Multi-View Learning over Structured and Non-Identical Outputs

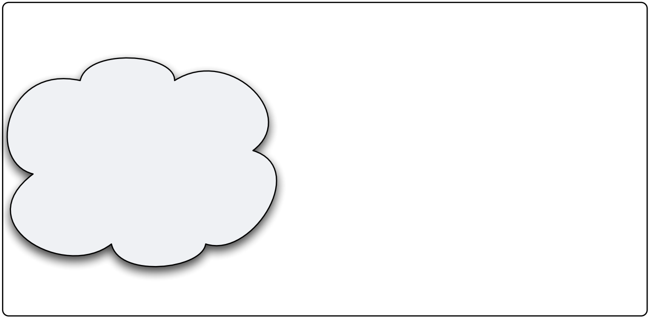
Kuzman Ganchev [1] João V. Graça [2] John Blitzer [1] Ben Taskar [1]

[1]Computer & Information Science
University of Pennsylvania

[2]INESC-ID
Lisboa, Portugal

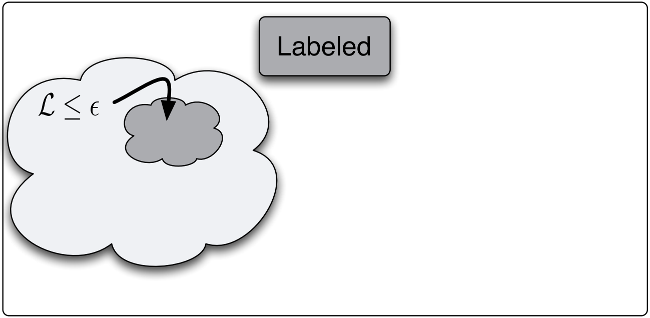
July 11, 2008

Supervised Learning



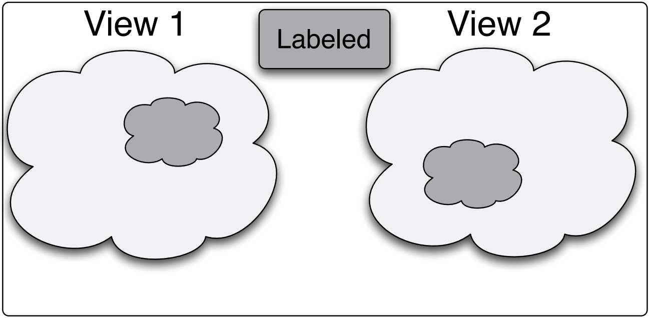
- We have a hypothesis class

Supervised Learning

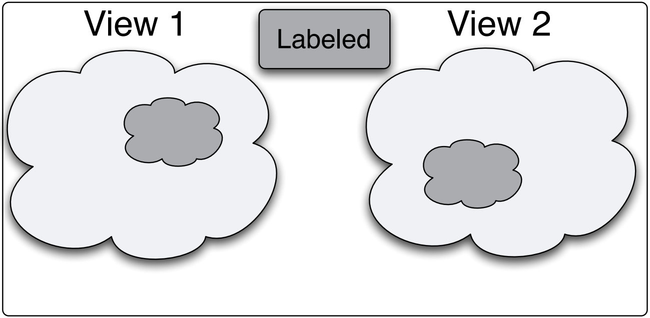


- We have a hypothesis class
  labeled data to choose hypothesis

Two View Learning

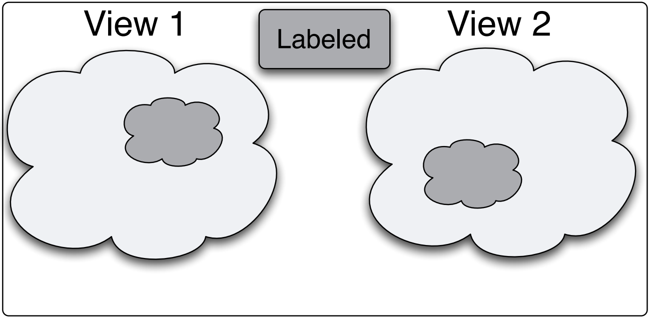


Two View Learning



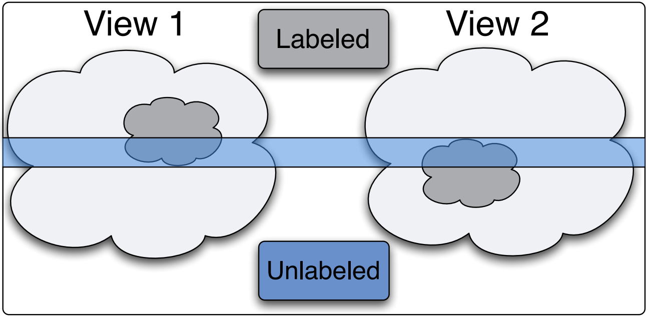
- each view performs well alone

Two View Learning



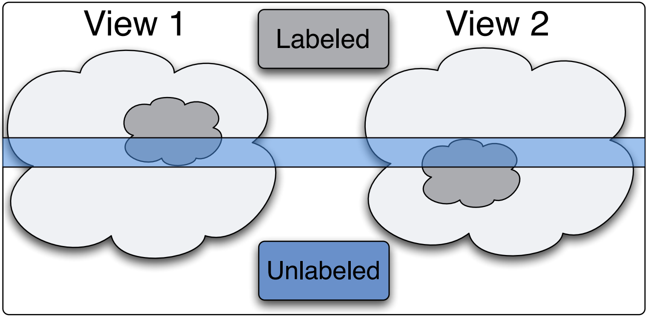
- each view performs well alone
  $\implies$ correct models should agree on unlabeled data

Two View Learning



- each view performs well alone
  $\Longrightarrow$ correct models should agree on unlabeled data
- views don't share too much extra information
  $\Longrightarrow$ can further reduce hypothesis space

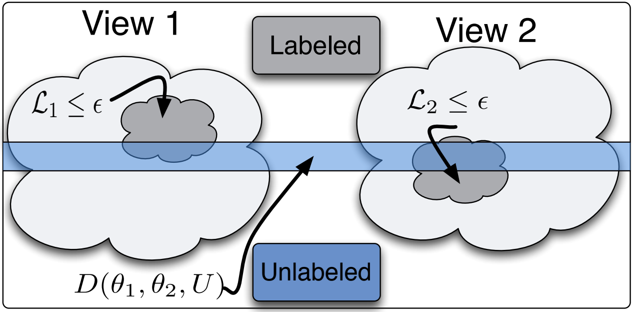
Two View Learning



- each view performs well alone
  $\Longrightarrow$ correct models should agree on unlabeled data
- views don't share too much extra information
  $\Longrightarrow$ can further reduce hypothesis space

Assumptions: (Blum & Mitchell, 1998; Balkan & Blum, 2006; Kakade and Foster, 2007)

How to learn models that agree

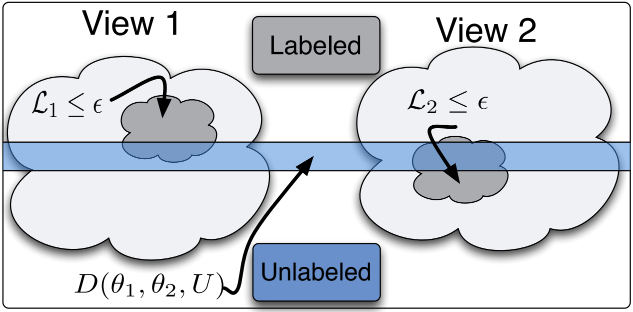


$$\min_{\theta_1,\theta_2} \mathcal{L}(\theta_1, L) + \mathcal{L}(\theta_2, L) + D(\theta_1, \theta_2, U)$$

- Learning probabilistic classifiers

Penn

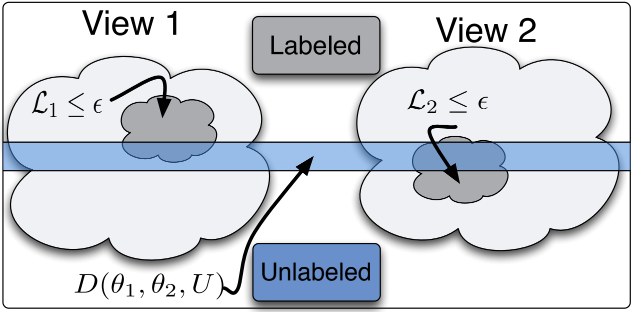
How to learn models that agree



$$\min_{\theta_1,\theta_2} \mathcal{L}(\theta_1, L) + \mathcal{L}(\theta_2, L) + D(\theta_1, \theta_2, U)$$

- Learning probabilistic classifiers
- $\mathcal{L}$: log-loss on labeled data L

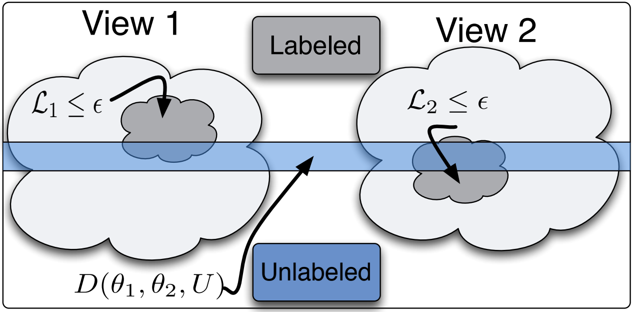
How to learn models that agree



$$\min_{\theta_1, \theta_2} \mathcal{L}(\theta_1, L) + \mathcal{L}(\theta_2, L) + D(\theta_1, \theta_2, U)$$

- Learning probabilistic classifiers
- $\mathcal{L}$: log-loss on labeled data L
- $\theta = \theta_1, \theta_2$: model paramters

How to learn models that agree



$$\min_{\theta_1, \theta_2} \mathcal{L}(\theta_1, L) + \mathcal{L}(\theta_2, L) + D(\theta_1, \theta_2, U)$$

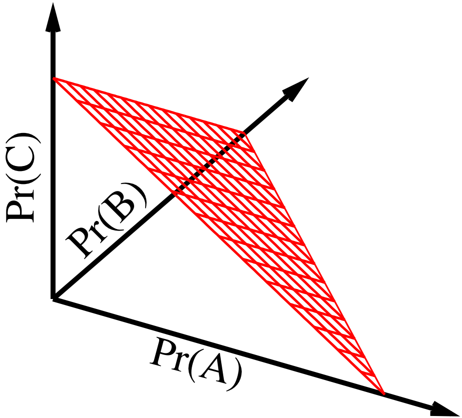
- Learning probabilistic classifiers
- $\mathcal{L}$: log-loss on labeled data L
- $\theta = \theta_1, \theta_2$: model paramters
- D: co-regularizer (encouraging agreement on unlabeled data U)

Co-regularizer

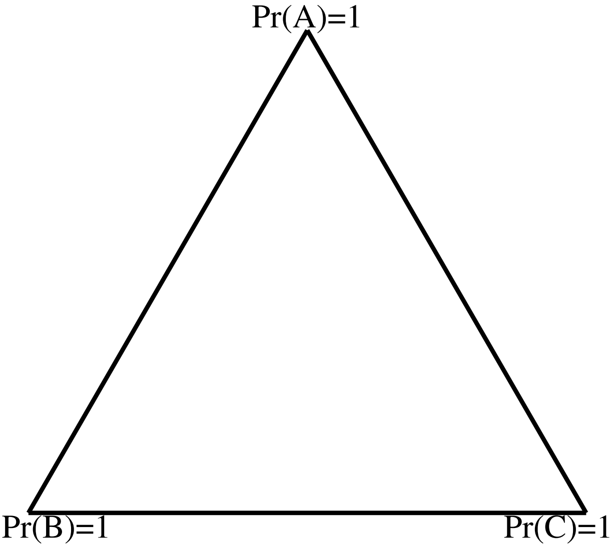
The coregularizer D...

- Based on KL distance to a consensus $q = \textbf{agree}(p_1, p_2)$
- p_i is distribution given by model i
- Illustrative to think in terms of consensus q

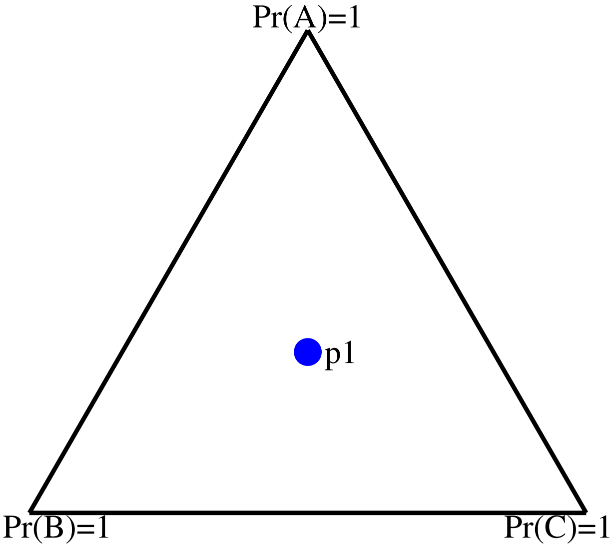
Probabilistic Coregularization



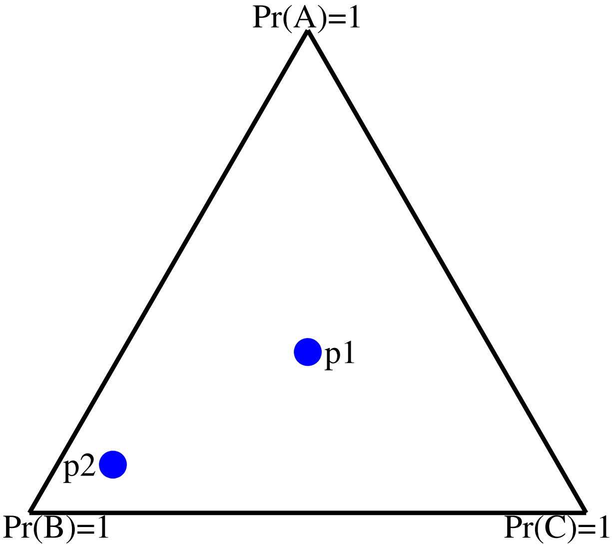
Probabilistic Coregularization



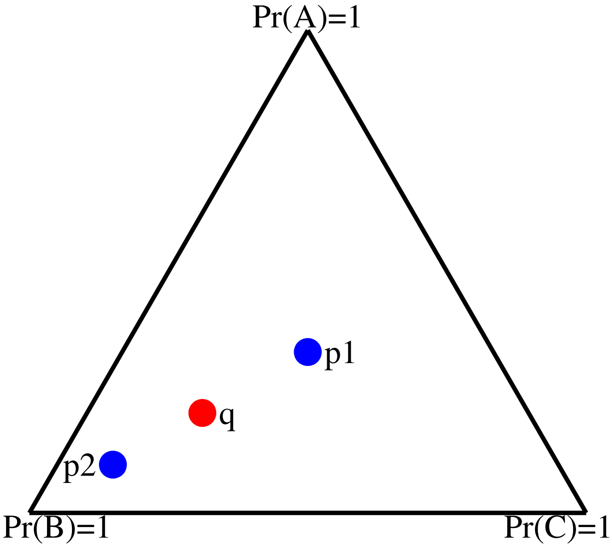
Probabilistic Coregularization



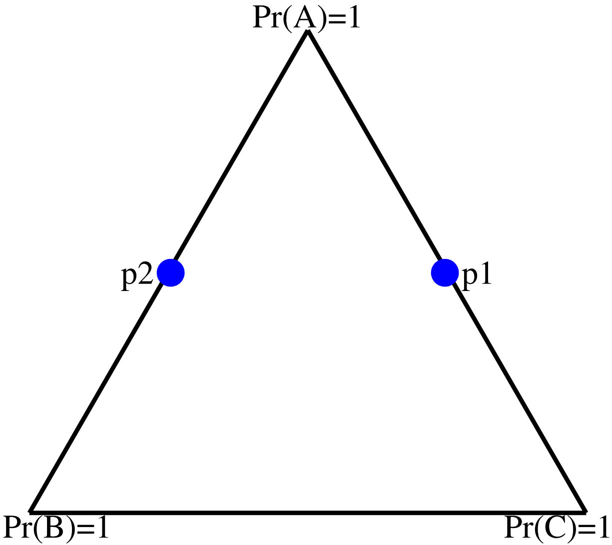
Probabilistic Coregularization



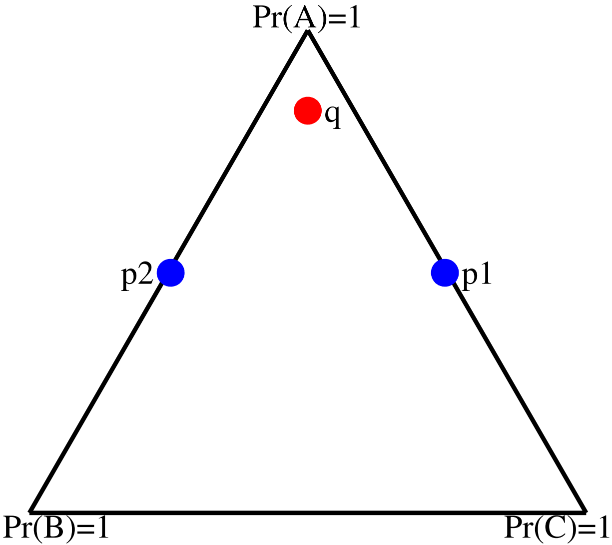
Probabilistic Coregularization



Probabilistic Coregularization



Probabilistic Coregularization



Our Agree Function

$$\textbf{agree}(p_1, p_2) = \arg\min_q \mathsf{KL}(q \,||\, p_1) + \mathsf{KL}(q \,||\, p_2)$$

Our Agree Function

$$\textbf{agree}(p_1, p_2) \quad = \quad \arg\min_q \mathrm{KL}(q \,||\, p_1) + \mathrm{KL}(q \,||\, p_2)$$

Theorem: $\textbf{agree}(p_1, p_2) \propto \sqrt{p_1 \times p_2}$

Algorithm

$\theta_i \leftarrow \min_\theta \mathcal{L}_i(\theta_i)$

for n iterations **do**

  $q(y_1|\mathbf{x}) \leftarrow \textbf{agree}(p_1(y_1|x), p_2(y_2|x)) \quad \forall x \in U$

  $\theta_i \leftarrow \min_{\theta} \; \mathcal{L}_i(\theta) - c \mathop{\mathbf{E}}_{x,y \sim U,q} [\log p_i(y_i|x;\theta)]$

end for

Algorithm

$\theta_i \leftarrow \mathsf{min}_\theta\, \mathcal{L}_i(\theta_i)$

for n iterations **do**

$\quad q(y_1|\mathbf{x}) \leftarrow \textbf{agree}(p_1(y_1|x), p_2(y_2|x)) \quad \forall x \in U$

$\quad \theta_i \leftarrow \min_{\theta} \; \mathcal{L}_i(\theta) - c \mathop{\mathbf{E}}_{x,y \sim U,q} [\log p_i(y_i|x;\theta)]$

end for

Algorithm

$\theta_i \leftarrow \mathsf{min}_\theta\, \mathcal{L}_i(\theta_i)$
for n iterations **do**
  $q(y_1|\mathbf{x}) \leftarrow \textbf{agree}(p_1(y_1|x), p_2(y_2|x)) \quad \forall x \in U$
  $\theta_i \leftarrow \min_\theta \; \mathcal{L}_i(\theta) - c \underset{x,y\sim U,q}{\mathbf{E}}[\log p_i(y_i|x;\theta)]$
end for

Algorithm

$\theta_i \leftarrow \min_\theta \mathcal{L}_i(\theta_i)$

for n iterations **do**

  $q(y_1|\mathbf{x}) \leftarrow \textbf{agree}(p_1(y_1|x), p_2(y_2|x)) \quad \forall x \in U$

  $\theta_i \leftarrow \min\limits_{\theta} \; \mathcal{L}_i(\theta) - c \mathop{\mathbf{E}}\limits_{x,y \sim U,q} [\log p_i(y_i|x;\theta)]$

end for

Algorithm

1: $\theta_i \leftarrow \min_\theta \mathcal{L}_i(\theta_i)$
2: **for** n iterations **do**
3: $\quad q(y_1|\mathbf{x}) \leftarrow \textbf{agree}(p_1(y_1|x), p_2(y_2|x)) \quad \forall x \in U$
4: $\quad \theta_i \leftarrow \min_\theta \; \mathcal{L}_i(\theta) - c \underset{x,y \sim U,q}{\mathbf{E}} [\log p_i(y_i|x;\theta)]$
5: **end for**

THEOREM: this minimizes co-regularized loss:

$$\mathcal{L}_1(\theta) + \mathcal{L}_2(\theta) + c\mathbf{E}_U \left[\min_q \mathrm{KL}(q \,||\, p_1) + \mathrm{KL}(q \,||\, p_1) \right] .$$

Algorithm

Theorem: this minimizes co-regularized loss:

$$\mathcal{L}_1(\theta) + \mathcal{L}_2(\theta) + c\mathbf{E}_U \left[\min_q \mathsf{KL}(q \,||\, p_1) + \mathsf{KL}(q \,||\, p_1) \right] .$$

Algorithm

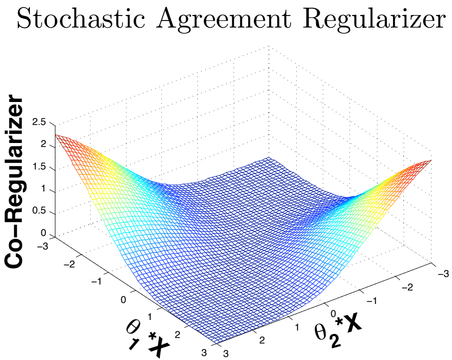
Theorem: this minimizes co-regularized loss:

$$\mathcal{L}_1(\theta) + \mathcal{L}_2(\theta) + c\mathbf{E}_U \left[\min_q \mathsf{KL}(q \,||\, p_1) + \mathsf{KL}(q \,||\, p_1) \right] .$$

$$= \mathcal{L}_1(\theta) + \mathcal{L}_2(\theta) + c\mathbf{E}_U \left[-\log \sum_y \sqrt{p(y;\theta_1)p(y;\theta_2)} \right] .$$

Bhattacharyya distance

Linear Model Coregularizer



- log-linear models:

$$p_i(1) \propto \exp(\theta_i \cdot x)$$
$$p_i(-1) \propto \exp(-\theta_i \cdot x)$$

Other aproaches

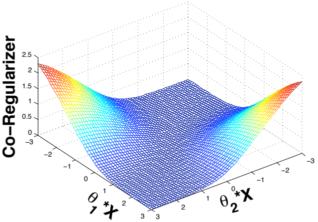
- CoBoosting (Collins and Singer, 1999), CoPerceptron (Brefeld et al., 2005)

Other aproaches

- CoBoosting (Collins and Singer, 1999), CoPerceptron (Brefeld et al., 2005)
    - Different regularized loss functions

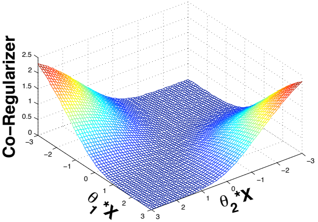
Different Loss Functions

Stochastic Agreement Regularizer

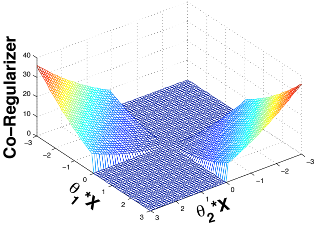


Different Loss Functions

Stochastic Agreement Regularizer

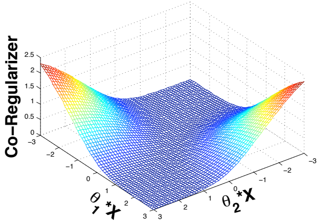


CoPerceptron

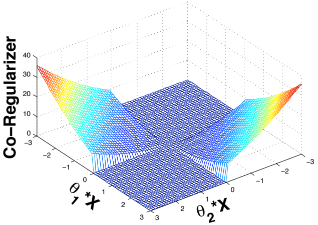


Different Loss Functions

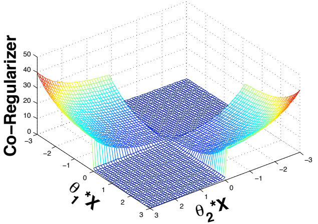
Stochastic Agreement Regularizer



CoPerceptron



CoBoosting



Other aproaches

- CoBoosting (Collins and Singer, 1999), CoPerceptron (Brefeld et al., 2005)
  - Different regularized loss functions

Other aproaches

- CoBoosting (Collins and Singer, 1999), CoPerceptron (Brefeld et al., 2005)
    - Different regularized loss functions
    - Hard assignment on unlabeled data

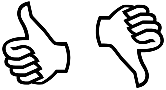
Other aproaches

- CoBoosting (Collins and Singer, 1999), CoPerceptron (Brefeld et al., 2005)
  - Different regularized loss functions
  - Hard assignment on unlabeled data
- Many others (Blum & Mitchell, 1998; Sindhwani et al., 2005; Kakade & Foster, 2007; Suzuki et al., 2007)

Sentiment Classification - Domain Adaptation (Blitzer et al, 2007)




- Product reviews from Amazon.com
  - Books, DVDs, Kitchen Appliances, Electronics
  - 2000 labeled, 3000 - 6000 unlabeled reviews per domain
- Binary classification problem
  - Positive if 4 stars or more, negative if 2 or less
- Transfer learning task
- Views: random split of features



Sentiment Classification

Domains	MIRA	SCL	CoBoost	CoPerc	SAR
books→dvds	77.2				
dvds→books	72.8				
books→electr	70.8				
electr→books	70.7				
books→kitchn	74.5				
kitchn→books	70.9				
dvds→electr	73.0				
electr→dvds	70.6				
dvds→kitchn	74.0				
kitchn→dvds	72.7				
electr→kitchn	84.0				
kitchn→electr	82.7				
Total					

Sentiment Classification

Domains	MIRA	SCL	CoBoost	CoPerc	SAR
books→dvds	77.2	-1.4			
dvds→books	72.8	6.9			
books→electr	70.8	5.1			
electr→books	70.7	**4.7**			
books→kitchn	74.5	4.4			
kitchn→books	70.9	-2.3			
dvds→electr	73.0	1.1			
electr→dvds	70.6	**5.6**			
dvds→kitchn	74.0	7.4			
kitchn→dvds	72.7	**4.2**			
electr→kitchn	84.0	1.9			
kitchn→electr	82.7	**4.1**			
Total		4			

Sentiment Classification

Domains	MIRA	SCL	CoBoost	CoPerc	SAR
books→dvds	77.2	-1.4	1.6		
dvds→books	72.8	6.9	7.0		
books→electr	70.8	5.1	**6.2**		
electr→books	70.7	**4.7**	0.3		
books→kitchn	74.5	4.4	3.5		
kitchn→books	70.9	-2.3	-1.1		
dvds→electr	73.0	1.1	2.3		
electr→dvds	70.6	**5.6**	2.9		
dvds→kitchn	74.0	7.4	5.0		
kitchn→dvds	72.7	**4.2**	-2.6		
electr→kitchn	84.0	1.9	1.0		
kitchn→electr	82.7	**4.1**	0.3		
Total		4	1		

Sentiment Classification

Domains	MIRA	SCL	CoBoost	CoPerc	SAR
books→dvds	77.2	-1.4	1.6	-1.7	
dvds→books	72.8	6.9	7.0	1.7	
books→electr	70.8	5.1	**6.2**	-1.5	
electr→books	70.7	**4.7**	0.3	-3.2	
books→kitchn	74.5	4.4	3.5	2.0	
kitchn→books	70.9	-2.3	-1.1	-4.3	
dvds→electr	73.0	1.1	2.3	-1.8	
electr→dvds	70.6	**5.6**	2.9	-7.3	
dvds→kitchn	74.0	7.4	5.0	4.3	
kitchn→dvds	72.7	**4.2**	-2.6	-12.2	
electr→kitchn	84.0	1.9	1.0	-0.7	
kitchn→electr	82.7	**4.1**	0.3	-2.2	
Total		4	1	0	

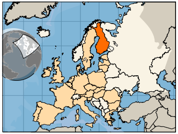
Sentiment Classification

Domains	MIRA	SCL	CoBoost	CoPerc	SAR
books→dvds	77.2	-1.4	1.6	-1.7	**2.6**
dvds→books	72.8	6.9	7.0	1.7	**8.5**
books→electr	70.8	5.1	**6.2**	-1.5	4.7
electr→books	70.7	**4.7**	0.3	-3.2	3.6
books→kitchn	74.5	4.4	3.5	2.0	**6.5**
kitchn→books	70.9	-2.3	-1.1	-4.3	**1.9**
dvds→electr	73.0	1.1	2.3	-1.8	**3.5**
electr→dvds	70.6	**5.6**	2.9	-7.3	2.4
dvds→kitchn	74.0	7.4	5.0	4.3	**8.8**
kitchn→dvds	72.7	**4.2**	-2.6	-12.2	0.1
electr→kitchn	84.0	1.9	1.0	-0.7	1.8
kitchn→electr	82.7	**4.1**	0.3	-2.2	2.8
Total		4	1	0	**6**

Sentiment Classification

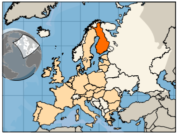
Domains	MIRA	SCL	CoBoost	CoPerc	SAR
books→dvds	77.2	-1.4	1.6	-1.7	**2.6**
dvds→books	72.8	6.9	7.0	1.7	**8.5**
books→electr	70.8	5.1	**6.2**	-1.5	4.7
electr→books	70.7	**4.7**	0.3	-3.2	3.6
books→kitchn	74.5	4.4	3.5	2.0	**6.5**
kitchn→books	70.9	-2.3	-1.1	-4.3	**1.9**
dvds→electr	73.0	1.1	2.3	-1.8	**3.5**
electr→dvds	70.6	**5.6**	2.9	-7.3	2.4
dvds→kitchn	74.0	7.4	5.0	4.3	**8.8**
kitchn→dvds	72.7	**4.2**	-2.6	-12.2	0.1
electr→kitchn	84.0	1.9	1.0	-0.7	1.8
kitchn→electr	82.7	**4.1**	0.3	-2.2	2.8
Total		4	1	0	**6**

Named entity disambiguation


- Classification of CoNLL 2003 named entites:
  - Person, location, organization, miscellaneous

Named entity disambiguation



- Classification of CoNLL 2003 named entites:
  - Person, location, organization, miscellaneous
- View 1 - Content
  - Features look only inside named entity
- View 2 - Context
  - Features look only outside named entity

Named entity disambiguation

Data size	mx-ent	SAR (RRE)
500	74.4	
1000	80.0	
2000	83.4	

- Prior variance and unlabeled weigh choose by cross-validation

Named entity disambiguation

Data size	mx-ent	SAR (RRE)
500	74.4	76.4 (9.2%)
1000	80.0	81.7 (8.5%)
2000	83.4	84.8 (8.4%)

- Prior variance and unlabeled weigh choose by cross-validation

How to generalize two view Idea

- Structured Output
- Partial Agreement Scenarios
- Both

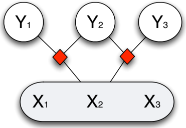
Structured Output

$$\mathbf{agree}(p_1, p_2) = \arg\min_q \mathrm{KL}(q \,||\, p_1) + \mathrm{KL}(q \,||\, p_2)$$

Structured Output

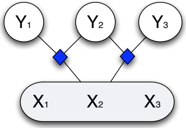
$$\textbf{agree}(p_1, p_2) = \arg\min_q \mathrm{KL}(q \,||\, p_1) + \mathrm{KL}(q \,||\, p_2)$$

View 1



$$p_1(y \mid x) \propto \prod_c \phi_1(y_c, x)$$

View 2

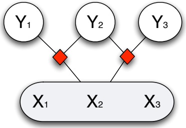


$$p_2(y \mid x) \propto \prod_c \phi_2(y_c, x)$$

Structured Output

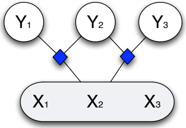
$$\textbf{agree}(p_1, p_2) = \arg\min_q \mathrm{KL}(q \,||\, p_1) + \mathrm{KL}(q \,||\, p_2)$$

View 1



$$p_1(y \mid x) \propto \prod_c \phi_1(y_c, x)$$

View 2



$$p_2(y \mid x) \propto \prod_c \phi_2(y_c, x)$$

Theorem: $$q_i(y) \propto \prod_c \sqrt{\phi_1(y_c, x)\phi_2(y_c, x)}$$

Structured Prediction

Small experiments on structured task.

- English NP-chunking from CoNLL 2000
- 500 sentences test data
- views are:
  - current word and POS tag
  - previous/next word and POS tag

Structured Prediction

size	Perc	coPerc	CRF	SAR(RRE)
10	69.4			
20	74.4			
50	80.1			
100	86.1			
200	89.3			
500	90.8			
1000	91.5			

Structured Prediction

size	Perc	coPerc	CRF	SAR(RRE)
10	69.4	71.2		
20	74.4	76.8		
50	80.1	84.1		
100	86.1	88.1		
200	89.3	**89.7**		
500	90.8	90.9		
1000	91.5	**91.8**		

Structured Prediction

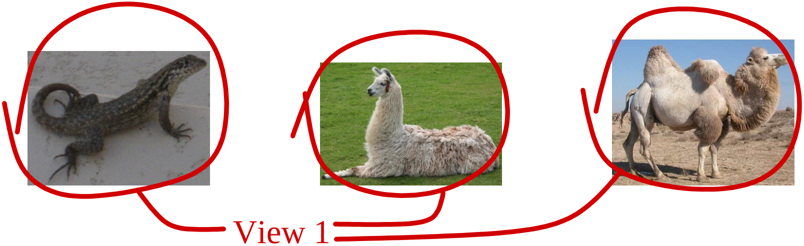
size	Perc	coPerc	CRF	SAR(RRE)
10	69.4	71.2	73.2	
20	74.4	76.8	79.4	
50	80.1	84.1	86.3	
100	86.1	88.1	88.5	
200	89.3	**89.7**	89.6	
500	90.8	90.9	**91.3**	
1000	91.5	**91.8**	91.6	

Structured Prediction

size	Perc	coPerc	CRF	SAR(RRE)
10	69.4	71.2	73.2	**78.2** (19%)
20	74.4	76.8	79.4	**84.2** (23%)
50	80.1	84.1	86.3	**86.9** (4%)
100	86.1	88.1	88.5	**88.9** (3%)
200	89.3	**89.7**	89.6	89.6 (0%)
500	90.8	90.9	**91.3**	90.6 (-8%)
1000	91.5	**91.8**	91.6	91.1 (-6%)

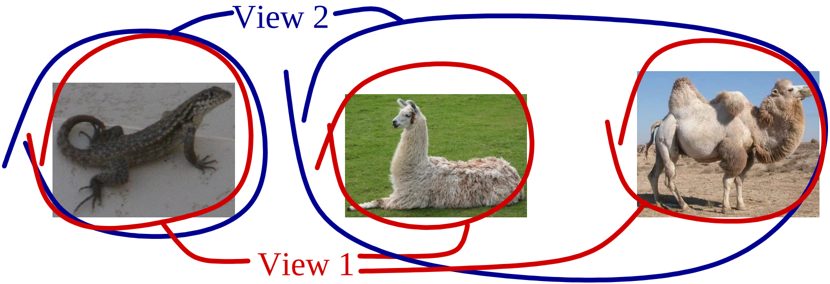
Different output spaces

Different output spaces



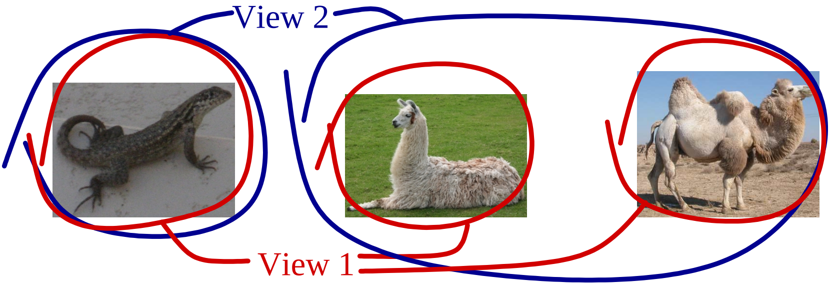
- Different Tag Sets between views

Different output spaces



- Different Tag Sets between views

Different output spaces



- Different Tag Sets between views
- Partial mapping between labels

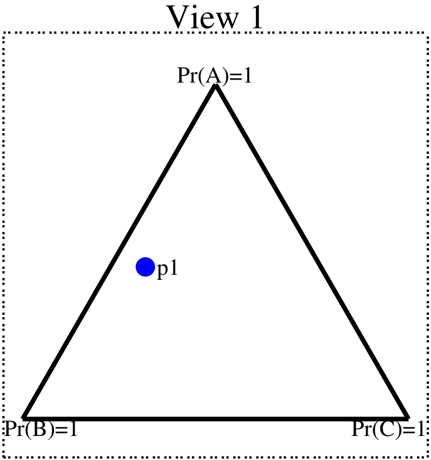
Algorithm

- Algorithm stays the same
- Consensus changes:

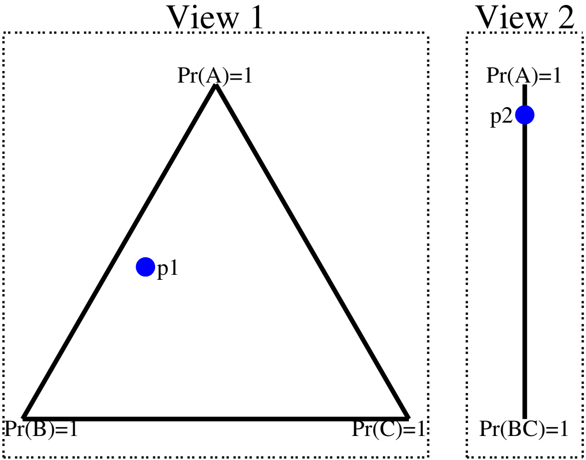
$$
\begin{aligned}
\textbf{agree}(p_1, p_2) \quad &= \quad \operatorname*{arg\,min}_q \mathsf{KL}\left(q(y_1, y_2) \,||\, p_1(y_1)p_2(y_2)\right) \\
&\quad\quad \textbf{s.t.} \quad q(z_1, z_2) = q(z_1)q(z_2) \\
&\quad\quad \textbf{where} \quad z_i = g(y_i)
\end{aligned}
$$

$g_i(y_i)$: mapping from output of model i to common space

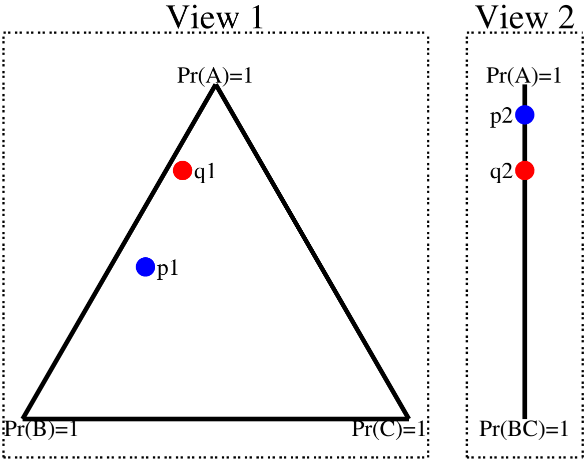
Different output spaces



Different output spaces



Different output spaces



Summary

- New two-view learning algorithm

Summary

- New two-view learning algorithm
- Probabilistic interpretation

Summary

- New two-view learning algorithm
- Probabilistic interpretation
- Generalizes naturally to structured data

Summary

- New two-view learning algorithm
- Probabilistic interpretation
- Generalizes naturally to structured data, partial agreement

Summary

- New two-view learning algorithm
- Probabilistic interpretation
- Generalizes naturally to structured data, partial agreement
- Empirically better than non-smooth alternatives

Thanks!