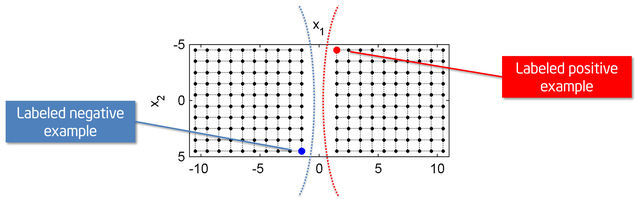
Semi-Supervised Learning with Max-Margin Graph Cuts

Branislav Kveton, Intel Labs Santa Clara
Michal Valko, University of Pittsburgh
Ali Rahimi, Intel Labs Berkeley
Ling Huang, Intel Labs Berkeley

Motivation

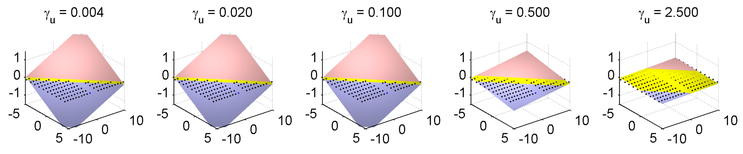
- Manifold regularization of SVMs (Belkin et al., 2006) is a state-of-art approach to semi-supervised max-margin learning
- Yet, the two clusters of points in the data adjacency graph:



cannot be separated by manifold regularization of linear SVMs:

$$\min_{\alpha_1,\alpha_2} \sum_{i \in l} V(f, \mathbf{x}_i, y_i) + \gamma[\alpha_1^2 + \alpha_2^2] + \gamma_u \mathbf{f}^\top L \mathbf{f}$$

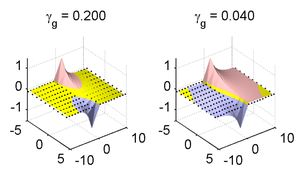
for any setting of the regularization parameter γ_u



Max-Margin Graph Cuts

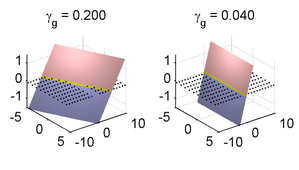
- Manifold regularization of SVMs (Belkin et al., 2006) combines the ideas of the harmonic function solution and semi-supervised SVMs in a single convex objective
- We investigate a different way of combining these two ideas
- First, we compute the regularized harmonic function solution on the data adjacency graph W

$$\min_{\boldsymbol{\ell} \in \mathbb{R}^n} \quad \boldsymbol{\ell}^\top (L + \gamma_g I) \boldsymbol{\ell}$$
$$\text{s.t.} \quad \ell_i = y_i \text{ for all } i \in l$$



- Second, we learn a max-margin discriminator which is conditioned on the labels induced by the harmonic function solution

$$\min_{f \in \mathcal{H}_K} \sum_{i:|\ell_i^*| \geq \varepsilon} V(f, \mathbf{x}_i, \text{sgn}(\ell_i^*)) + \gamma \|f\|_K^2$$
$$\text{s.t.} \quad \boldsymbol{\ell}^* = \arg\min_{\boldsymbol{\ell} \in \mathbb{R}^n} \boldsymbol{\ell}^\top (L + \gamma_g I) \boldsymbol{\ell}$$
$$\text{s.t. } \ell_i = y_i \text{ for all } i \in l$$



- Favorable properties of our solution:
  - Incorporates the kernel trick
  - Takes advantage of sparse data adjacency matrices
  - Its generalization error can be bounded

Existing Work

- In comparison to semi-supervised SVMs with the hat loss on unlabeled examples (Bennett & Demiriz, 1999):

$$\min_f \sum_{i \in l} V(f, \mathbf{x}_i, y_i) + \gamma \|f\|_K^2 + \gamma_u \sum_{i \in u} \widehat{V}(f, \mathbf{x}_i)$$

our optimization problem is convex and can be solved optimally

- In comparison to manifold regularization of SVMs (Belkin et al., 2006):

$$\min_{f \in \mathcal{H}_K} \sum_{i \in l} V(f, \mathbf{x}_i, y_i) + \gamma \|f\|_K^2 + \gamma_u \mathbf{f}^\top L \mathbf{f}$$

our algorithm yields better (and more intuitive) results when the class of discriminators is severely restricted, such as linear functions

Theoretical Analysis

- The generalization error of max-margin graph cuts can be bounded by combining existing transductive (Belkin et al., 2004; Cortes et al., 2008) and inductive (Vapnik, 1995) bounds

Proposition 2. *Let f be from a function class with the VC dimension h and n_ε be the number of examples such that $|\ell_i^*| < \varepsilon$. Then the inequality:*

$$R_P(f) \leq \frac{1}{n} \sum_{i:|\ell_i^*| \geq \varepsilon} \mathcal{L}(f(\mathbf{x}_i), \text{sgn}(\ell_i^*)) + \frac{2\varepsilon n_\varepsilon}{n} + \widehat{R}_P^W(\boldsymbol{\ell}^*) + \Delta_T(\beta, n_l, \delta) + \Delta_I(h, n, \eta)$$

holds with probability $1 - (\eta + \delta)$.

Experimental Results

- As γ_g increases, max-margin graph cuts gradually interpolate between semi-supervised learning on all data and supervised learning on labeled examples
- Max-margin graph cuts can solve problems which cannot be solved by manifold regularization of SVMs
- Max-margin graph cuts and manifold regularization of SVMs with RBF kernels yield similar results
- Max-margin graph cuts outperform manifold regularization of SVMs on 29 out of 36 experiments on UCI ML repository datasets

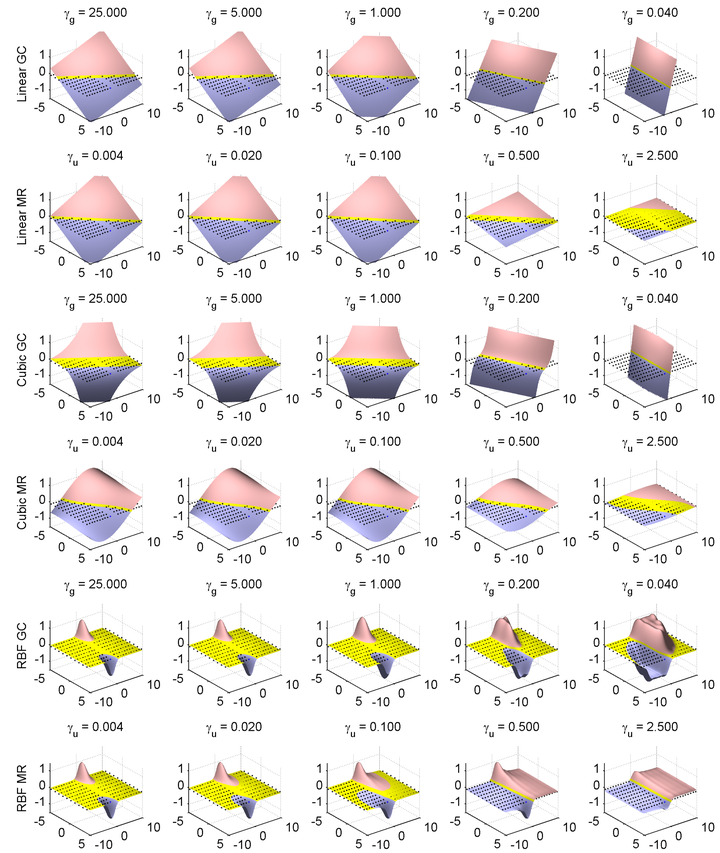


Figure 1: Linear, cubic, and RBF decision boundaries learned by manifold regularization of SVMs (MR) and max-margin graph cuts (GC). The regularization parameter γ_g is set as $\gamma_g = \gamma / \gamma_u$, $\gamma = 0.1$, and $\varepsilon = 0.1$. The pink and blue colors denote parts of the feature space where the discriminators f are positive and negative, respectively. The yellow color marks regions where $|f(\mathbf{x})| < 0.05$.

Dataset	L	Misclassification errors [%]								
		Linear kernel			Cubic kernel			RBF kernel		
		SVM	MR	GC	SVM	MR	GC	SVM	MR	GC
Letter recognition	1	18.90	30.94	15.79	20.54	25.96	17.45	20.06	17.61	16.01
	2	12.92	28.45	10.79	12.18	18.34	10.90	13.52	13.10	11.83
	5	8.21	27.13	5.65	5.49	18.77	4.80	6.81	8.06	5.65
	10	6.51	25.45	3.96	4.17	14.03	2.96	4.95	6.14	3.32
Digit recognition	1	7.06	9.59	6.88	9.62	5.29	8.55	8.22	6.36	7.65
	2	4.87	7.97	4.60	6.06	5.06	5.09	6.17	4.21	5.61
	5	2.97	3.68	2.29	3.04	2.27	2.36	2.74	2.29	2.19
	10	1.70	2.86	1.59	1.87	1.60	1.74	1.68	1.75	1.35
Image segmentation	1	14.02	11.81	10.27	23.30	12.02	14.10	14.02	11.60	9.51
	2	8.54	10.87	7.69	14.28	13.07	7.73	9.06	8.93	7.34
	5	4.73	7.83	4.49	8.32	8.79	7.17	5.87	5.43	5.31
	10	3.30	6.26	3.28	3.65	6.64	3.60	3.84	4.81	3.73

Figure 2: Comparison of SVMs, max-margin graph cuts (GC), and manifold regularization of SVMs (MR) on three datasets from the UCI ML repository. The fraction of labeled examples L varies from 1 to 10 percent.