An RKHS for Multi-View Learning and Manifold Co-Regularization

[1]Vikas Sindhwani and [2]*David S. Rosenberg*

July 7, 2008

[1]IBM T. J. Watson Research Center, Yorktown Heights, NY
[2]Department of Statistics, University of California, Berkeley, CA

Outline

Semi-Supervised Learning

- Input space $\mathcal{X}$
- Output space $\mathcal{Y}$
- Distribution $P_{\mathcal{X}\times\mathcal{Y}}$ on $\mathcal{X} \times \mathcal{Y}$
- Training Data:

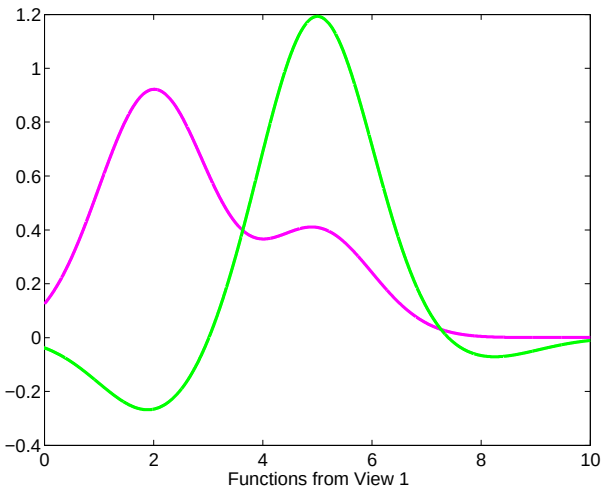
$$\begin{aligned} \text{labeled data :} \quad & (x_1, y_1), (x_2, y_2), \ldots, (x_\ell, y_\ell) \text{ i.i.d. } P_{\mathcal{X}\times\mathcal{Y}} \\ \text{unlabeled data :} \quad & x_{\ell+1}, x_{\ell+2}, \ldots, x_{\ell+u} \text{ i.i.d. } P_{\mathcal{X}} \end{aligned}$$

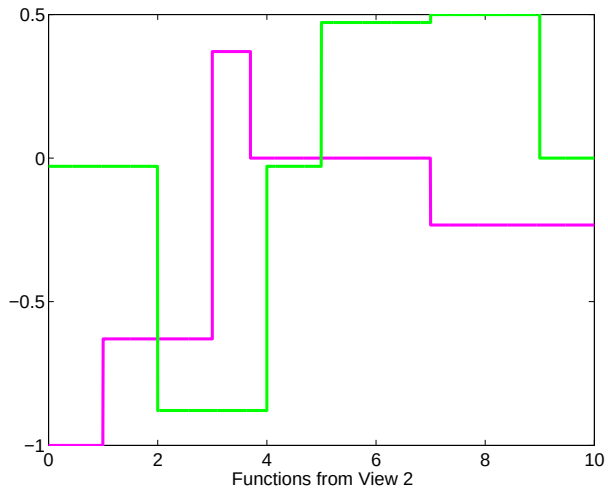
- Goal: Predict y given x

Two-View Learning

Basic Framework

- Two sets of prediction functions (the views): $\mathcal{F}$ and $\mathcal{G}$





- Want to find good functions $f^* \in \mathcal{F}$ and $g^* \in \mathcal{G}$
- Take final prediction to be

$$\varphi^* = \frac{1}{2}\left(f^* + g^*\right)$$

Two-View Learning

Motivating Idea

- Write $\eta(x)$ for the "target" prediction function.

Motivating Idea of Multi-View Learning

- Given "good" functions $f^* \in \mathcal{F}$ and $g^* \in \mathcal{G}$ s.t.

$$f^*(x) \approx \eta(x) \qquad g^*(x) \approx \eta(x),$$

- then

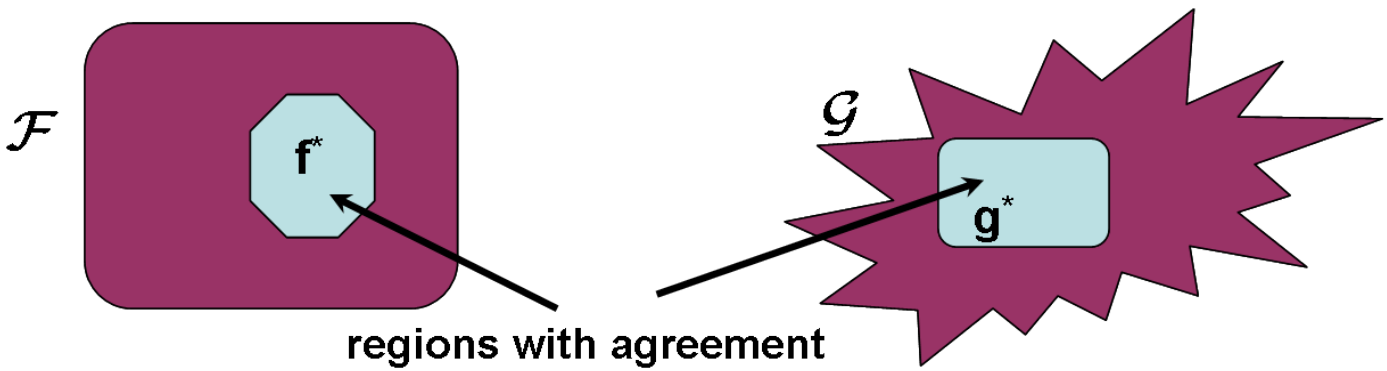
$$f^*(x) \approx g^*(x)$$

.

$\Longrightarrow$ "All good prediction functions are about equal."

Two-View Learning

- Restrict search to "agreement regions"



- How to find agreement regions?

Two-View Learning with L^2 Co-Regularization

- RKHS's $(\mathcal{F}, k_{\mathcal{F}})$ and $(\mathcal{G}, k_{\mathcal{G}})$
- Objective Function

$$(f^*, g^*) := \underset{f \in \mathcal{F}, g \in \mathcal{G}}{\arg\min} \; \hat{L}\left(\frac{f+g}{2}\right) + \gamma_{\mathcal{F}}||f||^2_{\mathcal{F}} + \gamma_{\mathcal{G}}||g||^2_{\mathcal{G}} + \lambda \sum_{i=\ell+1}^{\ell+u} |f(x_i) - g(x_i)|^2$$

- By Representer theorem,
  - If $\hat{L}(h) = \sum_{i=1}^{\ell} (h(x_i) - y_i)^2$, solve by minimizing a quadratic
  - If $\hat{L}(h) = \sum_{i=1}^{\ell} (1 - h(x_i)y_i)_+$, get a QP

Co-Regularization Reformulated

- Final prediction is

$$h^* = \frac{f^* + g^*}{2}$$

- Define

$$\tilde{\mathcal{H}} = \left\{ \tilde{h} = \frac{f+g}{2} \ : \ f \in \mathcal{F}, g \in \mathcal{G} \right\}$$

- Can show that

$$h^* = \arg\min_{\tilde{h} \in \tilde{\mathcal{H}}} \left[\hat{L}\left(\tilde{h}\right) + \|\tilde{h}\|^2_{\tilde{\mathcal{H}}} \right]$$

where

$$\|\tilde{h}\|^2_{\tilde{\mathcal{H}}} := \min_{\substack{f \in \mathcal{F}, g \in \mathcal{G} \\ \text{s.t. } \tilde{h} = \frac{1}{2}(f+g)}} \gamma_{\mathcal{F}} ||f||^2_{\mathcal{F}} + \gamma_{\mathcal{G}} ||g||^2_{\mathcal{G}} + \lambda \sum_{i=\ell+1}^{\ell+u} |f(x_i) - g(x_i)|^2$$

The Co-Regularized RKHS

Theorem

- *There exists an inner product on* $\tilde{\mathcal{H}}$ *for which* $\tilde{\mathcal{H}}$ *is an RKHS with norm* $\|\cdot\|_{\tilde{\mathcal{H}}}$ *(defined earlier).*
- *The reproducing kernel* $\tilde{k} : \mathcal{X} \times \mathcal{X} \to \mathbf{R}$ *is given by*

$$\tilde{k}(x, z) = \gamma_{\mathcal{F}}^{-1} k_{\mathcal{F}}(x, z) + \gamma_{\mathcal{G}}^{-1} k_{\mathcal{G}}(x, z) - \lambda d_x' H d_z$$

where

$$d_x = \gamma_{\mathcal{F}}^{-1} k_{\mathcal{F}}(x, \mathit{unlab}) - \gamma_{\mathcal{G}}^{-1} k_{\mathcal{G}}(x, \mathit{unlab})$$

and

$$H = \left(I + \frac{\lambda}{\gamma_{\mathcal{F}}} K_{\mathcal{F}}(\mathit{unlab}, \mathit{unlab}) + \frac{\lambda}{\gamma_{\mathcal{G}}} K_{\mathcal{G}}(\mathit{unlab}, \mathit{unlab}) \right)^{-1}$$

What have we gained?

- Can plug-in this "multi-view" kernel to any kernel algorithm
  - multiview semi-supervised KPCA
  - multiview semi-supervised ***
- Reproduce and improve results in [R. & Bartlett, AIStats'07]
  - Rademacher complexity bounds follow from standard theory RKHS theory
  - Localized Rademacher complexity theory improves the generalization bounds

Corollary

Rademacher Complexity of the L^2 Co-Regularized Function Class $\mathcal{J}$

- For simplicity, take $\gamma_{\mathcal{F}} = \gamma_{\mathcal{G}} = 1$.

Theorem

For the L^2 co-regularized function class $\mathcal{J}$,

$$\frac{1}{\sqrt[4]{2}} \frac{U}{\ell} \leq \hat{R}_\ell(\mathcal{J}) \leq \frac{U}{\ell},$$

where

$$U = \sqrt{\mathrm{tr}\left[K_{\mathcal{F}}(lab, lab)\right] + \mathrm{tr}\left[K_{\mathcal{G}}(lab, lab)\right] - \Delta(\lambda)},$$

and $\Delta(\lambda) \geq 0$ (explicit form given on next slide).

Interpretation of Complexity Reduction Term

Theorem (Continued)

For $\lambda = 0$,

$$\Delta(\lambda) = 0.$$

For $\lambda > 0$,

$$\Delta(\lambda) \quad = \quad \sum_{i=1}^{\ell} d^2 \left[k_{\mathcal{F}}(x_i, unlab) \, , \; k_{\mathcal{G}}(x_i, unlab) \right] ,$$

where

- $d(\cdot, \cdot)$ *is a metric on* $\mathbf{R}^u$ *defined by*

$$d^2(c, f) = (c - f)' \left(\lambda^{-1} I + M \right)^{-1} (c - f),$$

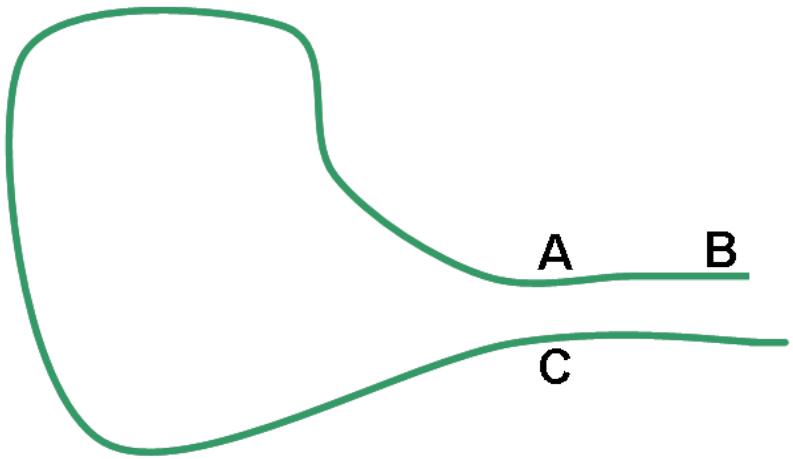
and

$$M := K_{\mathcal{F}}(unlab, unlab) + K_{\mathcal{G}}(unlab, unlab)$$

Manifold Regularization

Geodesic Distance

- Say data lie on a manifold $\mathcal{M}$ in $\mathbf{R}^2$



- In the ambient space $\mathbf{R}^2$: $d_{\mathcal{A}}(A, C) < d_{\mathcal{A}}(A, B)$
- On the manifold, $d_{\mathcal{M}}(A, B) \ll d_{\mathcal{M}}(A, C)$
- Hypothesis: Good prediction function is smooth w.r.t. "manifold distance"
- Lots of work with this idea
  - "Manifold learning": ISOMAP (Tenebaum et al., *Science* 2000), LLE (Roweis & Saul, *Science* 2000), Laplacian Eigenmaps (Belkin & Niyogi, NIPS 2001)
  - "Manifold transduction": Joachims et al. (ICML 2003), Belkin et al. (JLMR 2006), Zhou et al. (NIPS 2004), Sindhwani et al. (IMCL 2005)

Manifold Co-Regularization

The Two Views

- *Manifold regularization* (Belkin et al., 2006)

$$Q(f) \quad = \quad \hat{L}(f) + \gamma_{\mathcal{A}} \|f\|_{\mathcal{A}}^2 + \gamma_{\mathcal{I}} \boldsymbol{f}^T M \boldsymbol{f}$$

- Ambient View $\mathcal{H}_{\mathcal{A}}$: norm controls smoothness on $\mathcal{X}$

- Intrinsic View: controls smoothness w.r.t. $\mathcal{G} = \{x_1, \ldots x_\ell, x_{\ell+1}, \ldots, x_{\ell+u}\}$

- *Manifold co-regularization*
  Find $(f^*, \boldsymbol{g}^*) \in \mathcal{H}_{\mathcal{A}} \times \mathbf{R}^{\ell+u}$ minimizing

$$Q(f, \boldsymbol{g}) \quad = \quad \hat{L}(f, g) + \gamma_{\mathcal{A}} \|f\|_{\mathcal{A}}^2 + \gamma_{\mathcal{I}} \boldsymbol{g}^T M \boldsymbol{g} + \lambda \sum_{i=\ell+1}^{\ell+u} [f(x_i) - \boldsymbol{g}_i]^2$$

Experiment: Manifold Regularization vs CoMR

- 5 Datasets
  - LINES
  - G50C: 50-dim Gaussians s.t. Bayes error is 5%
  - USPS: 10-class digit recognition
  - COIL20: 32 x 32 gray scale images of 20 objects, varying angles
  - PCMAC: from 20 newsgroup dataset

Table: Datasets with d features and c classes. 10 random data splits were created with l labeled, u unlabeled, t test, and v validation examples.

Dataset	d	c	l	u	t	v
LINES	2	2	2	500	250	250
G50C	50	2	50	338	112	50
USPST	256	10	50	1430	477	50
COIL20	241	20	40	1320	40	40
PCMAC	7511	2	50	1385	461	50

Manifold Regularization vs CoMR

Table: Error Rates on Test Data

Dataset	MR	CoMR
LINES	7.7 (1.2)	**1.0** (1.5)
G50C	5.8 (2.8)	5.5 (2.3)
USPST	18.2 (1.5)	**14.1** (1.6)
COIL20	23.8 (11.1)	**14.8** (8.8)
PCMAC	11.9 (3.4)	**8.9** (2.6)

Table: $\lambda = 1$ for CoMR. Linear kernel for LINES, RBF for others.

Dataset	nn	σ	p	MR	CoMR
				γ_1, γ_2	γ_1, γ_2
LINES	10	–	1	$0.01, 10^{-6}$	$10^{-4}, 100$
G50C	50	17.5	5	$1, 100$	$10, 10$
USPST	10	9.4	2	$0.01, 0.01$	$10^{-6}, 10^{-4}$
COIL20	2	0.6	1	$10^{-4}, 10^{-6}$	$10^{-6}, 10^{-6}$
PCMAC	50	2.7	5	$10, 100$	$1, 10$

Generalization of RKHS Theorem

- RKHS views $\mathcal{H}^1, \ldots, \mathcal{H}^m$
- Arbitrary (fixed) linear combination of views: For $a_1, \ldots, a_m \in \mathbf{R}$,

$$\tilde{\mathcal{H}} := \left\{ a_1 f^1 + \cdots + a_m f^m \mid f^i \in \mathcal{H}^h, i = 1, \ldots, m \right\}$$

- Arbitrary quadratic "coupling" term.

$$\arg\min_{\varphi \in \tilde{\mathcal{H}}} \min_{\substack{f^1, \ldots, f^m \text{ s.t.} \\ \varphi = a_1 f^1 + \cdots a_m f^m}} L(\varphi) + \sum_{i=1}^{m} \gamma_i \|f^i\|^2_{\mathcal{H}^i} + \lambda \underline{\boldsymbol{f}}^T M \underline{\boldsymbol{f}} \qquad (1)$$

where

$$\underline{\boldsymbol{f}} := \left(f^1(x_1), \ldots, f^1(x_n), \ldots, f^m(x_1), \ldots, f^m(x_n)\right)^T$$

Generalization of Theorem

Theorem (Rosenberg et al. 2008)

- *There exists an inner product on $\tilde{\mathcal{H}}$ for which $\tilde{\mathcal{H}}$ is an RKHS with norm*

$$\|\varphi\|_{\tilde{\mathcal{H}}}^2 = \min_{\substack{f^1,\ldots,f^m \text{ s.t.} \\ \varphi = a_1 f^1 + \cdots a_m f^m}} \sum_{i=1}^{m} \gamma_i \|f^i\|_{\mathcal{H}^i}^2 + \lambda \underline{\boldsymbol{f}}^T M \underline{\boldsymbol{f}} \qquad (2)$$

- *The reproducing kernel $\tilde{k} : \mathcal{X} \times \mathcal{X} \to \mathbf{R}$ is given by*

$$\tilde{k}(z,x) = \sum_{j=1}^{m} \frac{a_j^2}{\gamma_j} k^j(z,x) - \lambda \underline{\boldsymbol{k}}_x^T \underline{A} \left(I + \lambda M \mathcal{K}\right)^{-1} M \underline{A} \underline{G}^{-1} \underline{\boldsymbol{k}}_z,$$

where

$$\mathcal{K} = \operatorname{diag}\left(K^1, \ldots, K^m\right) \in \mathbf{R}^{nm \times nm} \qquad \underline{\boldsymbol{k}}_x = \left(k^1(x_1,x), \ldots, k^1(x_n,x), \ldots, k^m(x_1,x), \ldots\right.$$

$$\underline{A} = \operatorname{diag}(\underbrace{a_1, \ldots, a_1}, \ldots, \underbrace{a_m, \ldots, a_m}) \qquad \underline{G} = \operatorname{diag}(\underbrace{\gamma_1, \ldots, \gamma_1}, \ldots, \underbrace{\gamma_m, \ldots, \gamma_m})$$

Summary and Conclusion

- Presented the co-regularized RKHS with closed form kernel
  - Easily convert any kernel method into a "multi-view semi-supervised" method
  - Can apply standard RKHS theory to get improved generalization bounds
- Presented Manifold Co-Regularization
  - reinterpretation of manifold regularization
  - improved empirical results
- Generalization to
  - multiple views
  - arbitrary linear combinations
  - quadratic coupling term