Mode: equiv. graph mincut [Blum & Chawla 01]; not unique.



?

Hard to compute: Boltzmann machines, MCMC...

[Zhu & Ghahramani 02]

Gaussian Random Fields

Why hard: $y_i \in \{0, 1\}$ combinatorial optimization

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

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

[Zhu & Ghahramani & Lafferty 03]

The Laplacian

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

The Laplacian $\Delta = D - W$

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

Gaussian Random Fields

$$p(\mathbf{y}) = \exp\left(-\mathbf{y}^\top \Delta \mathbf{y}\right) \; |_{\mathbf{y}_L = f_L}$$

Conditioned on labeled data f_L:

$$\mathbf{y}_U \sim \mathcal{N}\left(f_U, \frac{1}{2}(\Delta_{UU})^{-1}\right)$$

The mean on unlabeled data:

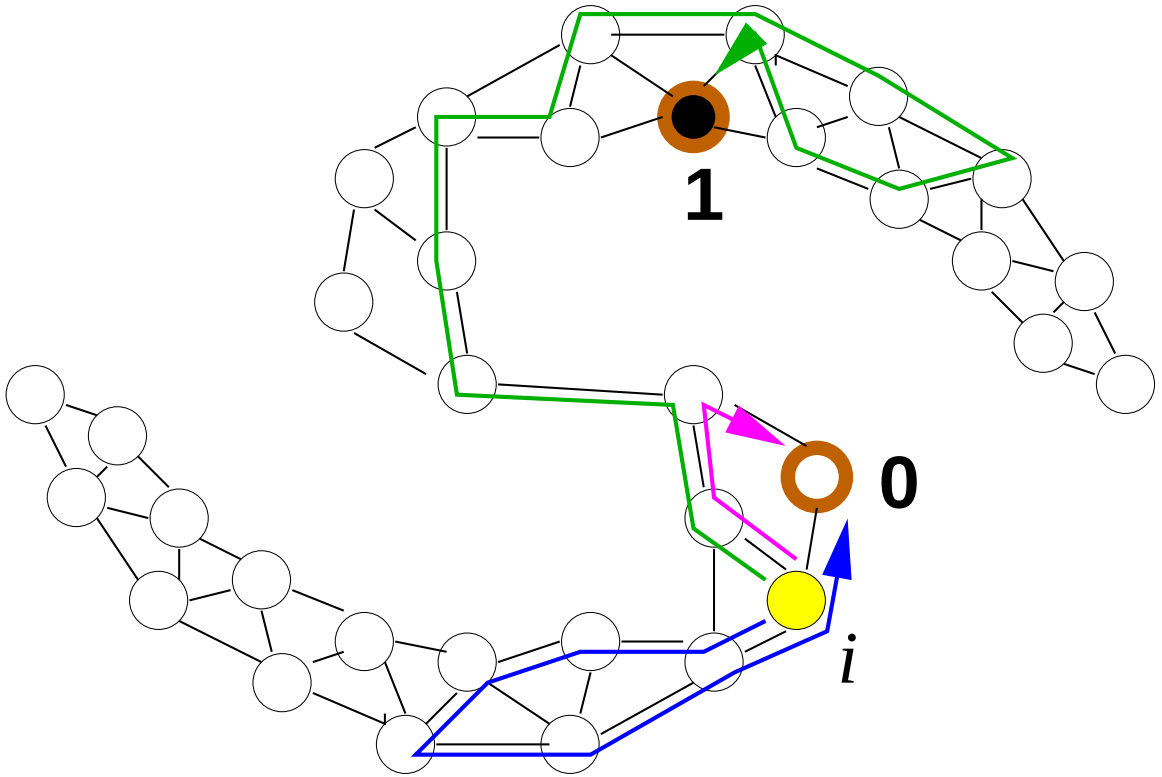
$$f_U = -(\Delta_{UU})^{-1}\Delta_{UL} f_L$$

The Mean f_U

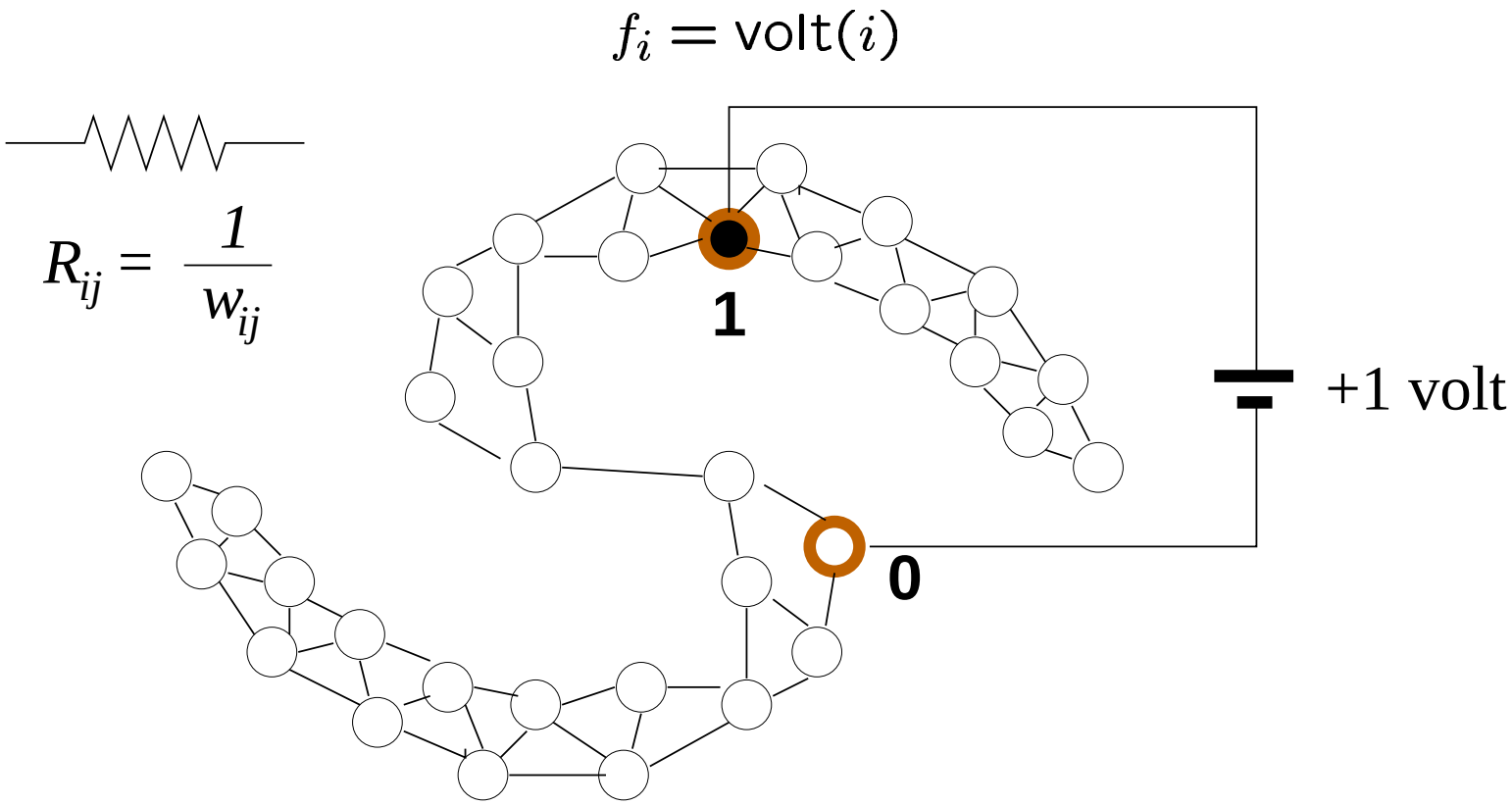
- mode of the Gaussian Random Field
- min energy state
- uniquely exists
- harmonic $\Delta \mathbf{f} = 0$ or $f_i = \frac{\sum_{j \sim i} w_{ij} f_j}{\sum_{j \sim i} w_{ij}},\ i \in U$
- $0 \leq f_i \leq 1$

f_U Interpretation: Random Walks

$f_i = P(\text{reach label } \mathbf{1} | \text{from } i)$



f_U Interpretation: Electric Networks



Classification Using Gaussian Fields

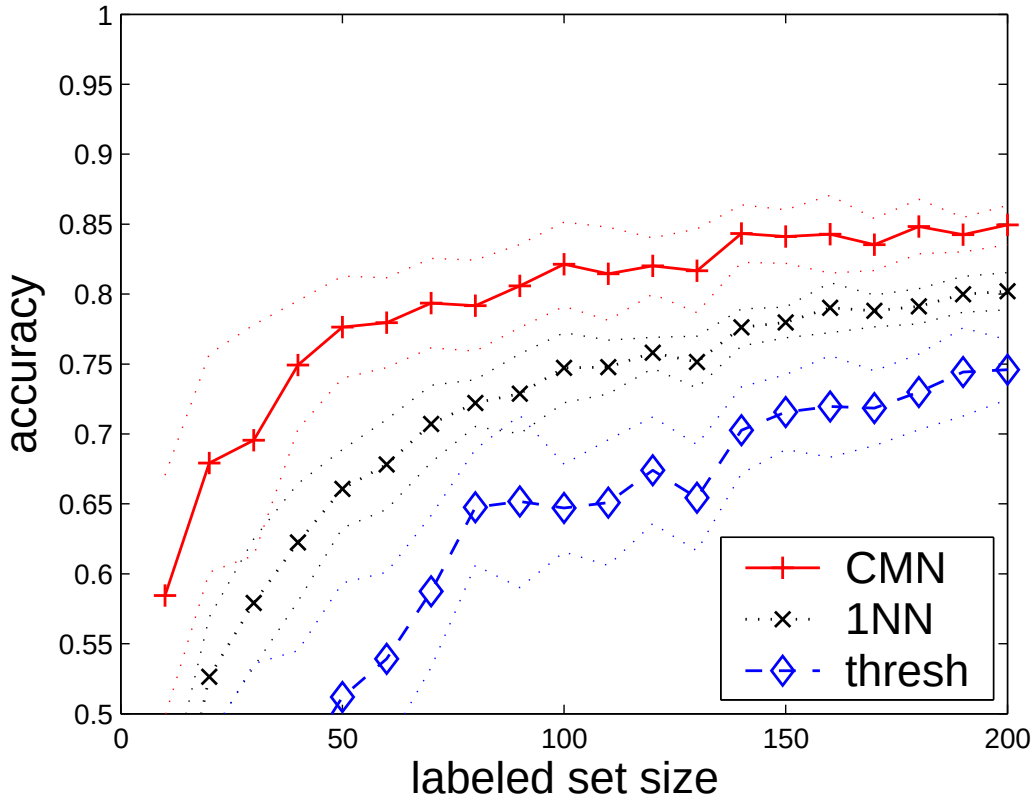
- naive: threshold f_U at 0.5

- heuristic: incorporating class proportions. E.g. prior knowledge: 90% class 1

$$\begin{aligned} \text{minimize} \quad & E(\mathbf{y}) = \mathbf{y}^\top \Delta \mathbf{y} \\ \text{subject to} \quad & y_L = f_L \\ & \frac{\sum f_U}{|U|} = 0.9 \end{aligned}$$

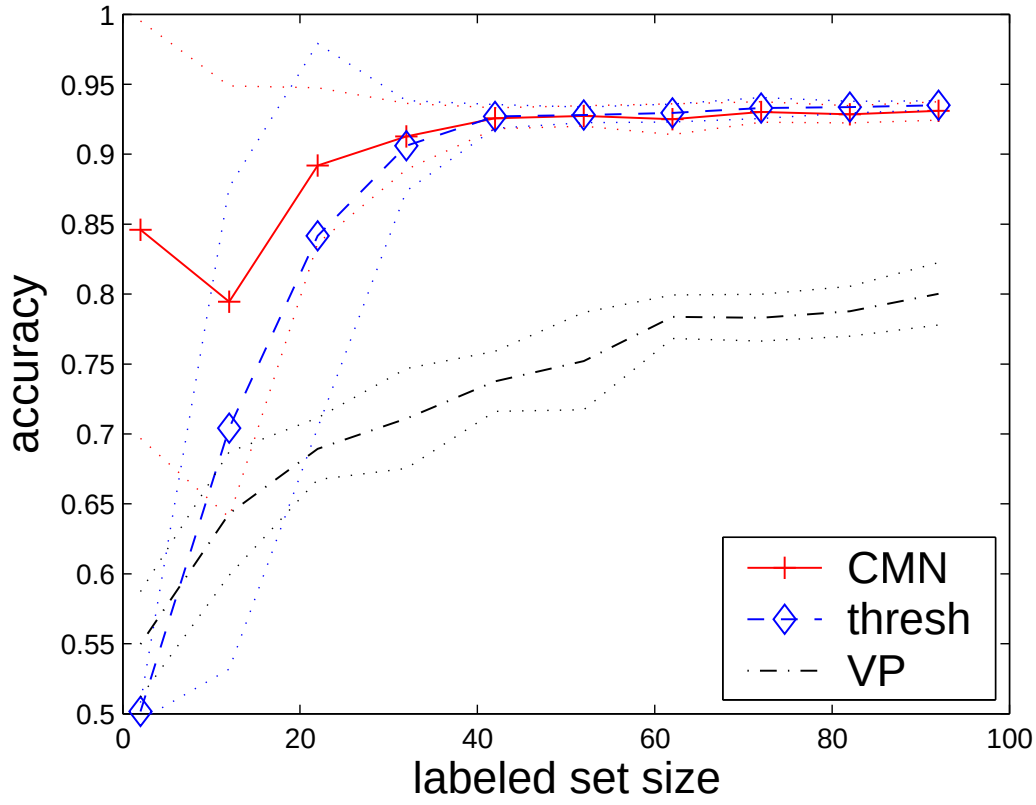
OCR Digits (0...9)

$|L \cup U| = 4000$



20 Newsgroups (PC vs. MAC)

$|PC| = 982, |MAC| = 961$



Learn the Graph

Parameterize W. For example:

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

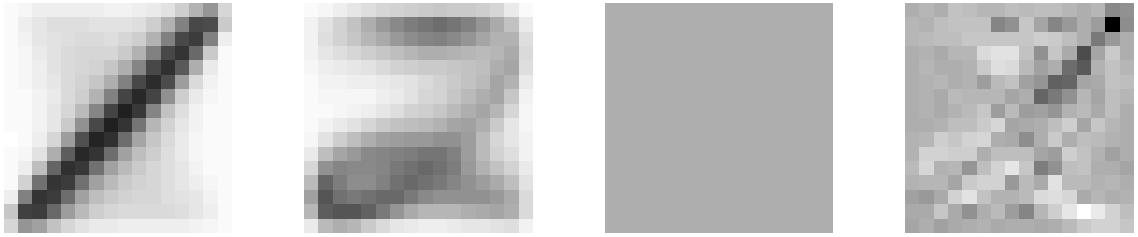
Hyperparameters $\Theta = \{\sigma_d^2 | d = 1, \ldots, m\}$: length scales.

Learn the Graph

- Minimize label entropy (maximize label confidence. Heuristic.)
- Evidence maximization (Local optima)
- Kernel alignment (research)

Learn the Graph

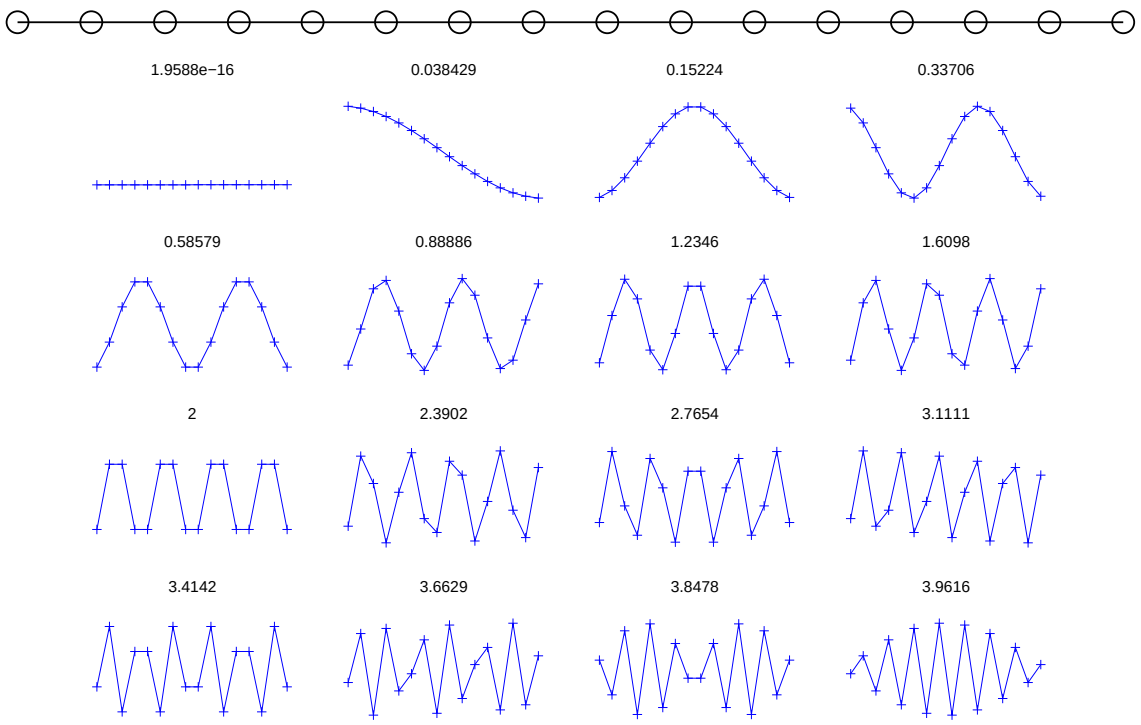
OCR Digits "1" vs. "2", minimize entropy



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

The Kernel View

Laplacian Spectrum $\Delta = \sum_i \lambda_i \phi_i \phi_i^\top$. λ_i: frequencies; ϕ_i: vibration modes. Low frequency = smoother.



The Kernel View

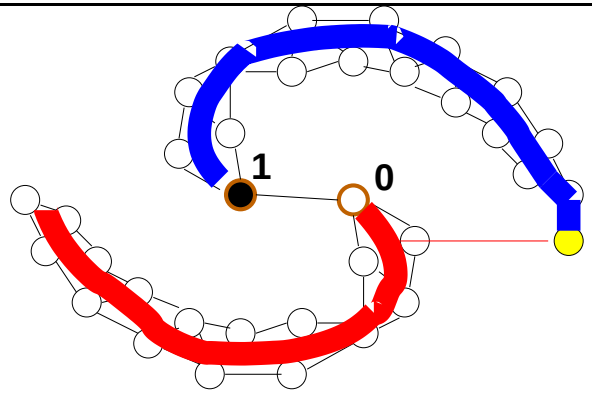
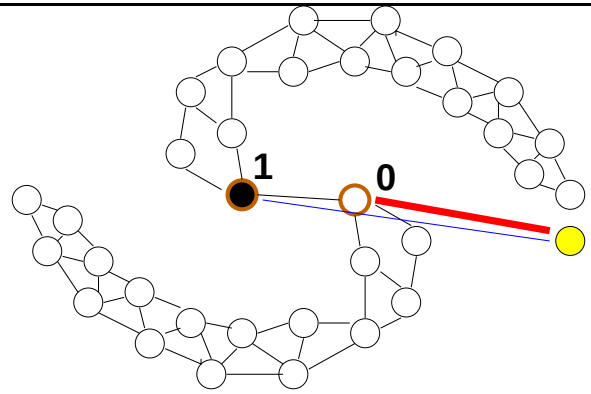
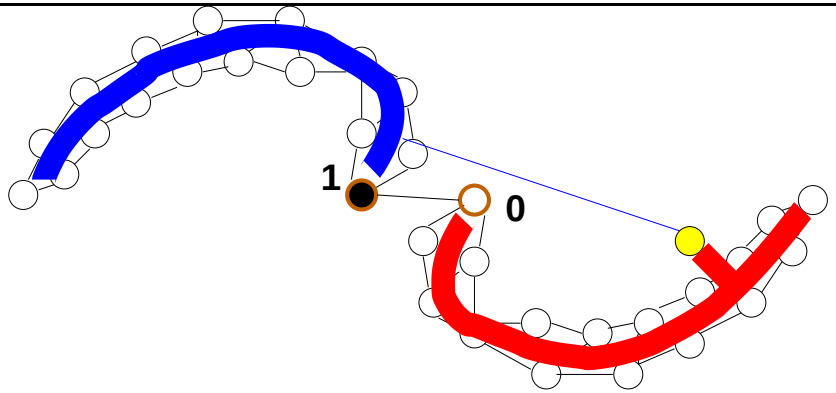
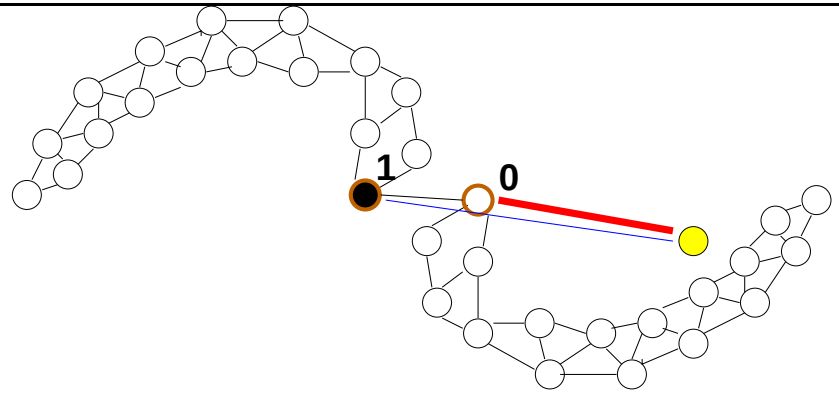
$$p(\mathbf{y}) = \exp\left(-\mathbf{y}^\top \Delta \mathbf{y}\right)$$

$$K \approx \Delta^{-1} = \sum_i \frac{1}{\lambda_i} \phi_i \phi_i^\top$$

$\frac{1}{\lambda_i}$ emphasizes smooth components.

In general, a non-negative, monotonic decreasing function on the spectrum of Δ gives a semi-supervised kernel. [Smola & Kondor 03]

Semi-supervised vs. 'standard' Kernels

Semi-supervised kernels	Linear, polynomial, RBF etc.
Use unlabeled data	Ignore unlabeled data
	
	

Active Learning

labeled data	unlabeled data
a few	A LOT

↑

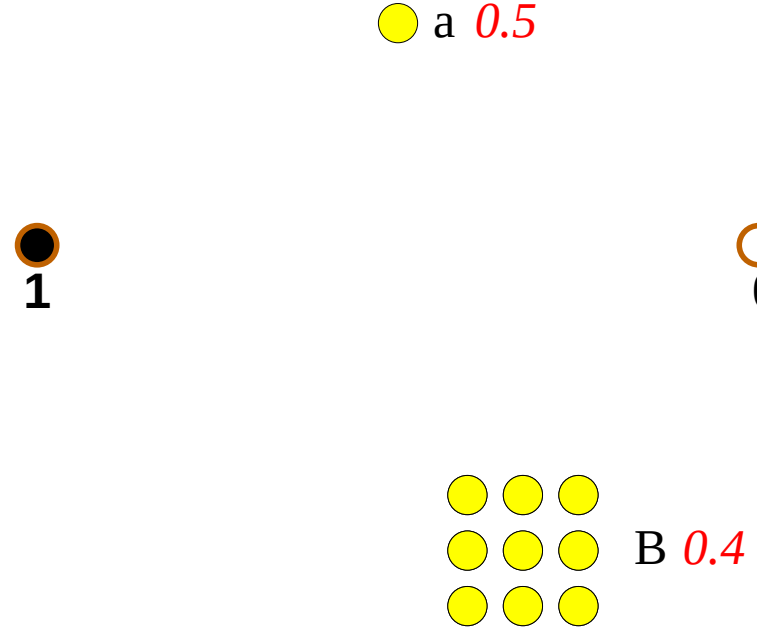
given vs. query

- pool based

- active + semi-supervised learning

Active Learning

Most ambiguous point not necessarily the best query:



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 error

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

Active Learning

estimated error after querying k

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

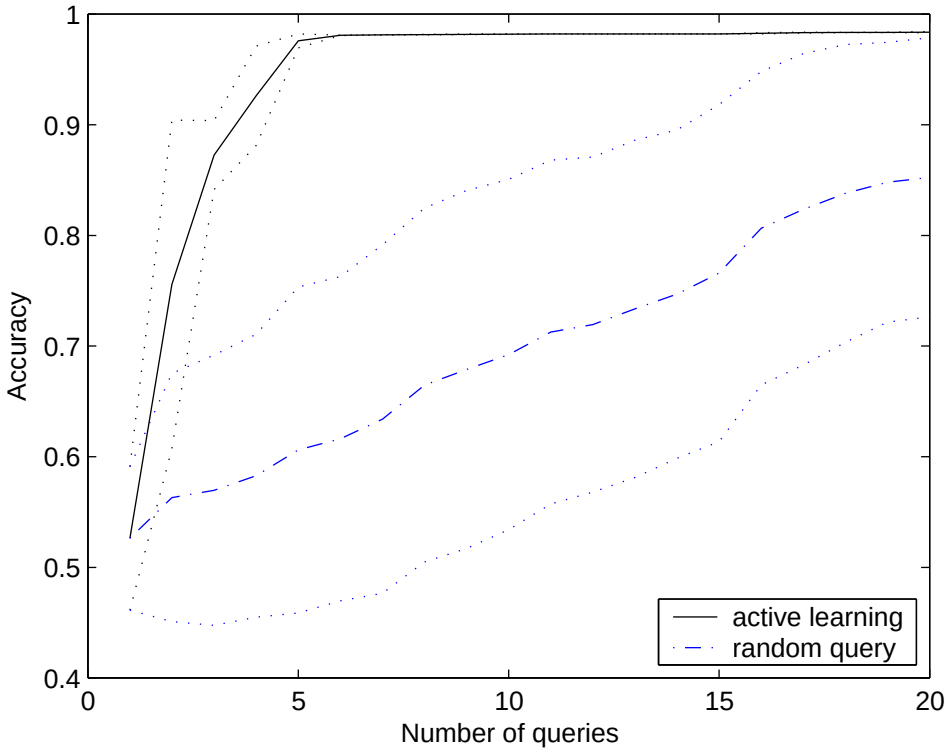
select query k^* to minimize the estimated error

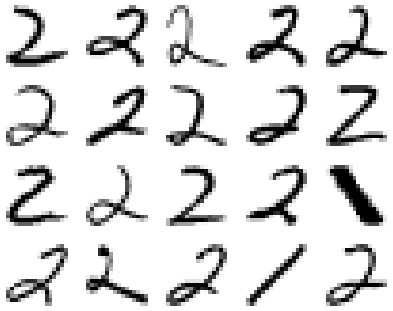
$$k^* = \arg\min_k (1 - f_k)\widehat{\text{err}}^{+(x_k,0)} + f_k \widehat{\text{err}}^{+(x_k,1)}$$

're-train' is fast

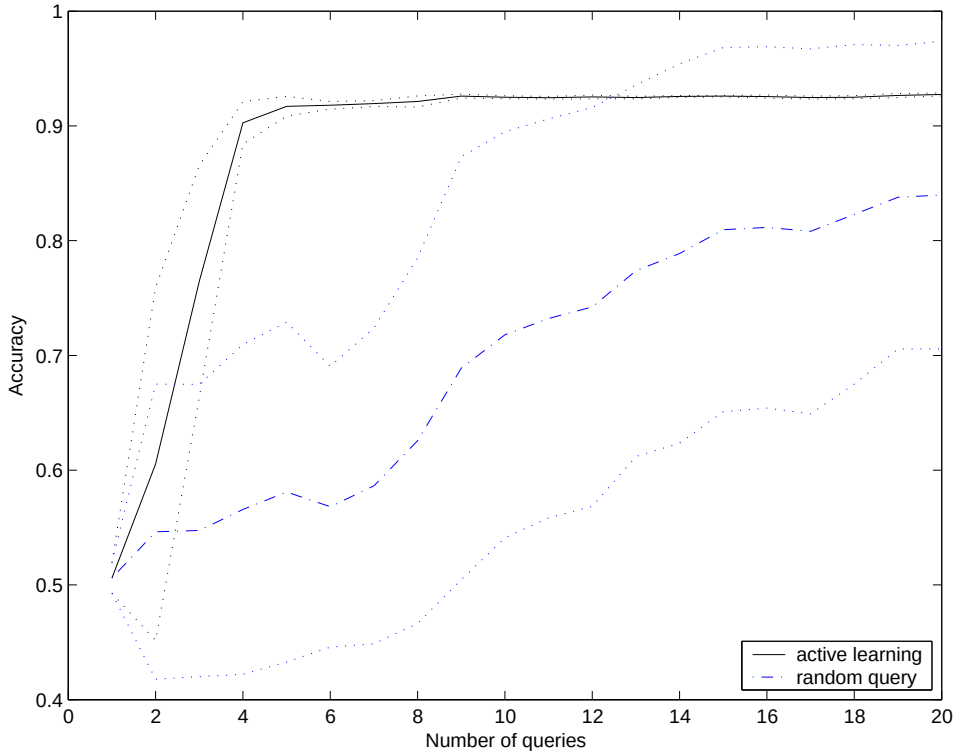
$$f_U^{+(x_k,y_k)} = f_U + (y_k - f_k)\frac{(\Delta_{UU})^{-1}_{\cdot k}}{(\Delta_{UU})^{-1}_{kk}}$$

Active Learning: OCR Digits "1" vs. "2"





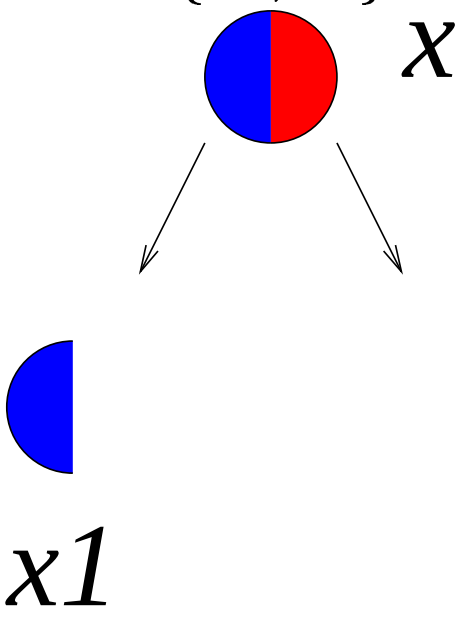
Active Learning: 20 Newsgroups PC vs. MAC



Co-training

split feature $x = \{x1, x2\}$

Method
mixture
trans. SVM
co-training
graph

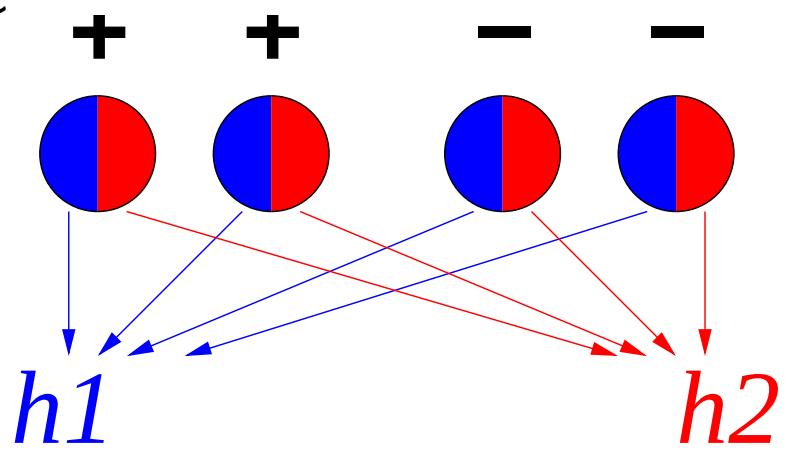


sub-features 'good' (conditional independence, sufficient)

Co-training

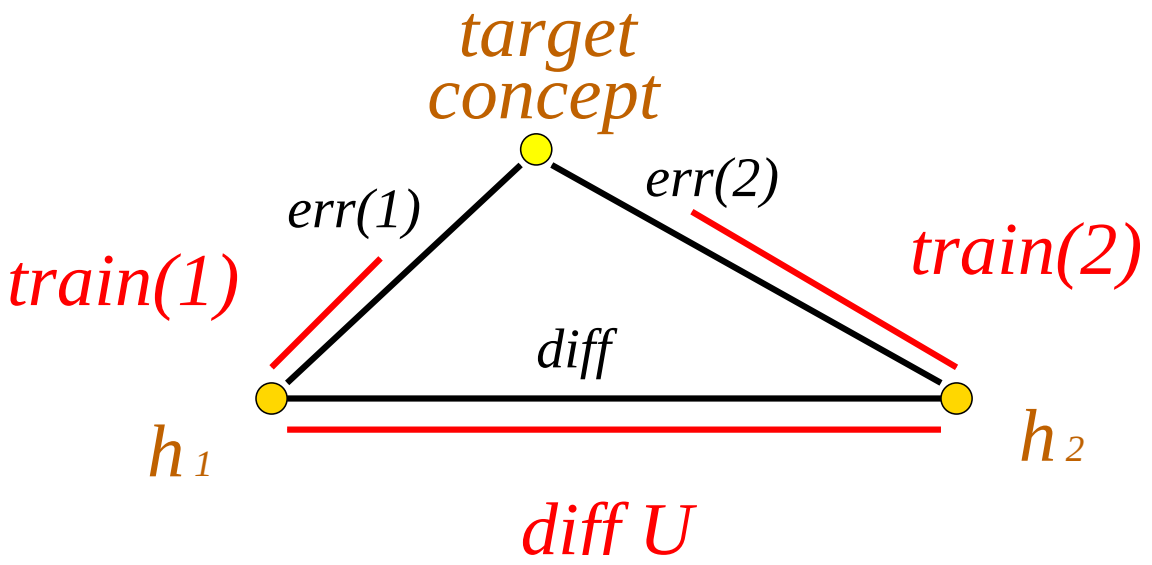
Method
mixture
trans. SVM
co-training
graph

train two classifiers, one on each sub-feature



Metric-based Model Selection

Method
mixture
trans. SVM
co-training
metric
graph



Weak.

[Schuurmans & Southey 01]

Metric-based Model Selection

Method
mixture
trans. SVM
co-training
metric
graph

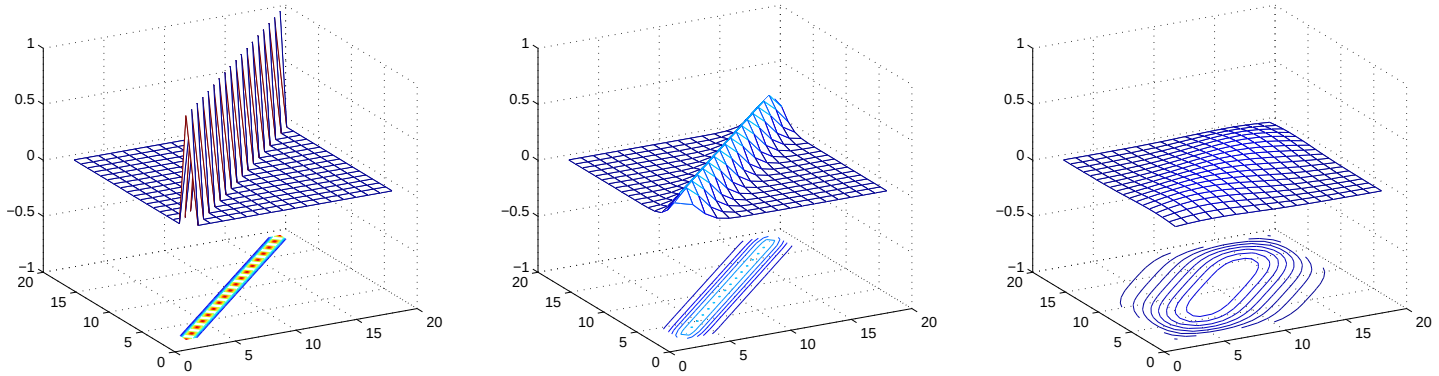
Consider 2 hypotheses h_1, h_2

- both have zero training error
- disagree a lot on unlabeled data

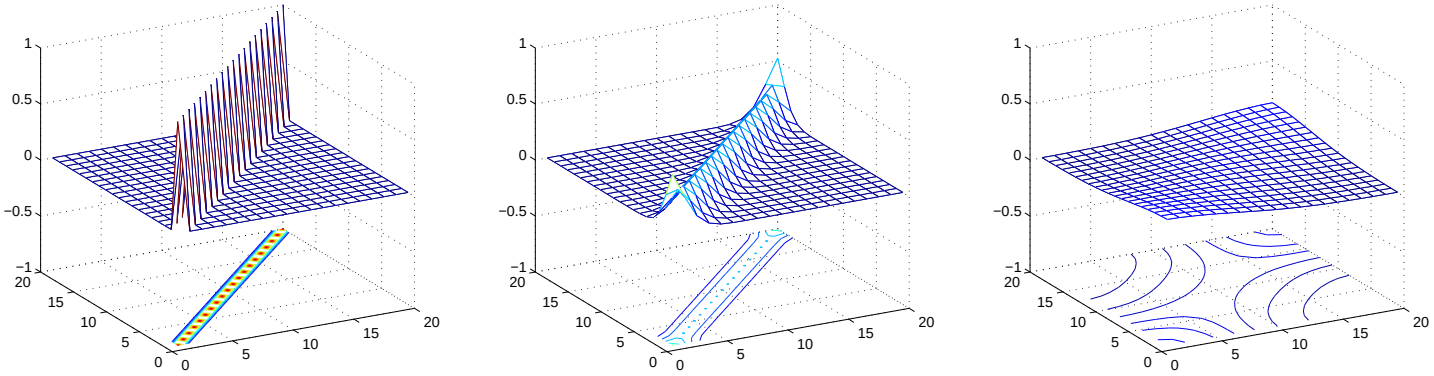
h_1, h_2 can't be both correct.

Connection: Graph Kernels

Diffusion kernel at time t: $e^{-t\Delta_{UU}}$ [Kondor & Lafferty 02]



t-step random walk: $(D^{-1}W)^t$ [Szummer & Jaakkola 01]



Connection: Graph Kernels

$$f_U = -(\Delta_{UU})^{-1}\Delta_{UL}f_L$$

Integrate diffusion kernels over t: $(\Delta_{UU})^{-1} = \int_0^\infty e^{-t\Delta_{UU}}\,dt$