

Transductive Learning via Spectral Graph Partitioning

Thorsten Joachims

Cornell University
Department of Computer Science

Overview

- 1) What is the transductive learning setting?
- 2) Why can unlabeled data help in supervised learning?
- 3) What does a good (reasonable) transductive learner look like?
- 4) A transductive learner based on graph cuts.
- 5) Connection to Co-Training.

Transductive Learning Process

Sampling Training data

- select random subset of l examples from DB of size n

$$\Rightarrow Z = [x_1, \dots, x_l]$$

- receive labels for these examples positive (+1) / negative (-1)

$$\Rightarrow Z = [(x_1, y_1), \dots, (x_l, y_l)]$$

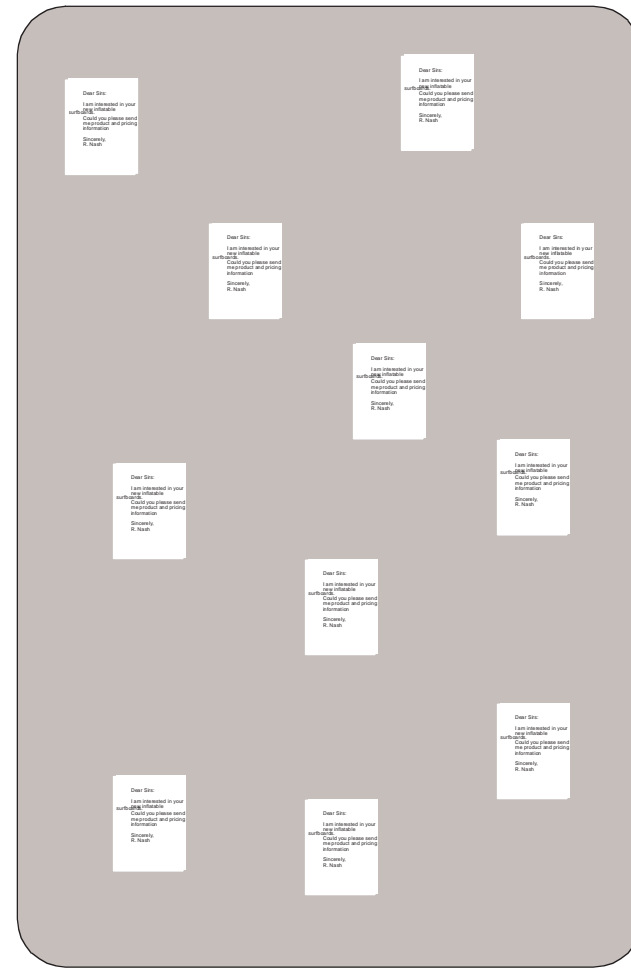
Goal of Learner

- predict the labels of the remaining examples $Z_x^* = [x_1^*, \dots, x_k^*]$

Opportunity

- Learning algorithm can study the test examples $Z_x^* = [x_1^*, \dots, x_k^*]$

Document DB



Transductive Learning Process

Sampling Training data

- select random subset of l examples from DB of size n

$$\Rightarrow Z = [x_1, \dots, x_l]$$

- receive labels for these examples positive (+1) / negative (-1)

$$\Rightarrow Z = [(x_1, y_1), \dots, (x_l, y_l)]$$

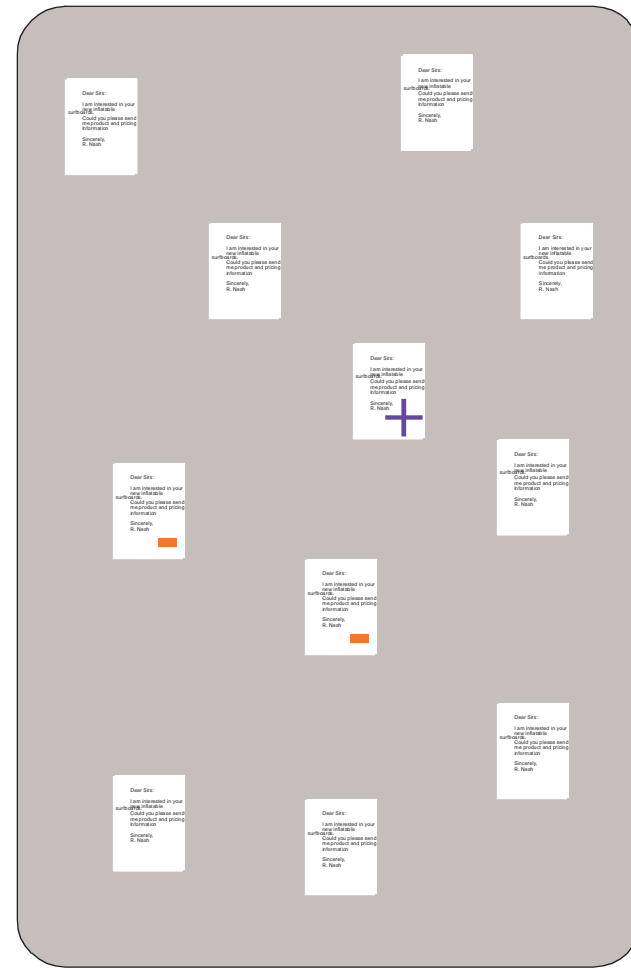
Goal of Learner

- predict the labels of the remaining examples $Z_x^* = [x_1^*, \dots, x_k^*]$

Opportunity

- Learning algorithm can study the test examples $Z_x^* = [x_1^*, \dots, x_k^*]$

Document DB



Transductive Learning Process

Sampling Training data

- select random subset of l examples from DB of size n
 $\Rightarrow Z = [x_1, \dots, x_l]$
- receive labels for these examples positive (+1) / negative (-1)
 $\Rightarrow Z = [(x_1, y_1), \dots, (x_l, y_l)]$

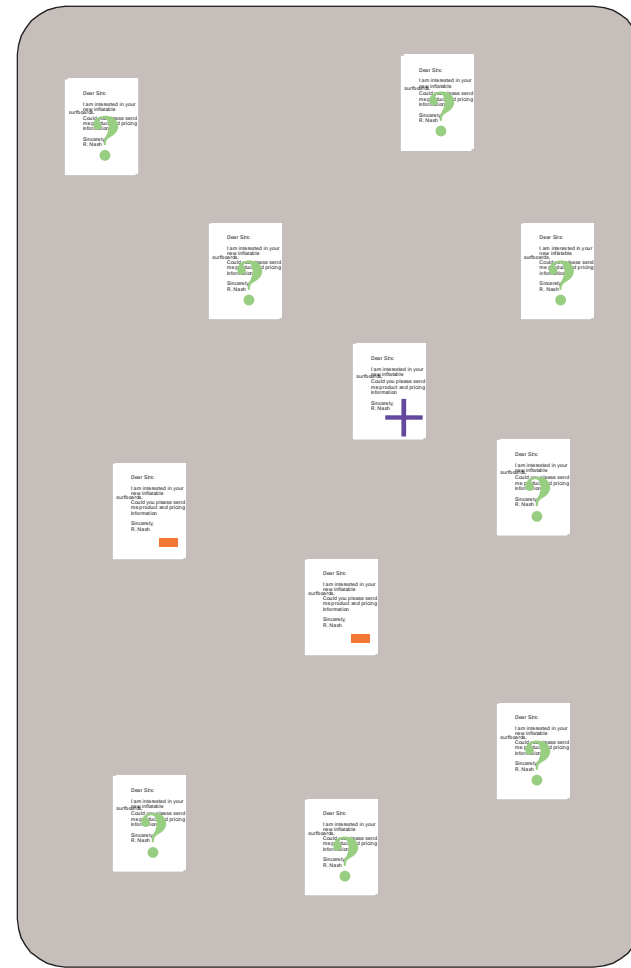
Goal of Learner

- predict the labels of the remaining examples $Z_x^* = [x_1^*, \dots, x_k^*]$

Opportunity

- Learning algorithm can study the test examples $Z_x^* = [x_1^*, \dots, x_k^*]$

Document DB



Example: Exploiting the Test Set

How would you classify the test set?

	nuclear	physics	atom	pepper	basil	salt	and
D1	1						1
D2	1	1	1				1
D3			1				1
D4				1	1		1
D5				1		1	1
D6					1	1	1

- training set {D1, D6}
- test set {D2, D3, D4, D5}

Example: Exploiting the Test Set

How would you classify the test set?

	nuclear	physics	atom	pepper	basil	salt	and
D1	1						1
D2	1	1	1				1
D3			1				1
D4				1	1		1
D5				1		1	1
D6					1	1	1

- training set {D1, D6}
- test set {D2, D3, D4, D5}

Transductive Support Vector Machines [Vapnik]

Objective: maximize margin δ on both training and test examples

Training sample: $Z = [(x_1, y_1), \dots, (x_l, y_l)]$

Test sample: $Z_x^* = [x_1^*, \dots, x_k^*]$

Optimization Problem:

$$\min_{y_1^\circ \dots y_k^\circ \in \{-1, 1\}} \min_{w \in \mathfrak{R}^d} w \cdot w + C \sum \xi_i$$

$$y_1 [w \cdot x_1 + b] \geq 1 - \xi_1$$

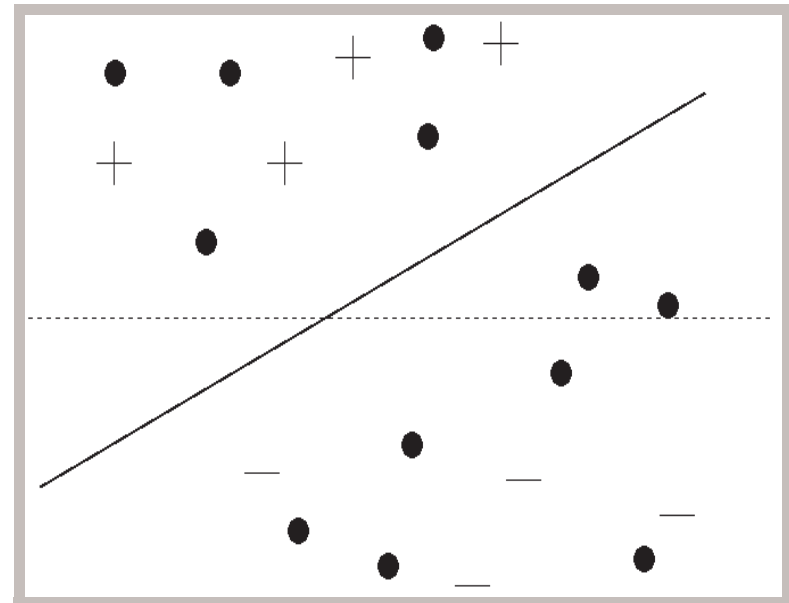
subject to ...

$$y_l [w \cdot x_l + b] \geq 1 - \xi_l$$

$$y_1^\circ [w \cdot x_1^* + b] \geq 1$$

...

$$y_k^\circ [w \cdot x_k^* + b] \geq 1$$



=> Integer prog. [Bennet et al., 98], local search [Joachims, 99]

Example: Exploiting the Test Set

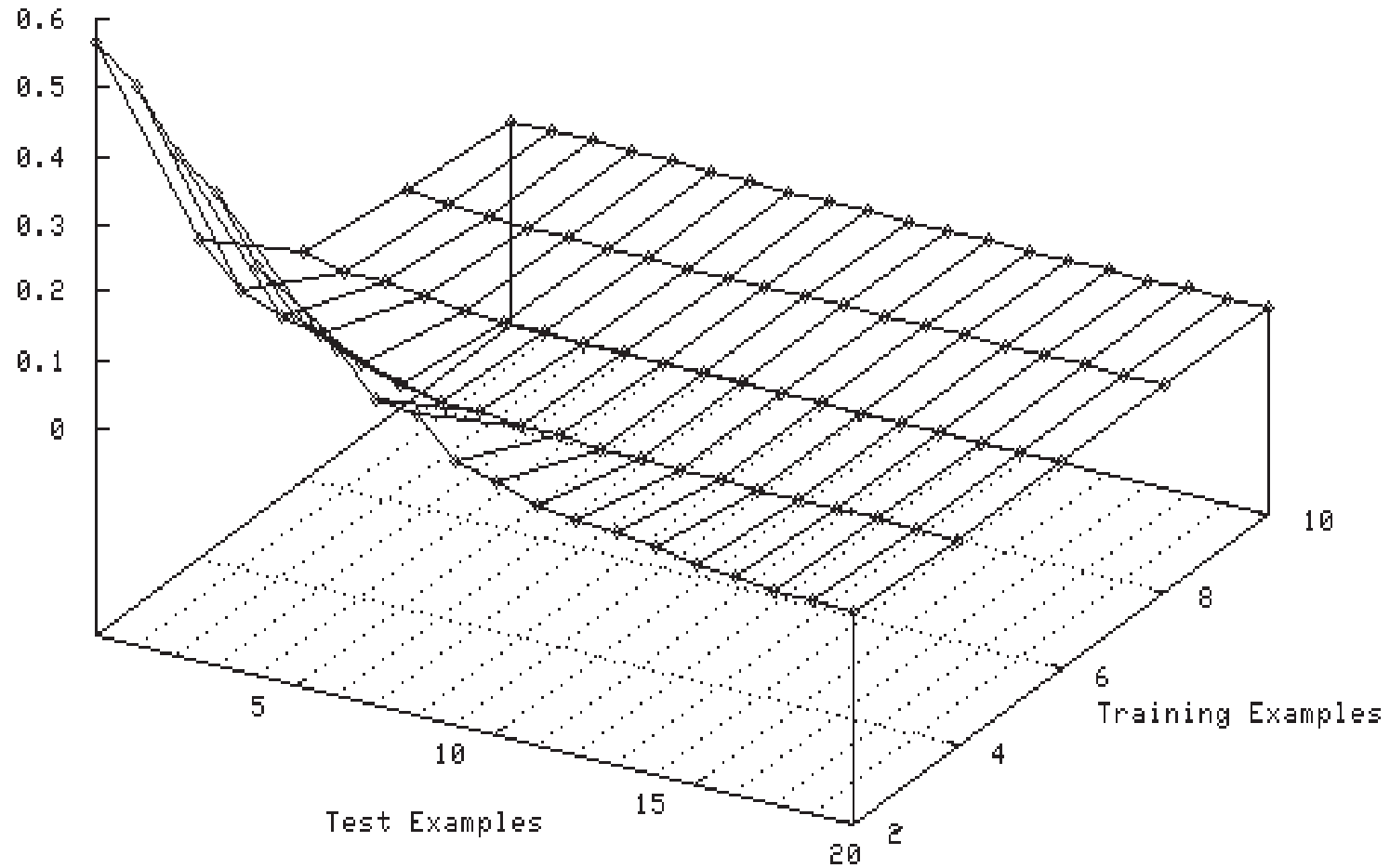
This is the labelling of the test set with the largest margin and 0 error!

	nuclear	physics	atom	pepper	basil	salt	and
D1	1						1
D2	1	1	1				1
D3			1				1
D4				1	1		1
D5				1		1	1
D6					1	1	1

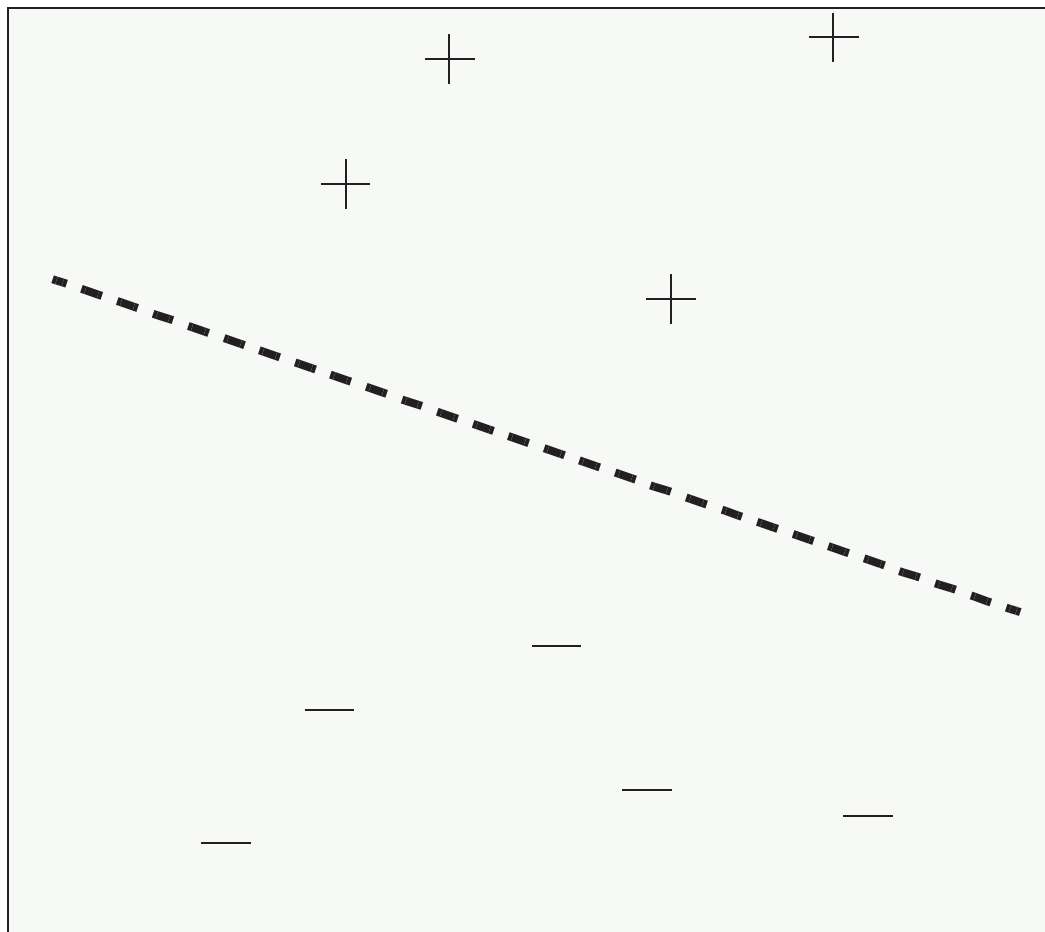
- training set {D1, D6}
- test set {D2, D3, D4, D5}

Simulation

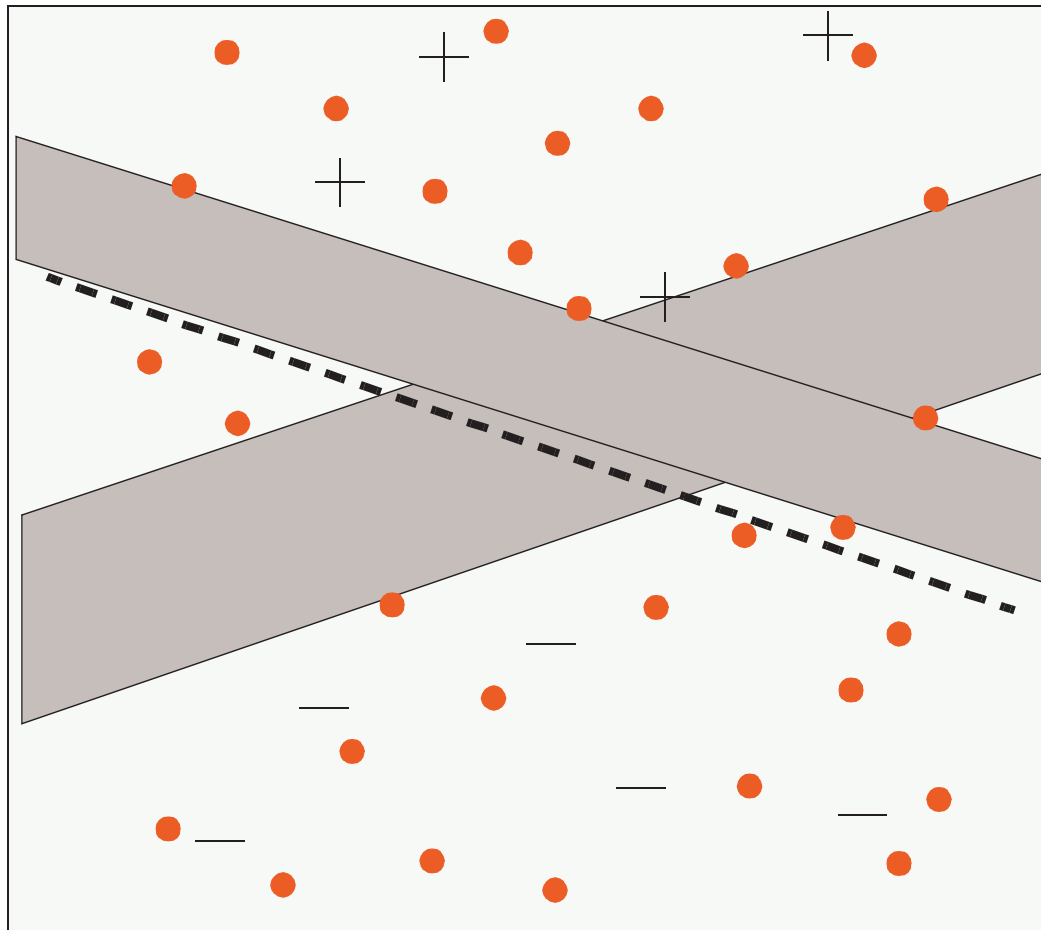
Target concept: $TCat([1:0:1],[0:1:1],[4:4:8])$



Why Does Adding Test Examples Reduce Error?

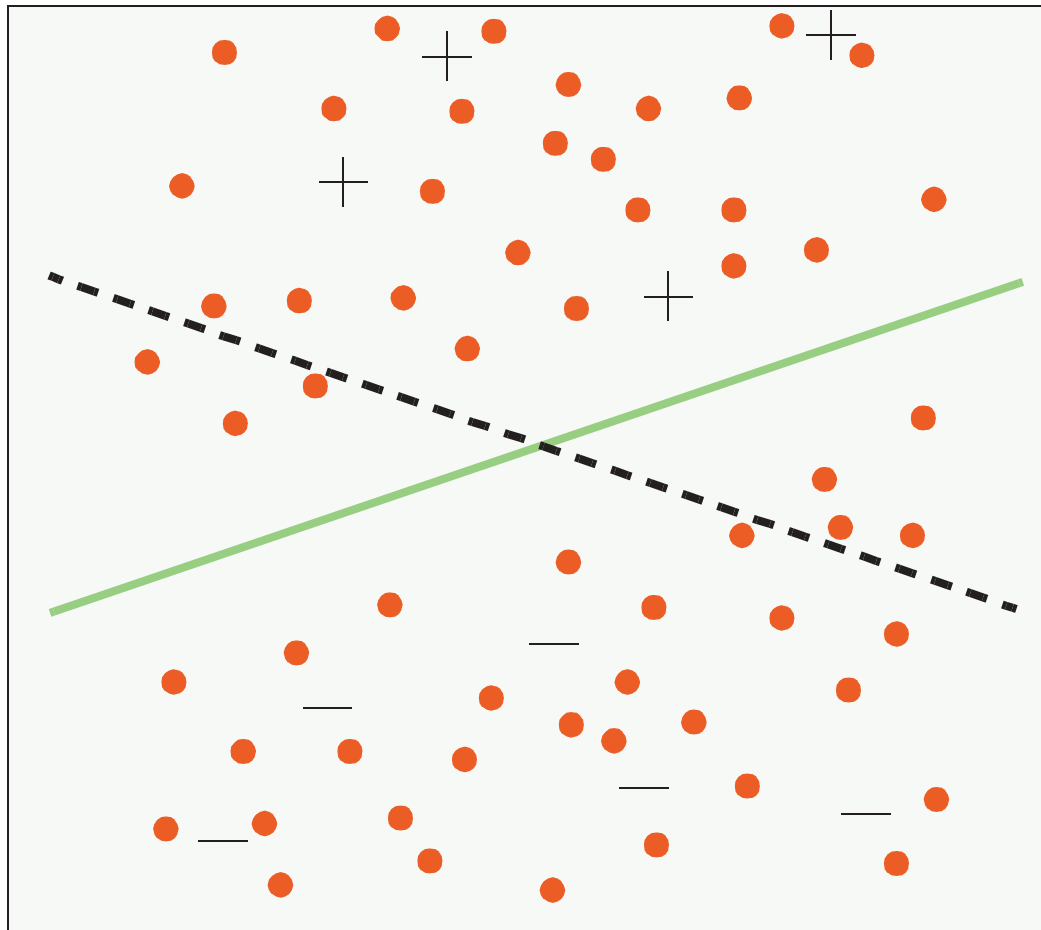


Why Does Adding Test Examples Reduce Error?



$$\text{Margin } \delta \geq \frac{1}{\sqrt{2}}$$

Why Does Adding Test Examples Reduce Error?



$$\text{Margin } \delta \geq \frac{1}{\sqrt{2}}$$

What Does it Mean to Maximize Margin on Test Examples?

Assumption: If we had much more labeled data, an inductive SVM would learn accurate classifier;

- then, this rule will have low leave-one-out error;
=> the perfect labeling of the data will be self-consistent.
- then (typically), this rule will have large margin.

Lemma: Inductive SVM (separable) $Err_{loo} \leq \frac{R^2}{\delta^2}$ [Vapnik, 1998].

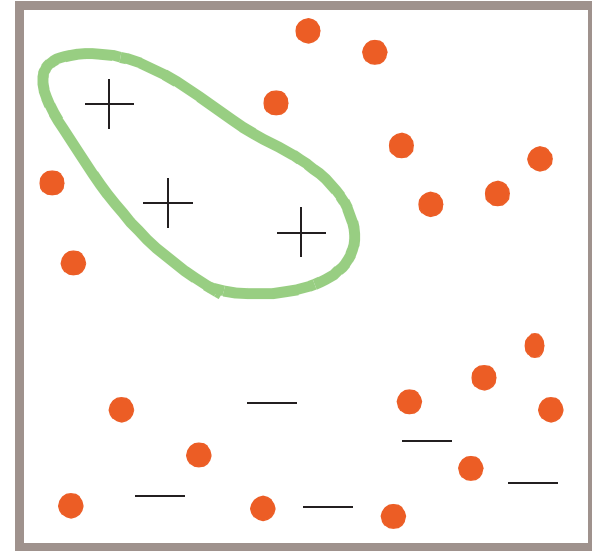
=> Transductive SVM finds a labeling of the test examples, so that an Inductive SVM would have low leave-one-out error [Joachims, 1999].

Ridge Regression [Chapelle et al. 1999], Nearest Neighbor [Blum and Chawla, 2001]

Taking Margin-Maximization Literally!

Problem 1:

In high dimensional spaces, the largest margin (with zero training error) is typically achieved when all test examples are in one class.



Solution:

Fix pos/neg ratio on the test set to that of the training set.

optimize via local search
[Joachims, 1999].

$$\begin{aligned} \min_{y^*_1 \dots y^*_k \in \{-1, 1\}} \quad & \min_{w \in \mathbb{R}^d} w \cdot w \\ \text{subject to} \quad & y_1 [w \cdot x_1 + b] \geq 1 \\ & y_l [w \cdot \overset{\dots}{x}_l + b] \geq 1 \\ & y^*_1 [w \cdot x^*_1 + b] \geq 1 \\ & y^*_k [w \cdot \overset{\dots}{x}^*_k + b] \geq 1 \\ & \sum y^*_i = c \end{aligned}$$

Experiment: Reuters-21578

Reuters Newswire Stories

- 90 categories
- 9603 training documents
- 3299 test documents

Experiment

- 10 most frequent categories
- 17 training documents
- 3299 test documents
- ca. 700-12000 features

PRBEP	Bayes	SVM	TSVM
earn	78.8	91.3	95.4
acq	57.4	67.8	76.6
money-fx	43.9	41.3	60.0
grain	40.1	56.2	68.5
crude	24.8	40.9	83.6

PRBEP	Bayes	SVM	TSVM
trade	22.1	29.5	34.0
interest	24.5	35.6	50.8
ship	33.2	32.5	46.3
wheat	19.5	47.9	54.4
corn	14.5	41.3	43.7

[Joachims, 1999]

Taking Margin-Maximization Literally!

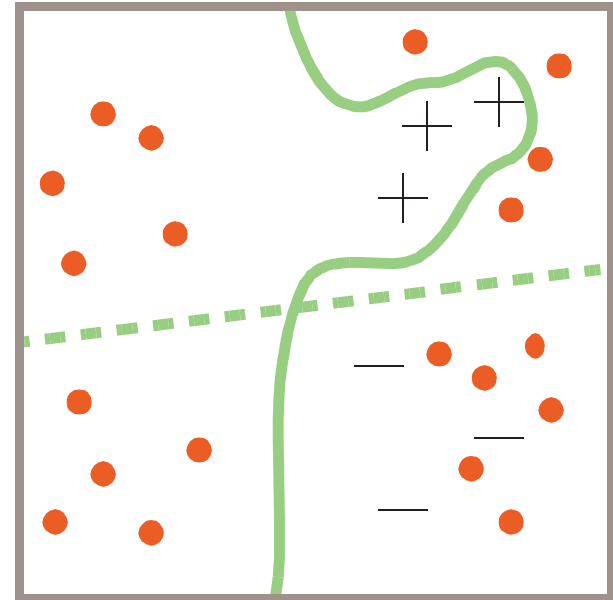
Problem 2:

A large margin on test data "overpowers"
the training data

=> Training examples are "outliers"!

Solution:

Use additional regularity criteria.



Lemma: Conditions for Leave-One-Out Error of Inductive SVM

- necessary condition: $\left[h_i(\hat{x}_i) \neq y_i \right] \Rightarrow \left[\rho \alpha_i R^2 + \xi_i \geq 1 \right] \quad 1 \leq \rho \leq 2$

[Jaakkola and Haussler, 1999] [Joachims, 2000]

- sufficient condition: $\left[\xi_i \geq 1 \right] \Rightarrow \left[h_i(\hat{x}_i) \neq y_i \right] \quad \text{[Joachims, 2002]}$

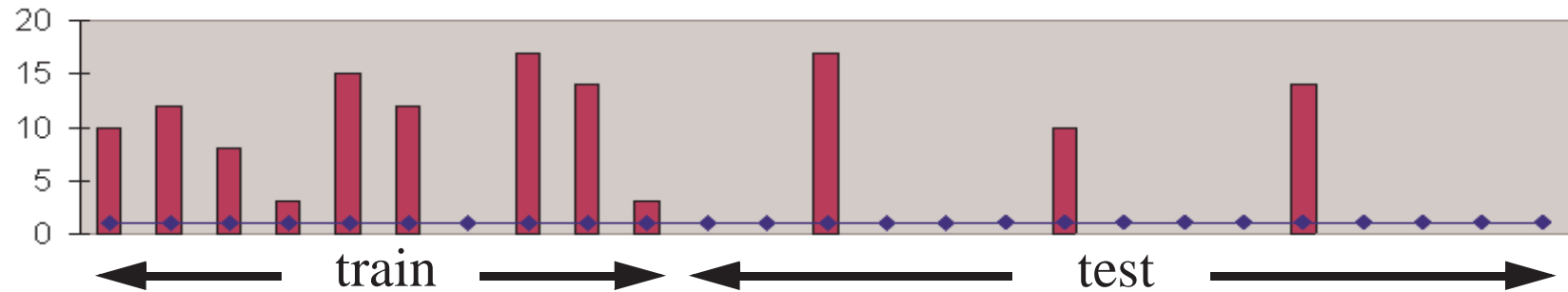
Constraints on the Transductive Hyperplane

Observation: For the optimal classification of the test examples

$$E_{S_{train}} \left(\frac{\left| \left\{ i \in S_{train}; (\alpha_i R^2 + \xi_i) \geq 1 \right\} \right|}{n} \right) = E_{S_{test}} \left(\frac{\left| \left\{ i \in S_{test}; (\alpha_i R^2 + \xi_i) \geq 1 \right\} \right|}{k} \right)$$

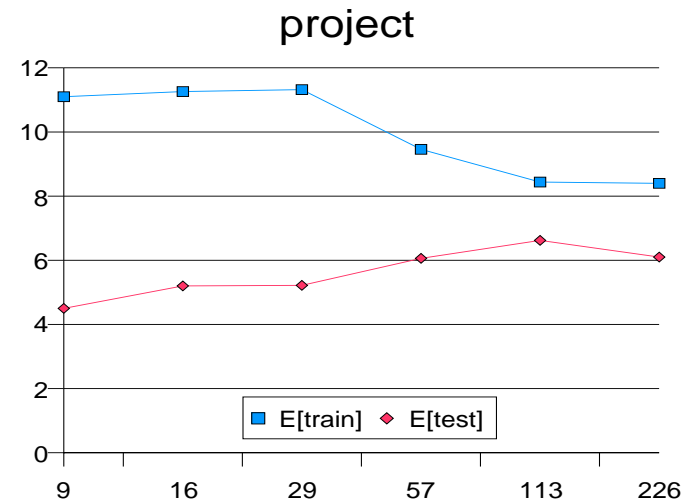
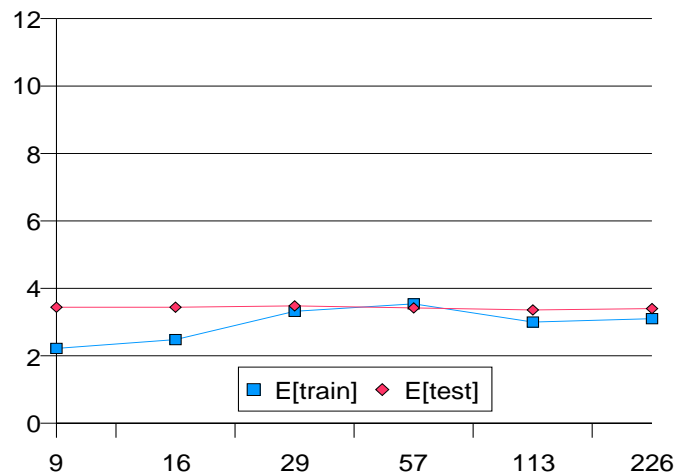
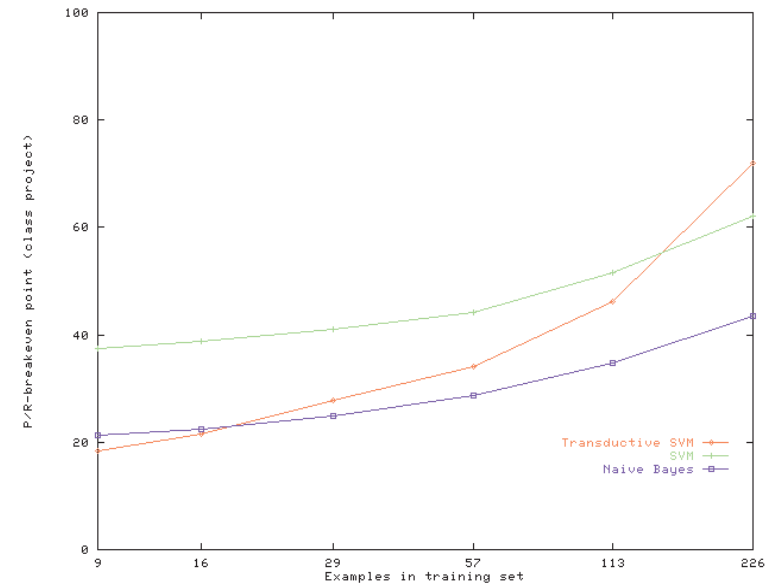
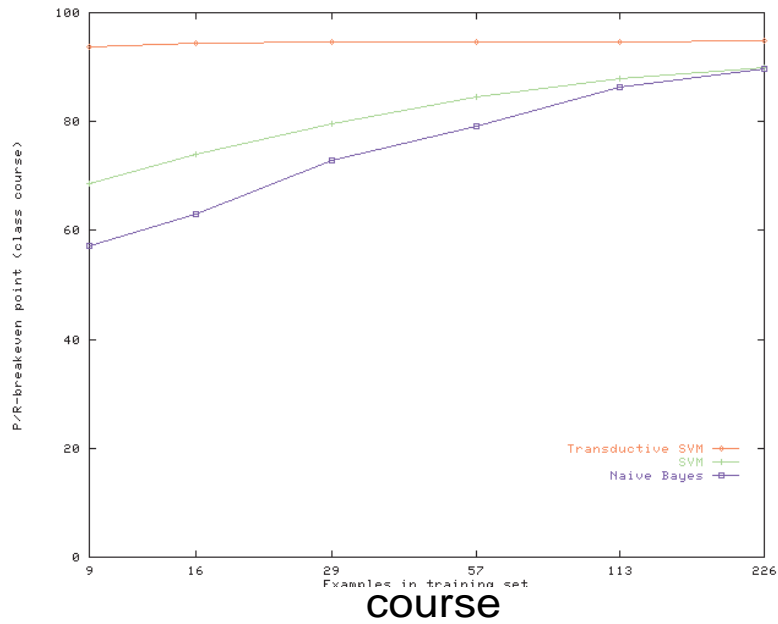
... and similarly for pos/neg examples separately [Joachims, 2002].

Example: Separable problem with the following α -values at solution.



$$f_{train} = 9 \quad f_{test} = 3 \quad P \left(\left| \frac{f_{train}}{10} - \frac{f_{test}}{15} \right| \geq \frac{7}{10} \right) \leq 0.01$$

Violation as Negative Evidence



What Should a Transductive Learner Do?

Postulate “low training error”.

Postulate "low leave-one-out": Find labeling of the test data that gives low leave-one-out error in inductive learner.

Postulate "uniform leave-one-out": Find a labeling, so that the leave-one-out errors are distributed equally between training and test examples.

Postulate "uniform pos/neg": Find a labeling, so that positive and negative examples are distributed equally between training and test examples.

... and potentially other constraints.

From TSVMs to Cuts

$$\min_{y^{\circ}_1 \dots y^{\circ}_k \in \{-1, 1\}} \max_{\alpha \in \mathbb{R}^n} 1' a - \frac{1}{2} \alpha' \text{diag}(y, y^{\circ}) H \text{diag}(y, y^{\circ}) \alpha$$

subject to

$$\begin{aligned} 0 &\leq \alpha_1 \\ &\vdots \\ 0 &\leq \alpha_l \\ &\vdots \\ 0 &\leq \alpha^{\circ}_1 \\ &\vdots \\ 0 &\leq \alpha^{\circ}_k \end{aligned}$$

Simplifying Assumption: All $\alpha_i, \alpha^{\circ}_j$ are equal at the solution

=> Closed form solution for quadratic program

Resulting Transductive Optimization Problem:

$$\max_{y^{\circ}_1 \dots y^{\circ}_k \in \{-1, 1\}} 1' \text{diag}(y, y^{\circ}) H \text{diag}(y, y^{\circ}) 1$$

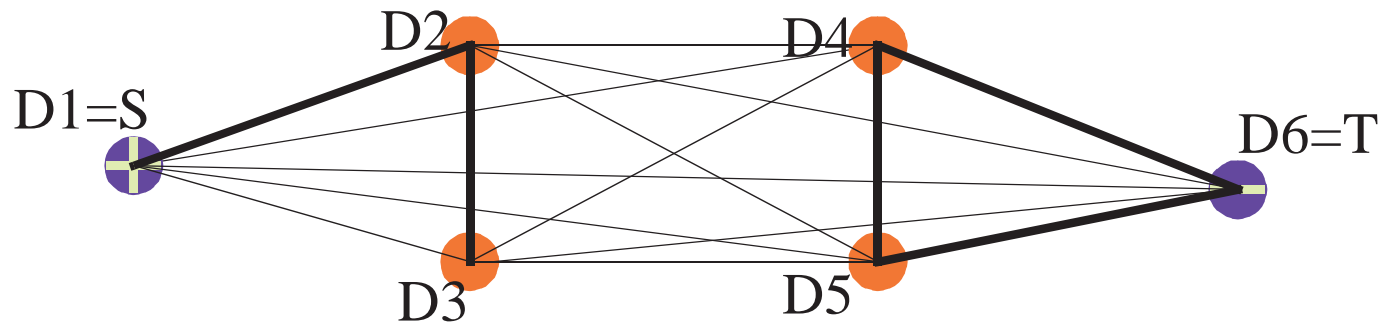
<=> s-t mincut with H as (symmetric) adjacency matrix.

$$\min_{y_1 \dots y_n \in \{-1, 1\}} \sum_{y_i \neq y_j} H_{ij} \quad \text{st} \quad \begin{aligned} y_i &= 1, \text{ if positive labeled} \\ y_i &= -1, \text{ if negative labeled} \end{aligned}$$

s-t Mincut for Transduction [Blum & Chawla, 01]

	nuclear	physics	atom	pepper	basil	salt	and
D1	1						1
D2	1	1	1				1
D3			1				1
D4				1	1		1
D5				1		1	1
D6					1	1	1

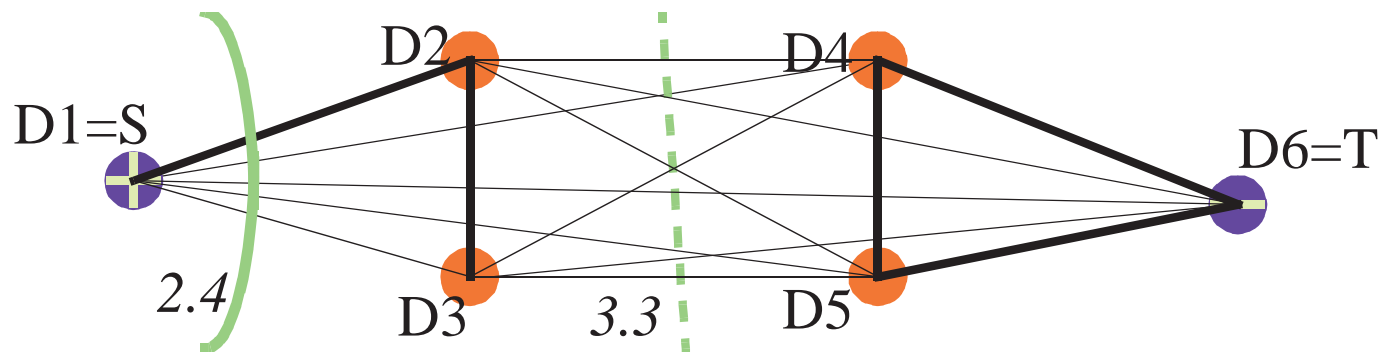
Adjacency matrix A based on cosine similarity:



s-t Mincut for Transduction [Blum & Chawla, 01]

	nuclear	physics	atom	pepper	basil	salt	and
D1	1						1
D2	1	1	1				1
D3			1				1
D4				1	1		1
D5				1		1	1
D6					1	1	1

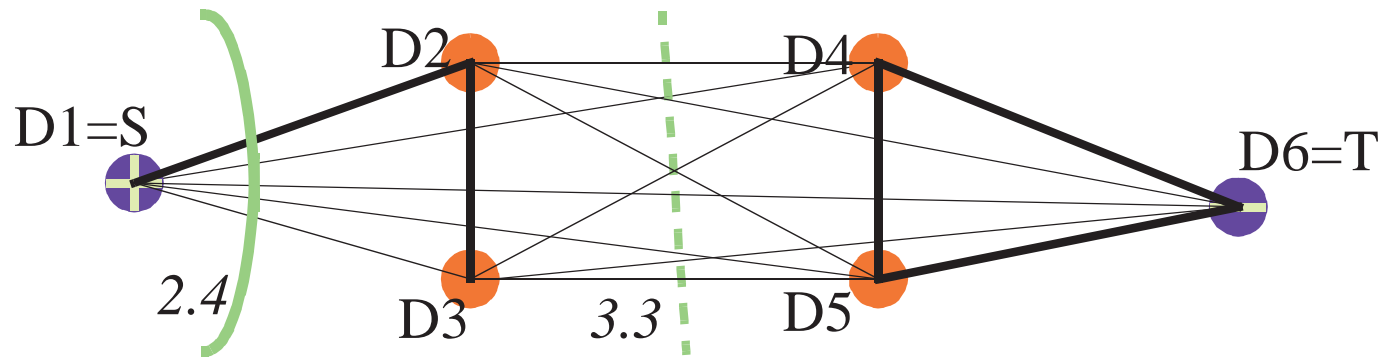
Adjacency matrix A based on cosine similarity:



Problem with s-t Mincuts

Postulate "low training error": OK

Postulate "uniform pos/neg": Violated

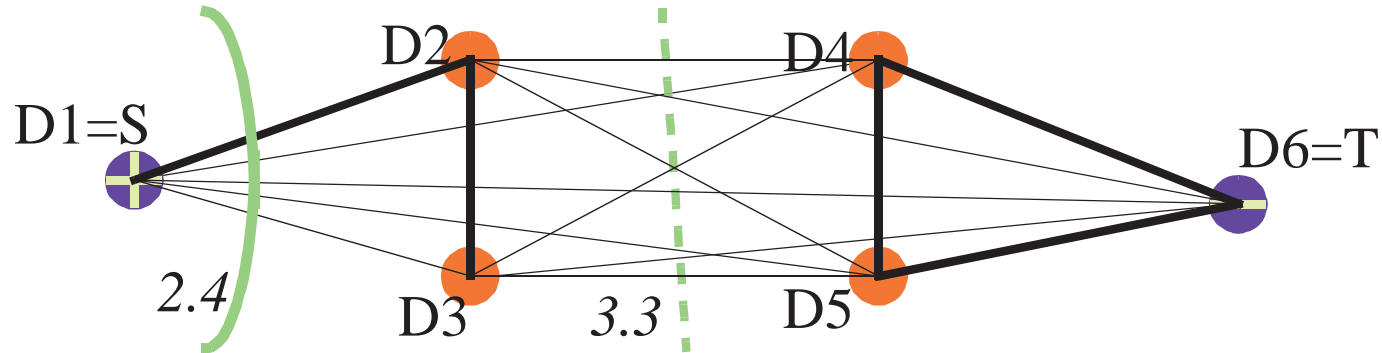


Postulate "low leave-one-out": OK [Blum and Chawla, 2001]

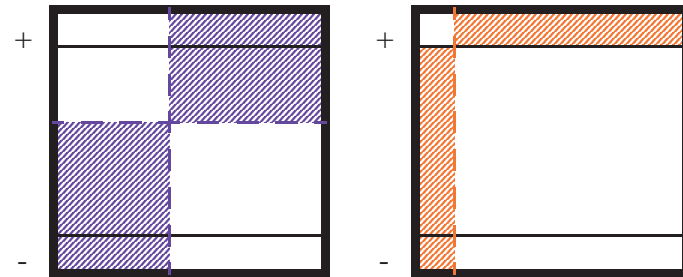
For K-NN graph H with edge-weight 1: $Err_{loo} \leq 2 \frac{mincut(A)}{k}$

Postulate "uniform leave-one-out": Violated

Ratio Cuts



s-t mincut: Minimize sum of edges.



Ratio Cut: Minimize average of edges => divide by "area"

$$\min_y \frac{cut(A, y)}{|\{i; y = 1\}| |\{i; y = -1\}|} \quad st \ y \in \{-1, 1\}^n$$

with symmetric adjacency matrix A and $cut(A, y) = \sum_{y_i \neq y_j} A_{ij}$.

Computing the Ratio Cut

Optimization Problem:

$$\min_y \frac{\text{cut}(A, y)}{|\{i; y_i = 1\}| |\{i; y_i = -1\}|} \quad \text{st } y \in \{-1, 1\}^n$$

Real Relaxation (see [Dhillon, 2001]):

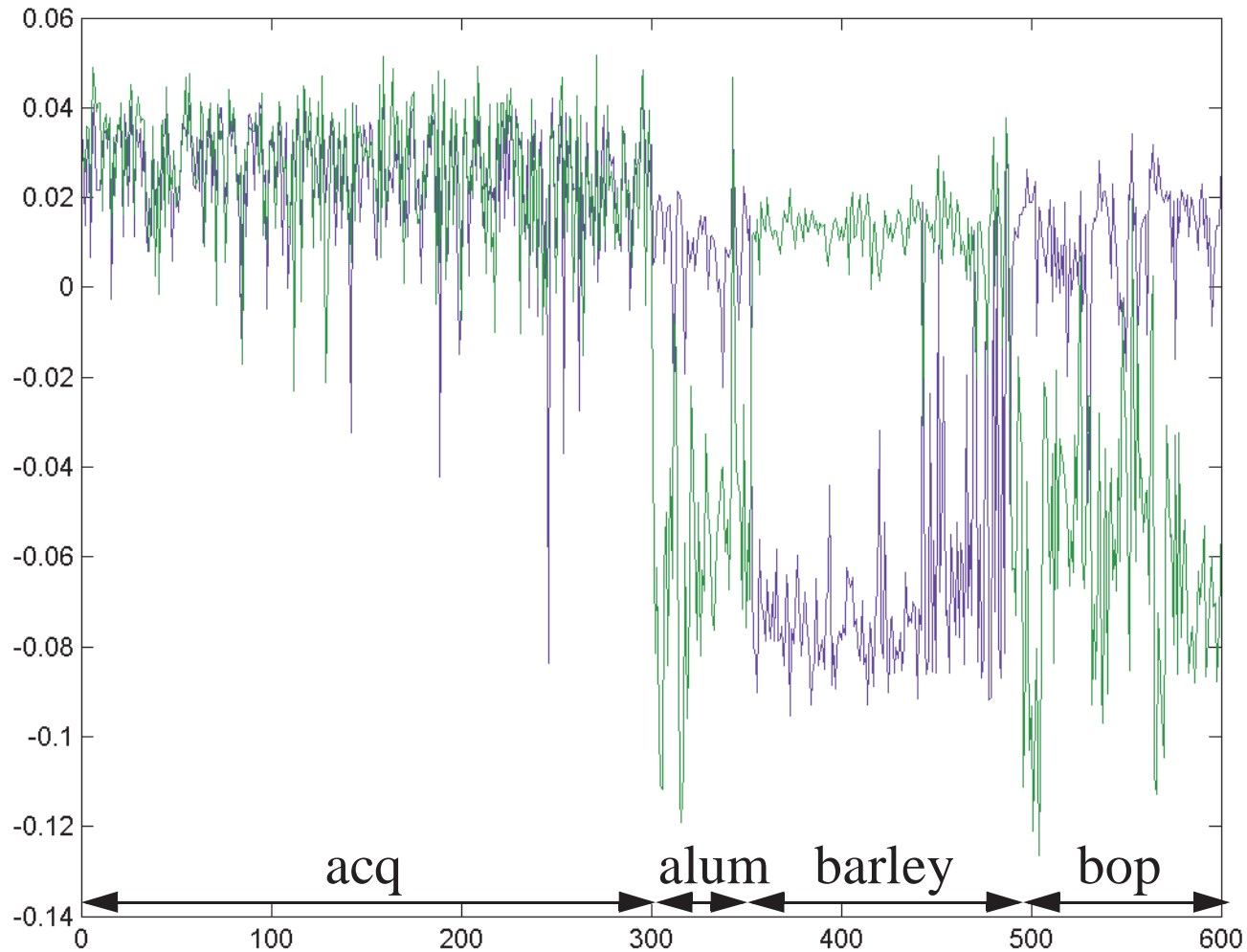
- Laplacian $L = B - A$, with degree matrix $B_{ii} = \sum_j A_{ij}$
- The solution y of the ratio cut has same sign as the solution z of

$$\min_z \frac{z' L z}{z' z} \quad \text{st } z \in \{\gamma_{pos}, \gamma_{neg}\}^n$$

$$\text{with } \gamma_{pos} = \sqrt{\frac{|\{i; z_i < 0\}|}{|\{i; z_i > 0\}|}} \text{ and } \gamma_{neg} = -\sqrt{\frac{|\{i; z_i > 0\}|}{|\{i; z_i < 0\}|}}.$$

- It holds that $z' z = n$ and $z' 1 = 0$.
- Solve real relaxation via 2nd lowest eigenvalue of L . Eigenvector is approximation of z [Hagen and Kahng, 92] [Shi and Malik, 2000].

Experiment: 20-NN Graph Text Clustering



see also [Dhillon, 2001]

Supervised Ratio Cut

Optimization problem:

$$\min_y \frac{\text{cut}(A, y)}{|\{i; y_i=1\}| |\{i; y_i=-1\}|} \quad \text{st } y \in \{-1, 1\}^n \quad \text{and} \quad \begin{array}{l} y_i = 1, \text{ if positive labeled} \\ y_i = -1, \text{ if negative labeled} \end{array}$$

=> eigenvalue problem [Gander et al, 1989] [Yu and Shi, 2002].

Optimization Problem with Training Error:

$$\min_{z \in \mathcal{R}^n} z' L z + c(z - \gamma)' C(z - \gamma) \quad \text{st } z' \mathbf{1} = 0, \quad z' z = n$$

- Estimate target values $\gamma_i = \sqrt{\frac{l_{neg}}{l_{pos}}}$ (pos), $\gamma_i = -\sqrt{\frac{l_{pos}}{l_{neg}}}$ (neg) from labeled data.
- c scalar trading off training error versus cut value.
- C diagonal matrix with cost per example.

Spectral Graph Transducer (SGT)

Preprocessing:

- Compute k-NN graph A_{knn} and symmetricize $A = A_{knn} + A'_{knn}$.
- Compute eigendecomposition of (normalized) Laplacian $(A - B) = VDV'$ ($B^{-1}(A - B) = VDV'$).
- Replace eigenvalues with $diag(D_{sparse}) = (0, 1, 4, 9, \dots, d^2, 0, \dots, 0)$ [Chapelle et al., 2002].

Prediction:

- Estimate $\gamma_i = \sqrt{\frac{l_{neg}}{l_{pos}}}$, $C_{ii} = \frac{l}{2l_{pos}}$ (pos), $\gamma_i = -\sqrt{\frac{l_{pos}}{l_{neg}}}$, $C_{ii} = \frac{l}{2l_{neg}}$ (neg).
- Solve

$$\min_{z \in \mathbb{R}^n} z'(VD_{sparse}V')z + c(z - \gamma)'C(z - \gamma) \quad st \quad z'z = n$$

via eigenvalue problem of size $2d$ [Gander et al, 1989].

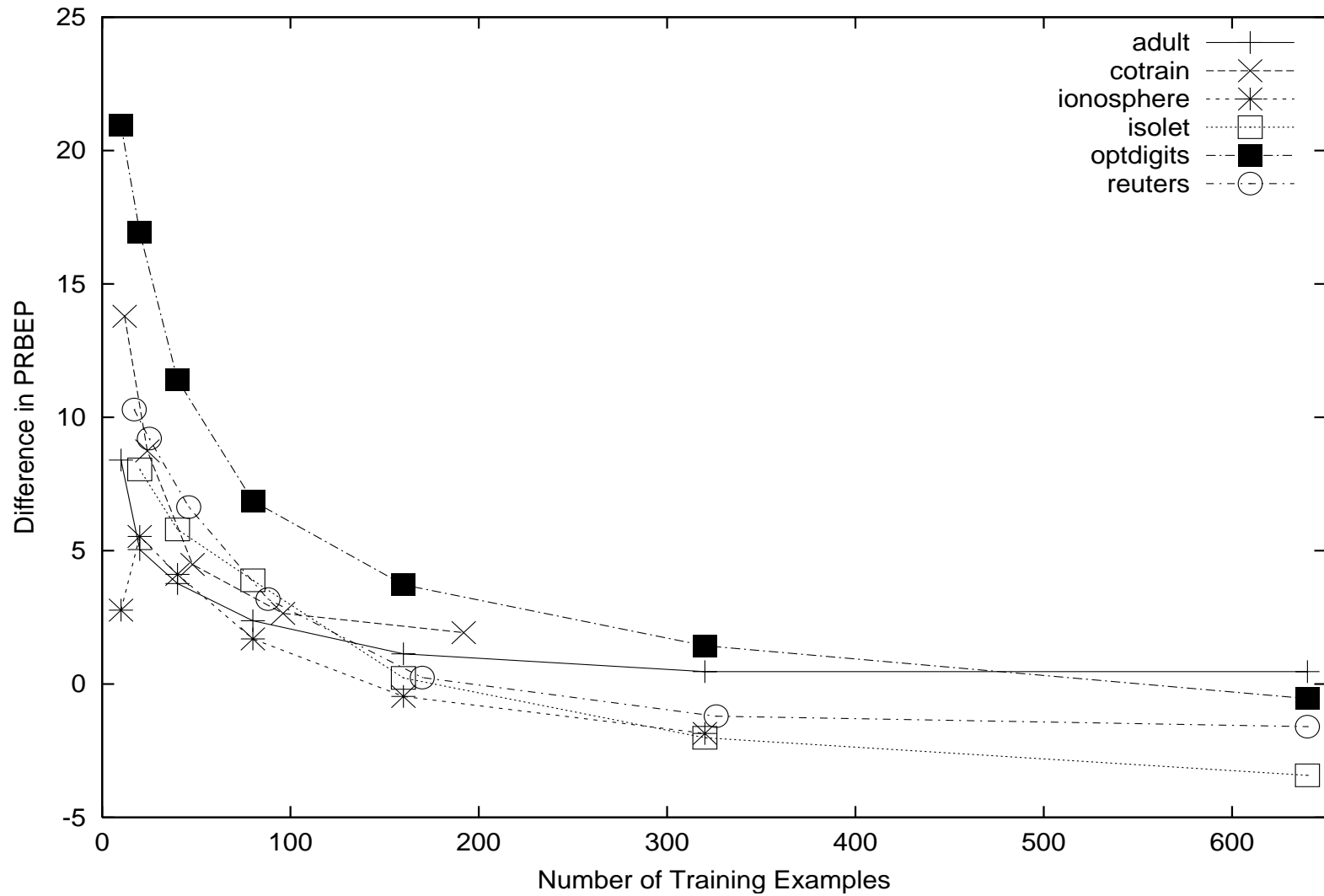
- prediction $y_i^\circ = sign(z_i - \Theta)$ with $\Theta = \frac{1}{2}(\gamma_{pos} - \gamma_{neg})$

Experiments

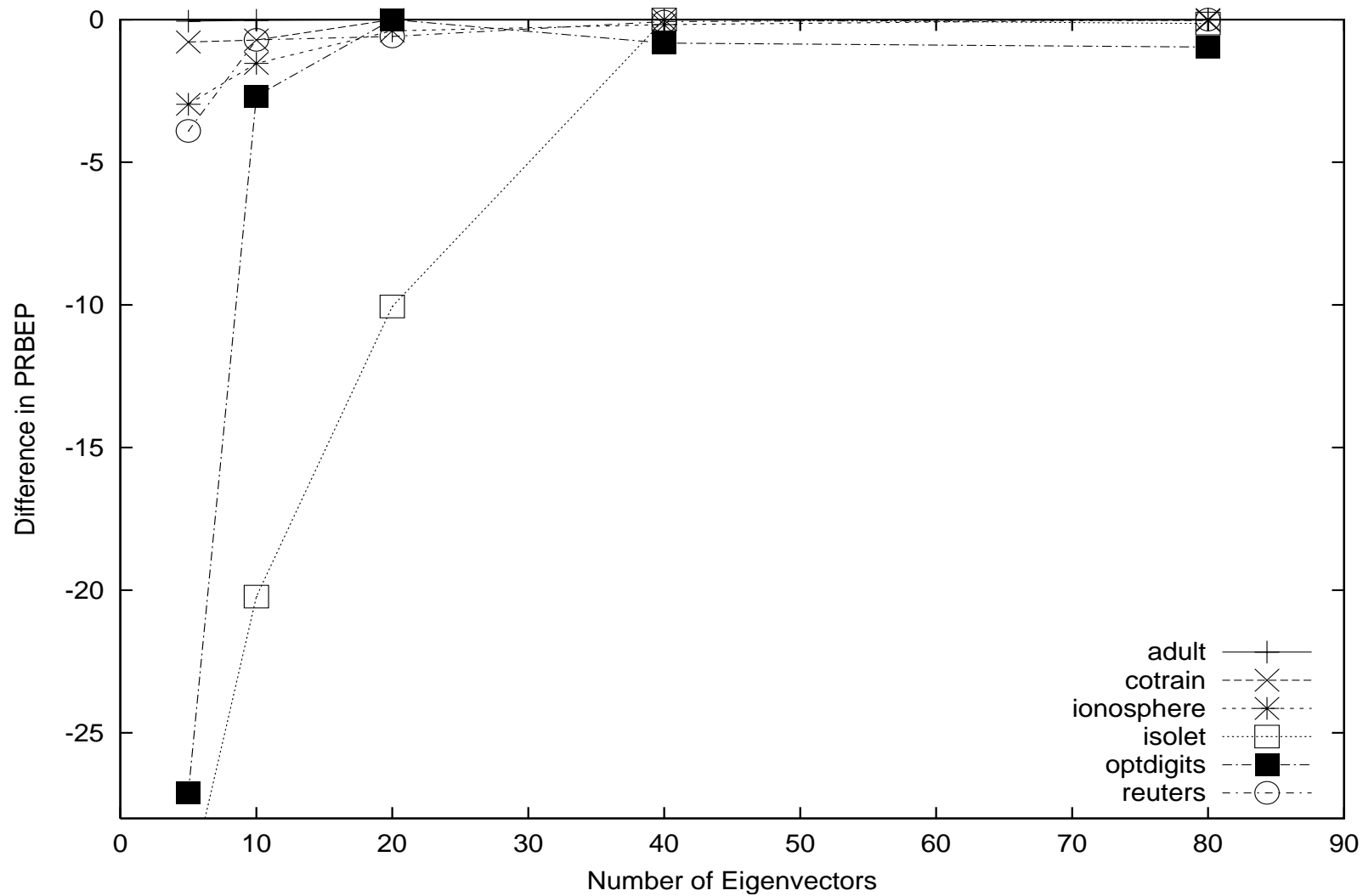
Task	n	l	SGT	KNN	TSVM	SVM
Reuters	3299	17	57.0	46.7	51.5	49.1
Optdigits	1797	10	83.4	62.4	61.5	61.6
Isolet	7797	20	55.5	47.4	-	46.3
Adult	32561	10	54.5	46.0	-	46.2
Ionosphere	351	10	79.6	76.7	80.6	80.6

- Macro-averaged precision/recall break-even point
- Average over 100 random training samples
- Parameters:
 - SGT: cosine similarity, $c = 3200$, $d = 80$
 - others: optimized on the test set

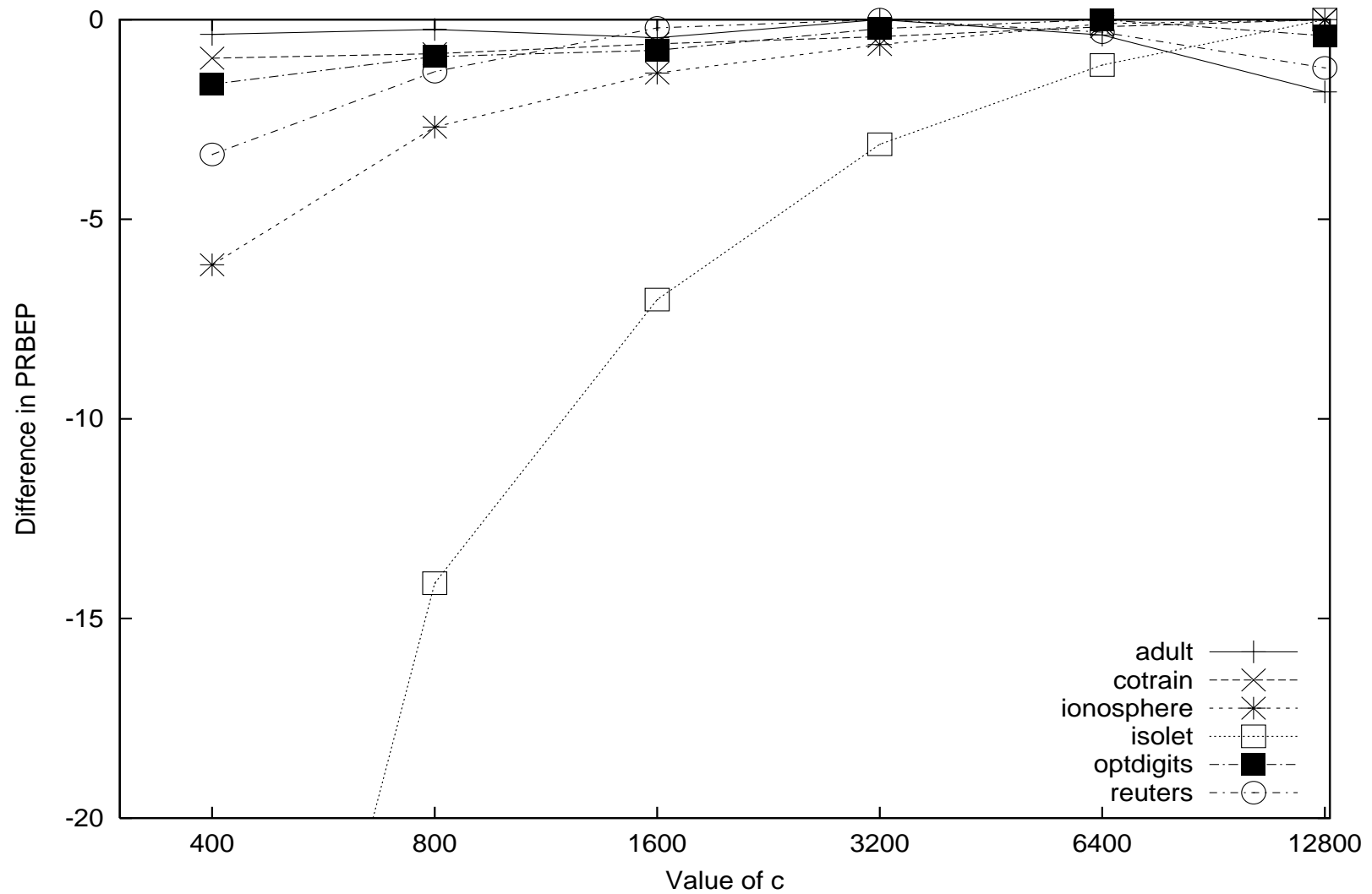
Size of the Training Set?



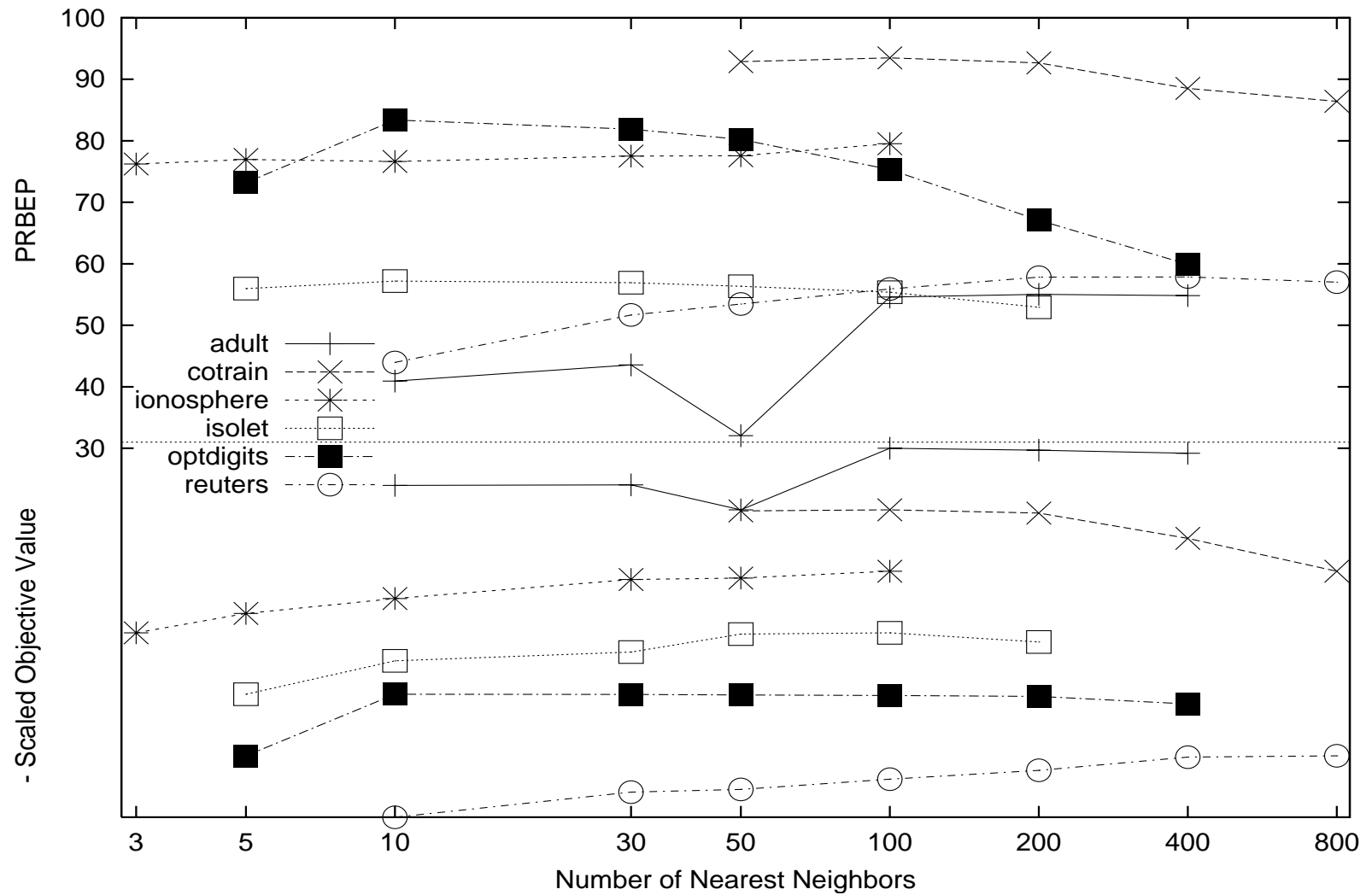
Choice of Number of Eigenvalues d ?



Choice of Training Error Penalty c ?



Choice of k for k-Nearest-Neighbor Graph?



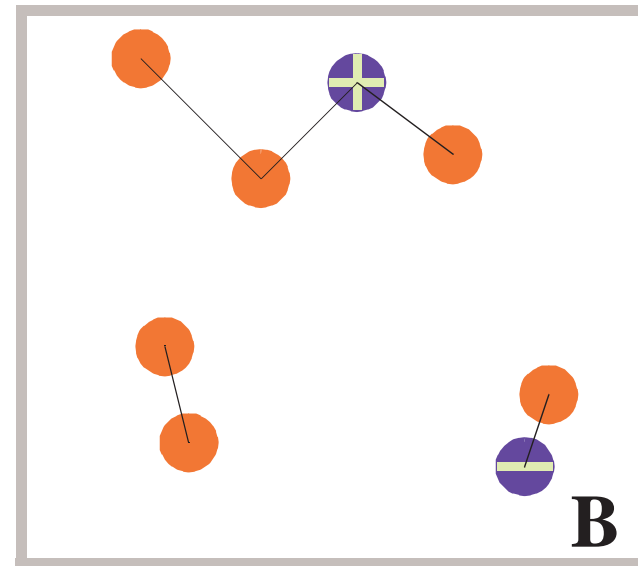
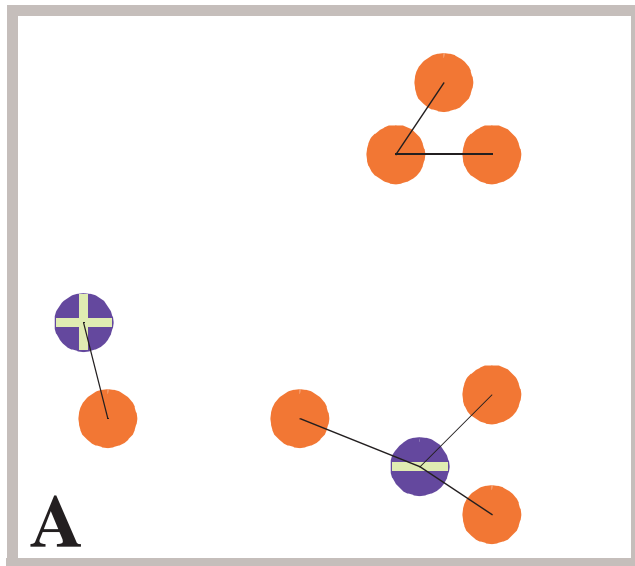
Co-Training [Blum & Mitchell]

Idea: Exploit two sufficiently redundant representations $X = A \times B$.

Scenarios:

- Web-page body text (A) / Hyperlinks pointing to page (B)
- Sound of person saying “hello” (A) / Image of lip movements (B)

Compatible: Perfect classifiers on A and B do not disagree!



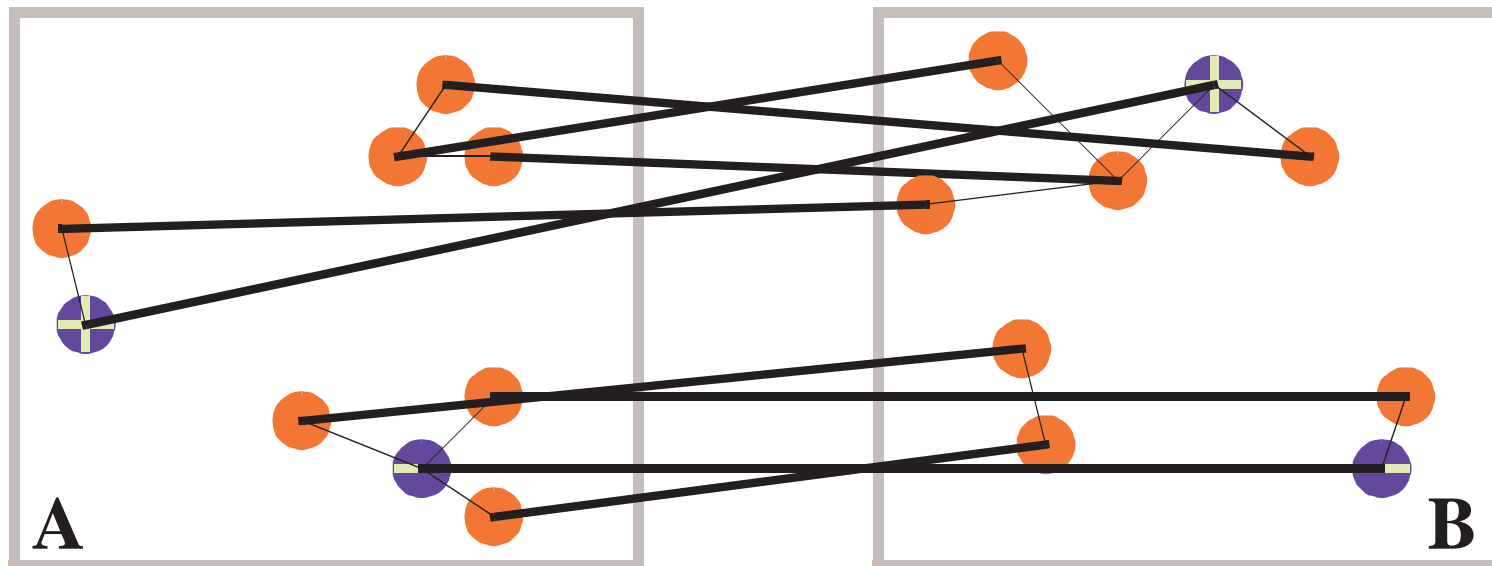
Co-Training [Blum & Mitchell]

Idea: Exploit two sufficiently redundant representations $X = A \times B$.

Scenarios:

- Web-page body text (A) / Hyperlinks pointing to page (B)
- Sound of person saying “hello” (A) / Image of lip movements (B)

Compatible: Perfect classifiers on A and B do not disagree!



Co-Training Experiment

	SGT	KNN	TSVM	SVM	B&M
cotrain	3.3	-	-	-	5.0
page+link	5.9	10.1	4.3	20.3	-
page	6.2	13.3	4.6	21.6	12.9
link	22.1	13.1	8.9	18.5	12.4

- Dataset: classifying course homepages from Blum and Mitchell
- 12 training examples, 1039 test examples
- 200 neighbors in each representation
- Error on test set averaged over 100 random test/training splits

Template for a Transductive Learner

Training Sample: $Z = [(x_1, y_1), \dots, (x_l, y_l)]$

Test Sample: $Z_x^* = [x_1^*, \dots, x_k^*]$

Predicted Labels $y^\circ = [y_1^\circ, \dots, y_n^\circ]$ optimize

$$\min_{y_1^\circ, \dots, y_n^\circ \in \{-1, 1\}} (\text{training error}) + (\text{alignment})$$

	training error	alignment
TSVM	$\sum \xi_i$	$\delta = 1/\ w\ $
Min s-t Cut	0	$cut(A, y^\circ)$
SGT	$(y^\circ - y)^2$	$\frac{cut(A, y^\circ)}{ \{i; y_i^\circ = 1\} \{i; y_i^\circ = -1\} }$

Conclusions

- Transductive learning problem.
- Constraints that a good transductive prediction should fulfill.
- Connection between
 - Inductive/Transductive SVMs
 - Inductive/Transductive KNN
 - graph cuts.
- Ratio cuts as a transductive version of k Nearest Neighbors.
- Co-Training as a special case of transductive learning.

Open Questions

- How to best formulate the optimization problems.
- Computing error bounds for transduction.
- How many (substantially different) cuts with cut-value less than δ exist?