Geometry of Semi-Supervised Learning

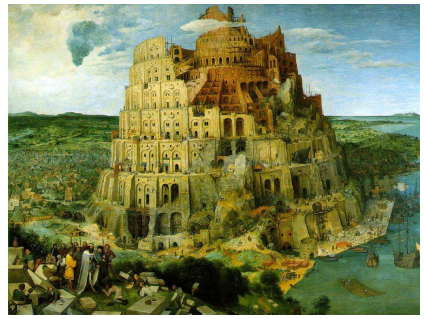
Mikhail Belkin
Partha Niyogi

The University of Chicago

Collaborators: V. Sindhwani, I. Matveeva, D. Surendran

Machine learning vs human learning

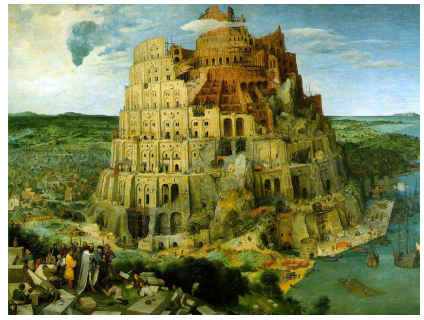
Human learning:



- Complex stimuli.
- Impoverished inputs.
- Robust.
- Extensive use of prior knowledge.
- Learning through mostly unlabeled data. Inference from few labeled examples.

Machine learning vs human learning

Human learning:



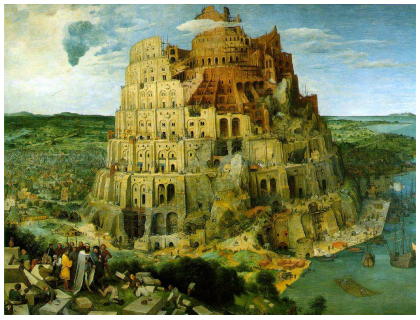
- Complex stimuli.
- Impoverished inputs.
- Robust.
- Extensive use of prior knowledge.
- Learning through mostly unlabeled data. Inference from few labeled examples.

water

?

Machine learning vs human learning

Human learning:



- Complex stimuli.
- Impoverished inputs.
- Robust.
- Extensive use of prior knowledge.
- Learning through mostly unlabeled data. Inference from few labeled examples.

water

?

Machine learning: (almost) none of the above!

Unlabeled data.

Unlabeled data

Reasons to use unlabeled data in inference:

- Pragmatic:

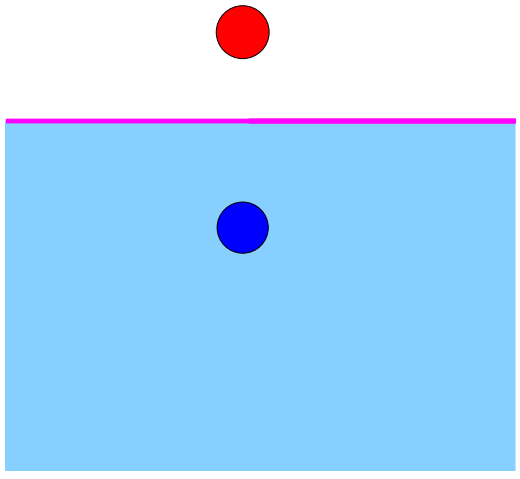
Unlabeled data is everywhere. Need a way to use it.

- Philosophical:

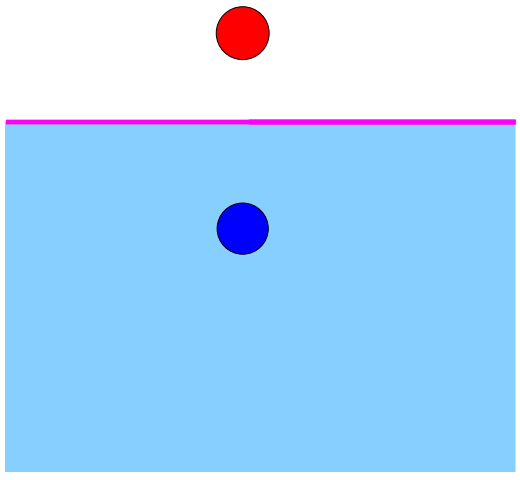
The brain uses unlabeled data.

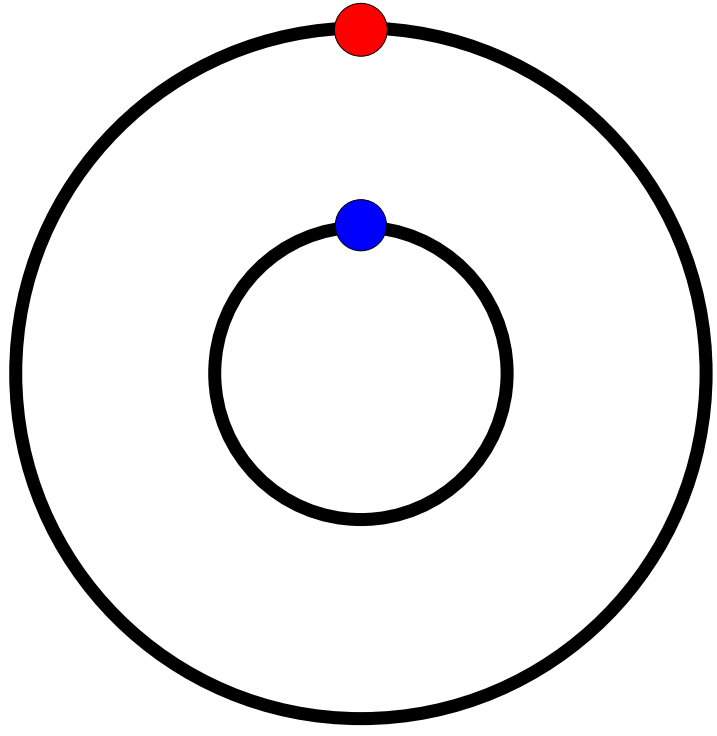
Intuition

Intuition

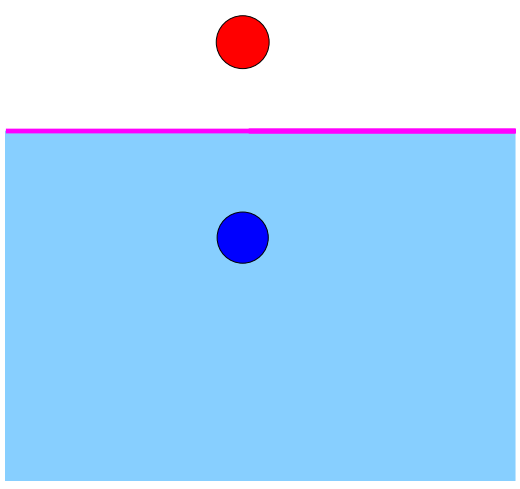


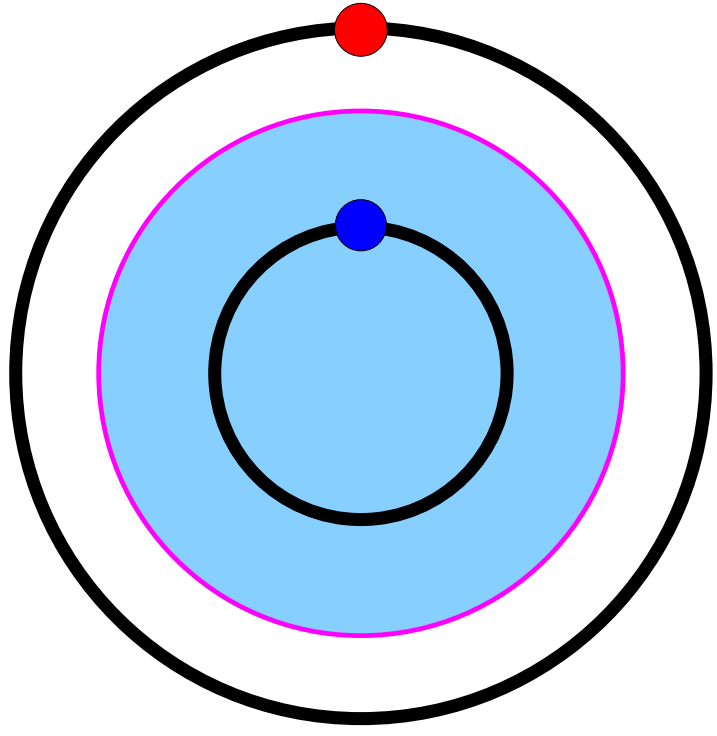
Intuition





Intuition

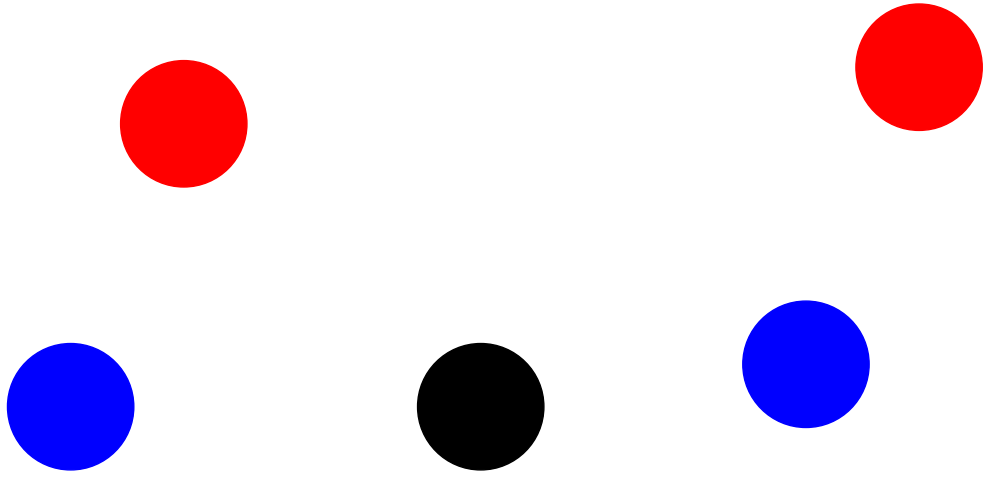




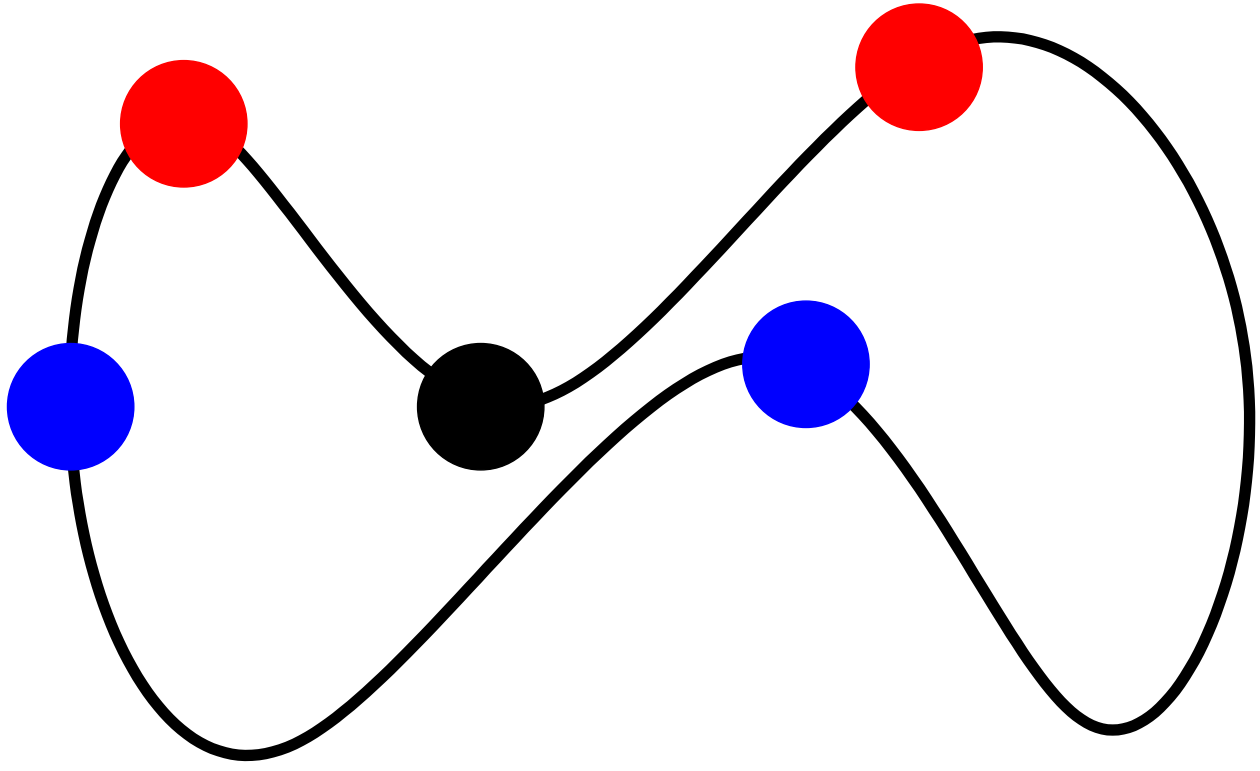
Geometry of data changes our notion of similarity.

Manifold assumption

Manifold assumption

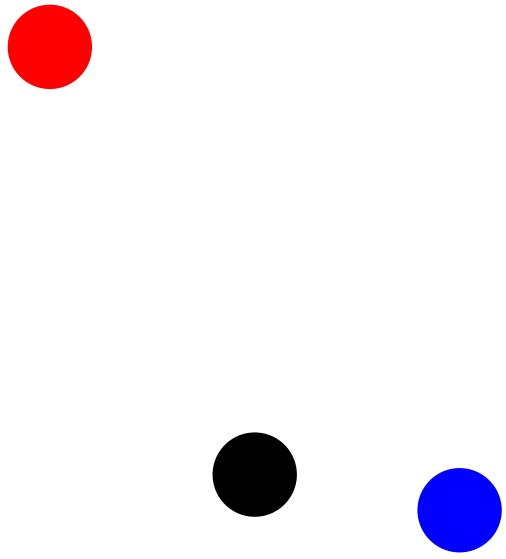


Manifold assumption

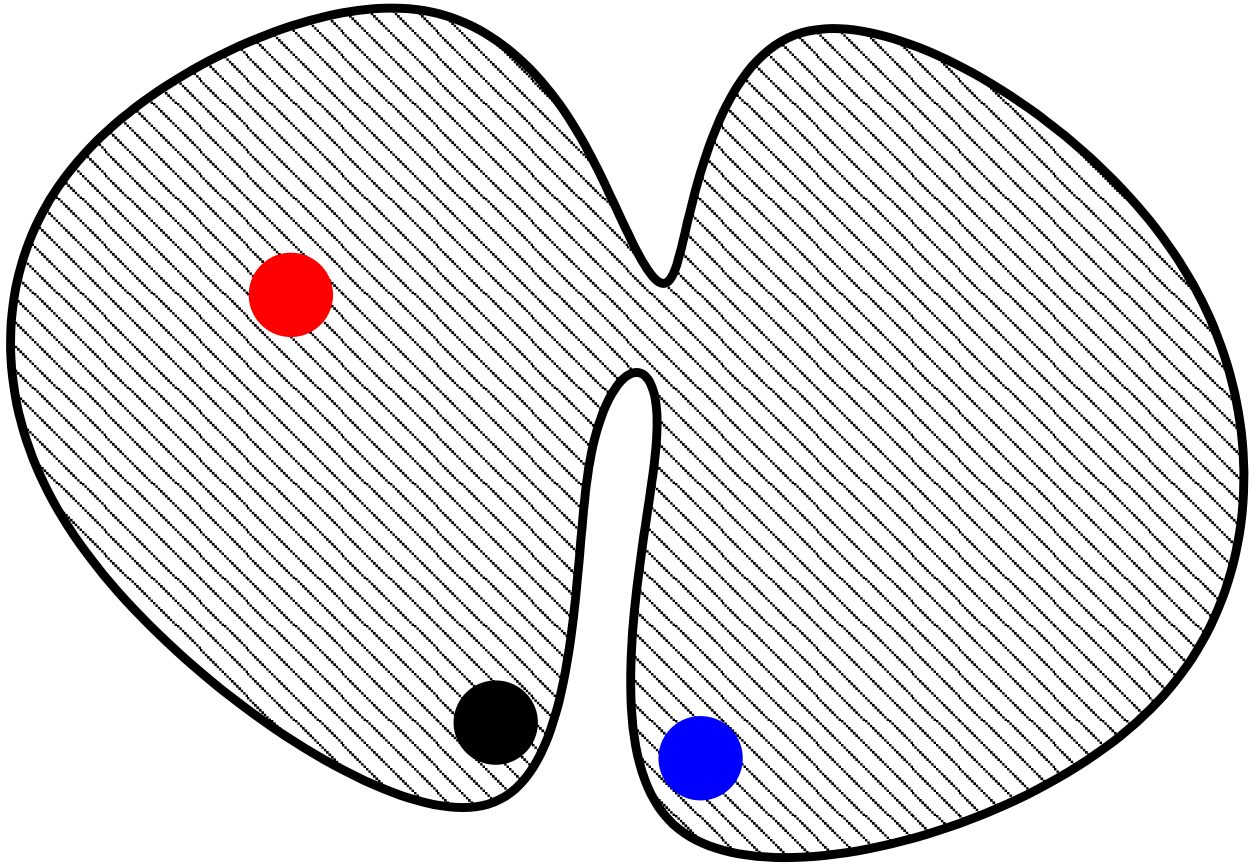


Geometry is important.

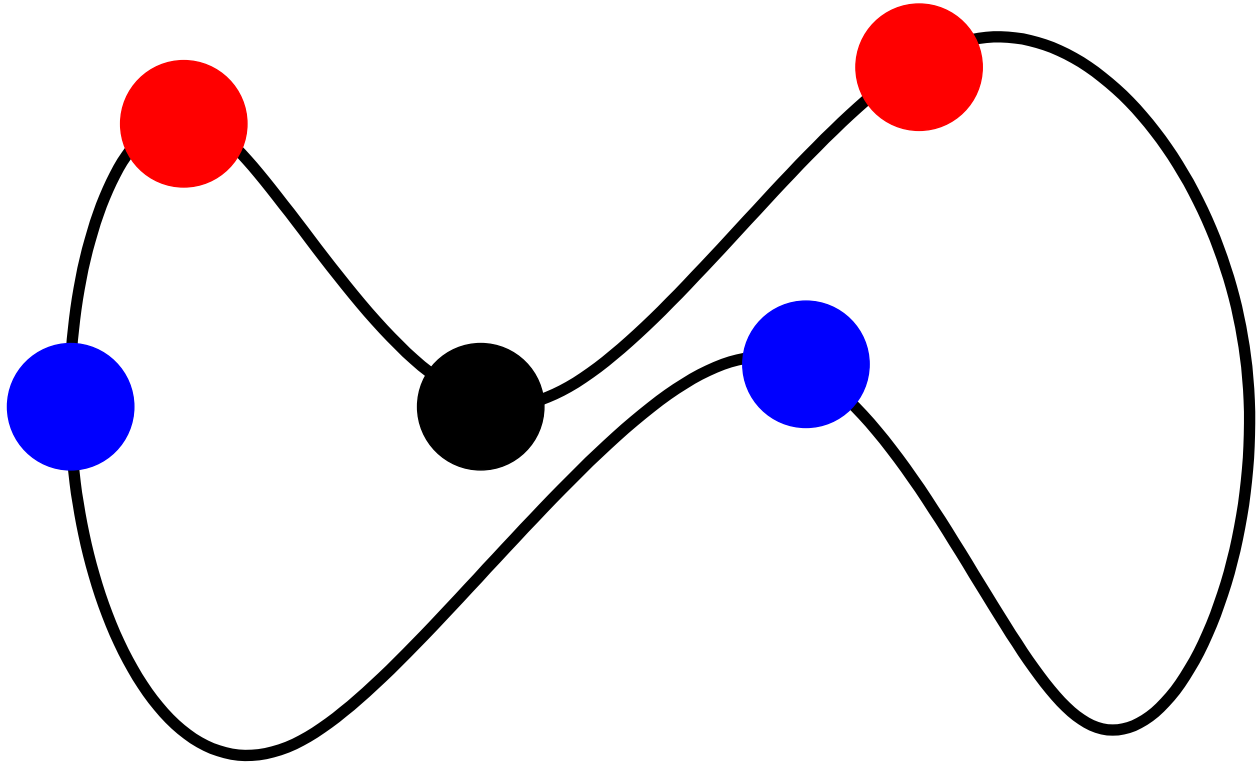
Cluster assumption



Cluster assumption

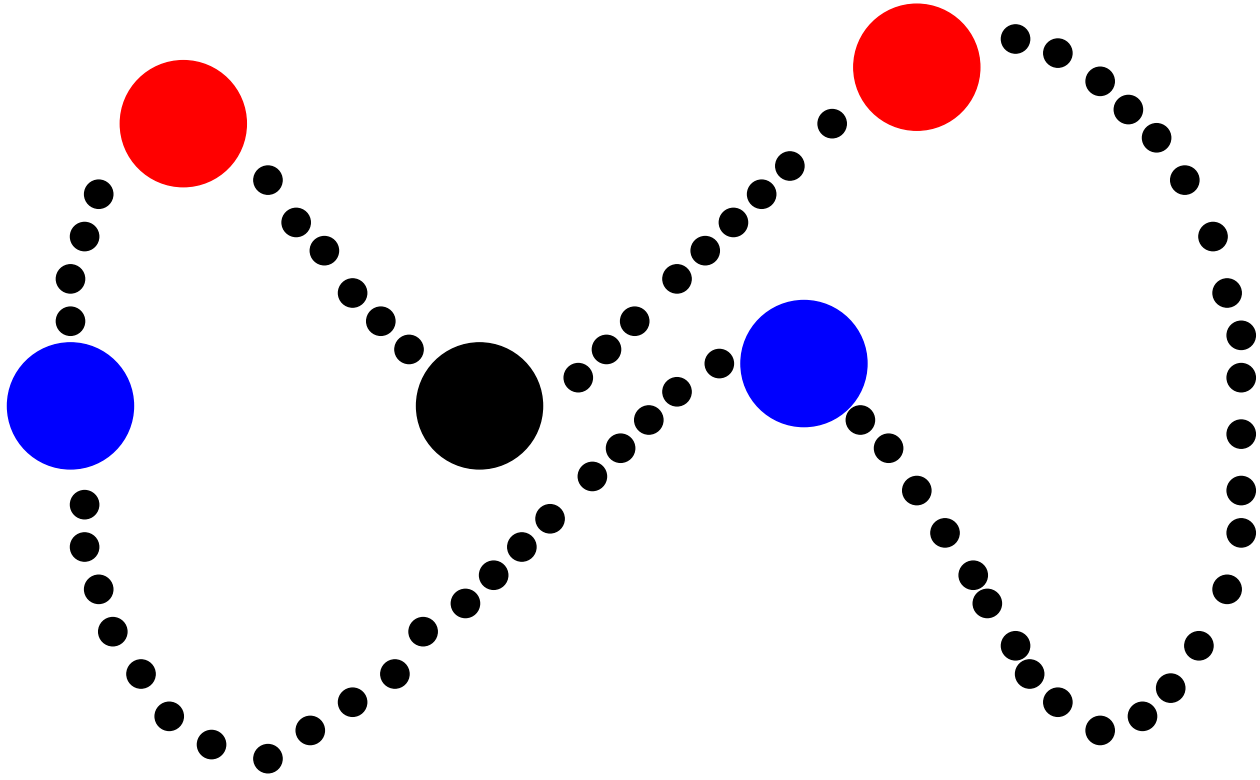


Unlabeled data



Geometry is important.

Unlabeled data



Geometry is important.
Unlabeled data to estimate geometry.

Manifold assumption

Manifold/geometric assumption:
functions of interest are smooth with respect to the underlying geometry.

Manifold assumption

Manifold/geometric assumption:
functions of interest are smooth with respect to the underlying geometry.

Probabilistic setting:
Map $X \to Y$. Probability distribution P on $X \times Y$.

Regression/(two class)classification: $X \to \mathbb{R}$.

Manifold assumption

Manifold/geometric assumption:
functions of interest are smooth with respect to the underlying geometry.

Probabilistic setting:
Map $X \to Y$. Probability distribution P on $X \times Y$.

Regression/(two class)classification: $X \to \mathbb{R}$.

Probabilistic version:
conditional distributions $P(y|x)$ are smooth with respect to the marginal $P(x)$.

What is smooth?

Function $f : X \to \mathbb{R}$. Penalty at $x \in X$:

$$\frac{1}{\delta^k} \int\limits_{\text{small } \delta} \left(f(x) - f(x+\delta)\right)^2 p(x) d\,\delta \approx \|\nabla f\|^2 p(x)$$

Total penalty – Laplace operator:

$$\int_X \|\nabla f\|^2 p(x) = \langle f, \mathcal{L}_p f \rangle_X$$

What is smooth?

Function $f : X \to \mathbb{R}$. Penalty at $x \in X$:

$$\frac{1}{\delta^k} \int\limits_{\text{small } \delta} \left(f(x) - f(x+\delta)\right)^2 p(x) d\,\delta \approx \|\nabla f\|^2 p(x)$$

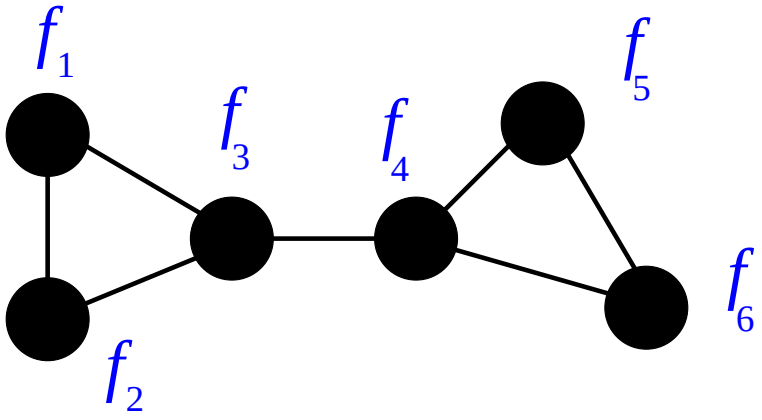
Total penalty – Laplace operator:

$$\int_X \|\nabla f\|^2 p(x) = \langle f, \mathcal{L}_p f \rangle_X$$

Two-class classification – conditional $P(1|x)$.

Manifold assumption: $\langle P(1|x), \mathcal{L}_p P(1|x) \rangle_X$ is small.

Algorithmic framework: Laplacian



$$\mathbf{L} = \begin{pmatrix} 2 & -1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & -1 & 2 \end{pmatrix}$$

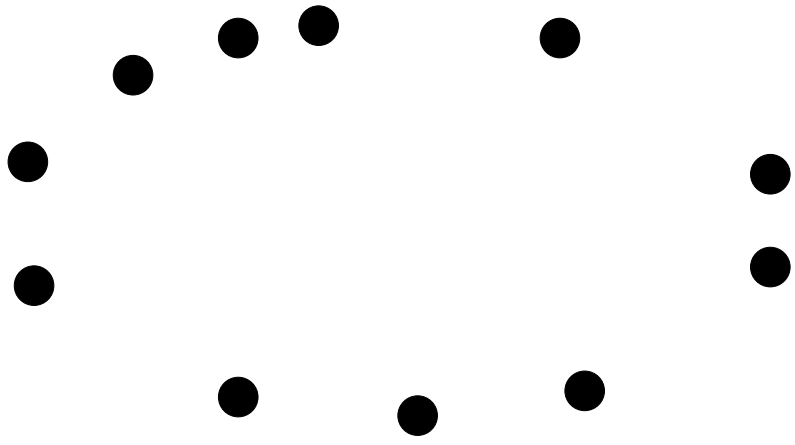
Natural smoothness functional (analogue of grad):

$$\mathcal{S}(\mathbf{f}) = (f_1-f_2)^2+(f_1-f_3)^2+(f_2-f_3)^2+(f_3-f_4)^2+(f_4-f_5)^2+(f_4-f_5)^2+(f_5-f_6)^2$$

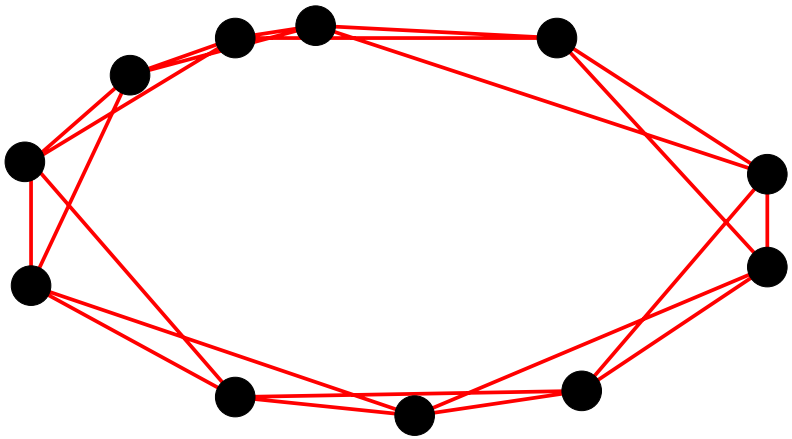
Basic fact:

$$\mathcal{S}(\mathbf{f}) = \sum_{i\sim j}(f_i - f_j)^2 = \frac{1}{2}\,\mathbf{f}^t\,\mathbf{L}\,\mathbf{f}$$

Algorithmic framework



Algorithmic framework

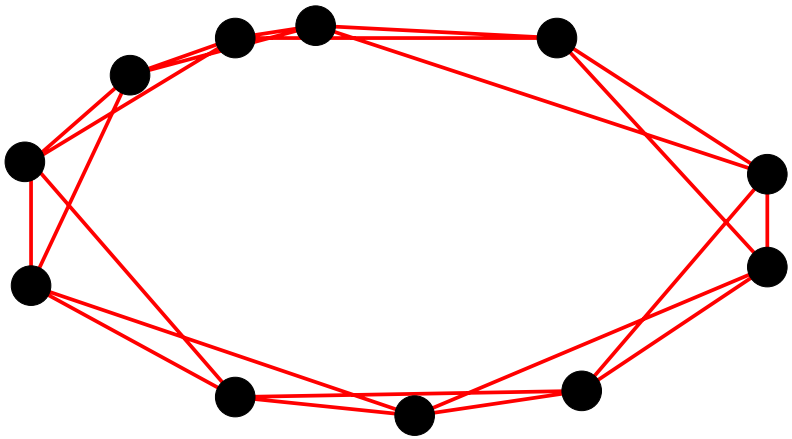


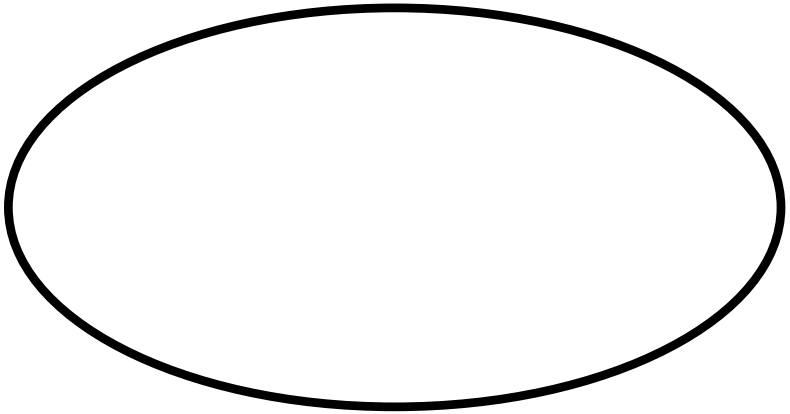
$$W_{ij} = e^{-\frac{\|x_i - x_j\|^2}{t}}$$

$$Lf(x_i) = f(x_i) \sum_j e^{-\frac{\|x_i - x_j\|^2}{t}} - \sum_j f(x_j) e^{-\frac{\|x_i - x_j\|^2}{t}}$$

$$\mathbf{f}^t \mathbf{L} \mathbf{f} = 2 \sum_{i \sim j} e^{-\frac{\|x_i - x_j\|^2}{t}} (f_i - f_j)^2$$

Algorithmic framework





$$W_{ij} = e^{-\frac{\|x_i - x_j\|^2}{t}}$$

$$Lf(x_i) = f(x_i) \sum_j e^{-\frac{\|x_i - x_j\|^2}{t}} - \sum_j f(x_j) e^{-\frac{\|x_i - x_j\|^2}{t}}$$

$$\mathbf{f}^t \mathbf{L} \mathbf{f} = 2 \sum_{i \sim j} e^{-\frac{\|x_i - x_j\|^2}{t}} (f_i - f_j)^2$$

Semi-supervised learning

Learning from labeled and unlabeled data.

- Unlabeled data is everywhere. Need to use it.
- Natural learning is semi-supervised.

Labeled data: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l) \in \mathbb{R}^N \times \mathbb{R}$
Unlabeled data: $\mathbf{x}_{l+1}, \ldots, \mathbf{x}_{l+u} \in \mathbb{R}^N$

Need to reconstruct

$$f_{L,U} : \mathbb{R}^N \to \mathbb{R}$$

Regularization

Estimate $f : \mathbb{R}^N \rightarrow \mathbb{R}$
Data: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)$

Regularized least squares (hinge loss for SVM):

$$f^* = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{l} \sum (f(\mathbf{x}_i) - y_i)^2 + \lambda \| f \|_K^2$$

fit to data + smoothness penalty

$\| f \|_K$ incorporates our smoothness assumptions.
Choice of $\| \ \|_K$ is important.

Algorithm: RLS/SVM

$$\textbf{Solve}: \quad f^* = \operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{l} \sum (f(\mathbf{x}_i) - y_i)^2 + \lambda \|f\|_K^2$$

$\|f\|_K$ is a Reproducing Kernel Hilbert Space norm with kernel $K(\mathbf{x}, \mathbf{y})$.

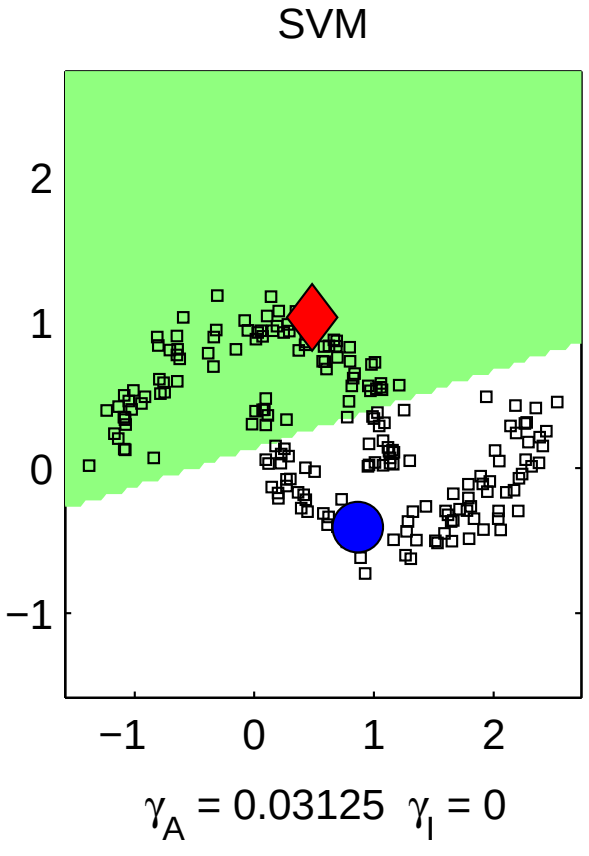
Can solve explicitly (via Representer theorem):

$$f^*(\cdot) = \sum_{i=1}^{l} \alpha_i K(\mathbf{x}_i, \cdot)$$

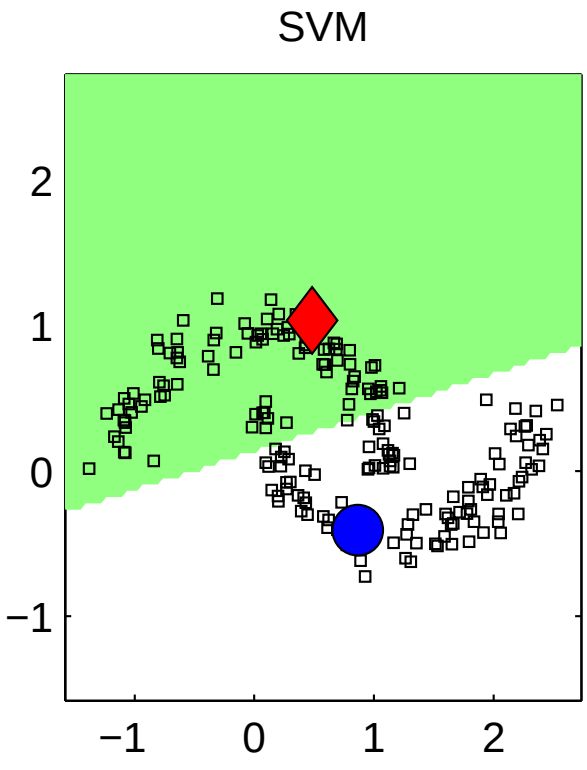
$$[\alpha_1, \ldots, \alpha_l]^t = (\mathbf{K} + \lambda I)^{-1} [y_1, \ldots, y_l]^t$$

$$(\mathbf{K})_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$$

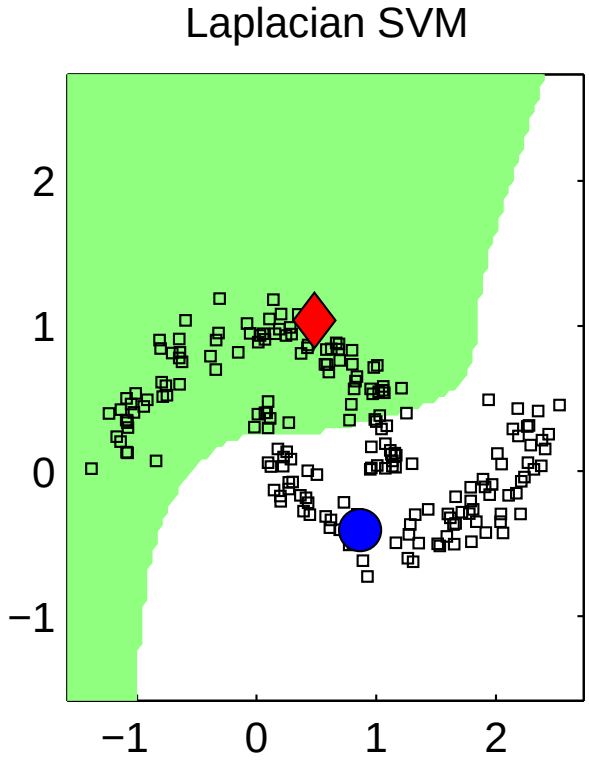
Toy example



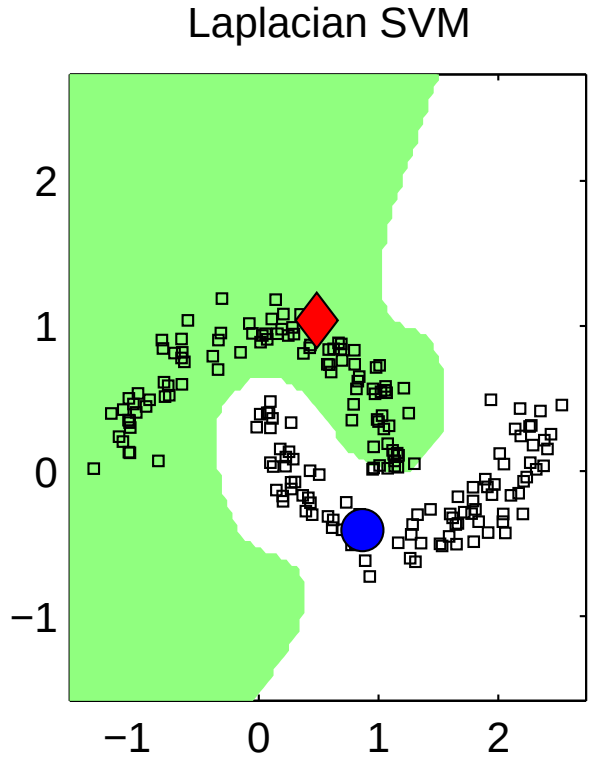
Toy example



$\gamma_A = 0.03125$ $\gamma_I = 0$



$\gamma_A = 0.03125$ $\gamma_I = 0.01$



$\gamma_A = 0.03125$ $\gamma_I = 1$

Data space X.

$$f^* = \operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{l} \sum_{i=1}^{l} (f(\mathbf{x}_i) - y_i)^2 + \lambda_A \|f\|_K^2 + \lambda_I \|f\|_I^2$$

fit to data + extrinsic smoothness + intrinsic smoothness

Data space X.

$$f^* = \operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{l} \sum_{i=1}^{l} (f(\mathbf{x}_i) - y_i)^2 + \lambda_A \|f\|_K^2 + \lambda_I \|f\|_I^2$$

fit to data + extrinsic smoothness + intrinsic smoothness

$\|f\|_I^2 = \langle f, Df \rangle \quad D : \mathbf{RKHS} \to L^2$ is bounded.

Theorem [Intrinsic Representer theorem]

$$f^*(\cdot) = \sum_{i=1}^{l} \alpha_i K(\mathbf{x}_i, \cdot) + \int_X \alpha(\mathbf{x}) K(\mathbf{x}, \cdot) \, d\mu_{\mathbf{x}}$$

Manifold regularization

What is the nature of $\|f\|_I$?

For example:

$$\|f\|_I^2 = \int_X \|\mathrm{grad}_X f\|_X^2 \, d\mu_{\mathbf{x}}$$

Any differential operator on the space X, e.g. $\mathcal{L}^n$.
Diffusions and other kernels on the manifold.

Problem: X is usually not known!

Data-dependent regularization

Estimate $f : \mathbb{R}^N \rightarrow \mathbb{R}$

Labeled data: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)$
Unlabeled data: $\mathbf{x}_{l+1}, \ldots, \mathbf{x}_{l+u}$

$$f^* = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{l} \sum (f(\mathbf{x}_i) - y_i)^2 + \lambda_A \|f\|_K^2 + \lambda_I \|f\|_I^2$$

Empirical estimate:

$$\|f\|_I^2 = \frac{1}{(l+u)^2} [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_{l+u})] \, L \, [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_{l+u})]^t$$

Representer theorem (discrete case):

$$f^*(\cdot) = \sum_{i=1}^{l+u} \alpha_i K(\mathbf{x}_i, \cdot)$$

Explicit solution for quadratic loss:

$$\bar{\alpha} = (J\mathbf{K} + \lambda_A l I + \frac{\lambda_I l}{(u+l)^2}\mathbf{L}\mathbf{K})^{-1}[y_1, \ldots, y_l, 0, \ldots, 0]^t$$

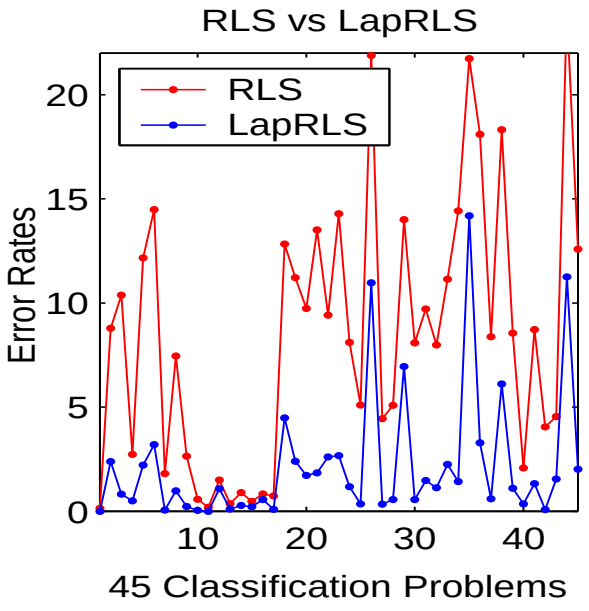
$$(\mathbf{K})_{ij} = K(\mathbf{x}_i, \mathbf{x}_j), \qquad J = diag\,(\underbrace{1, \ldots, 1}_{l}, \underbrace{0, \ldots, 0}_{u})$$

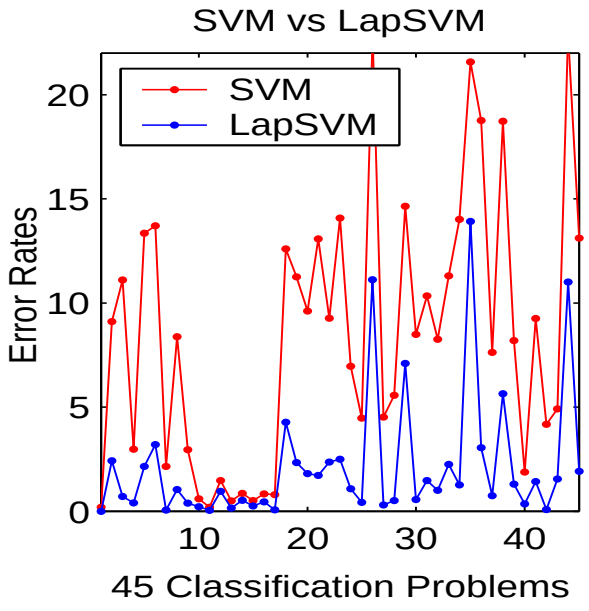
Laplacian RLS demo

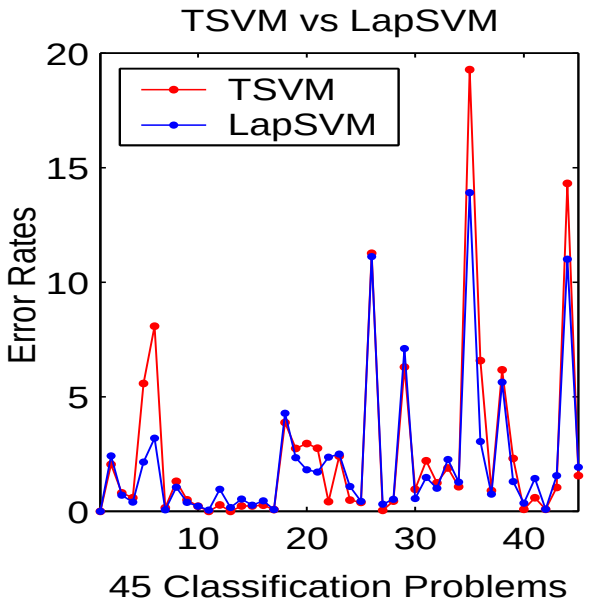
Laplacian Regularized Least Squares demo [link]

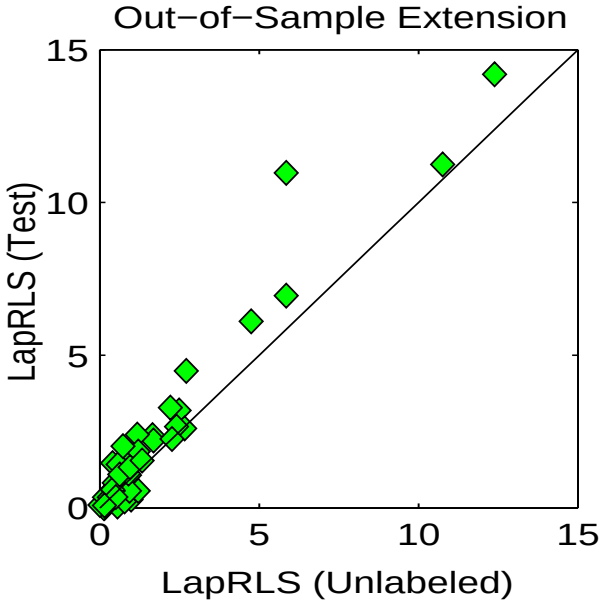
Available at
http://people.cs.uchicago.edu/∼mrainey/jlapvis/JLapVis.html

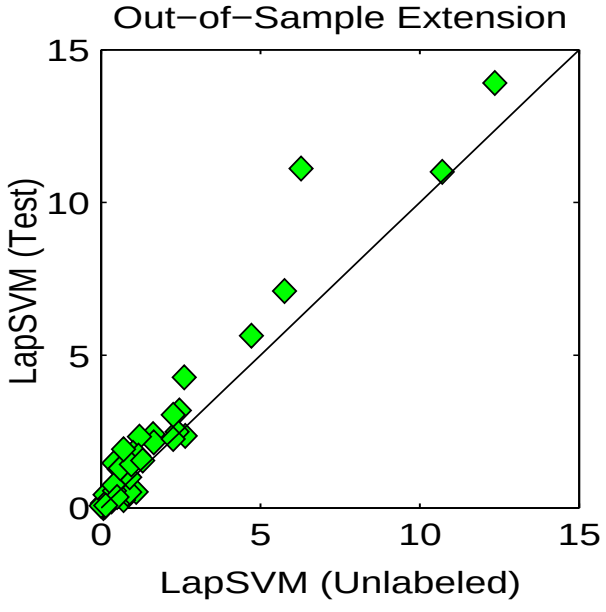
Experimental results: USPS

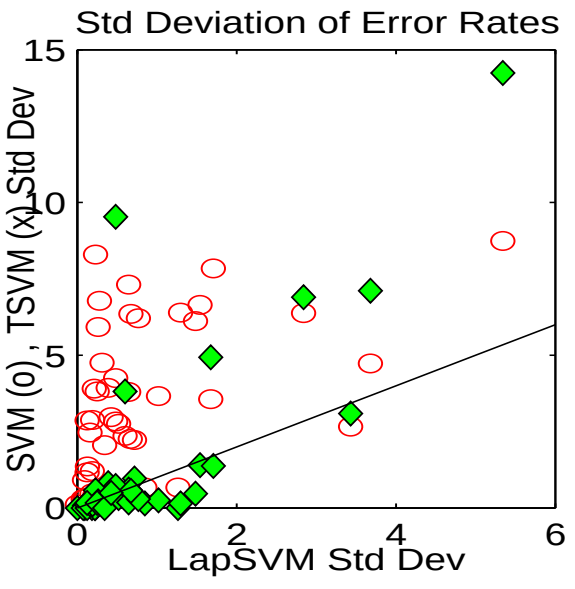










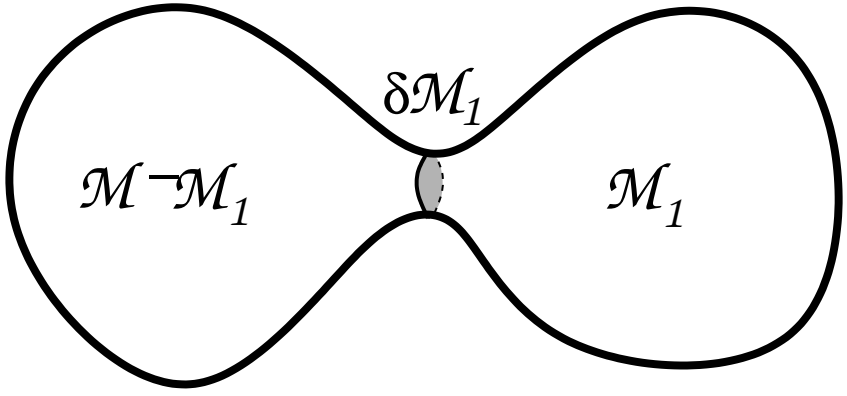


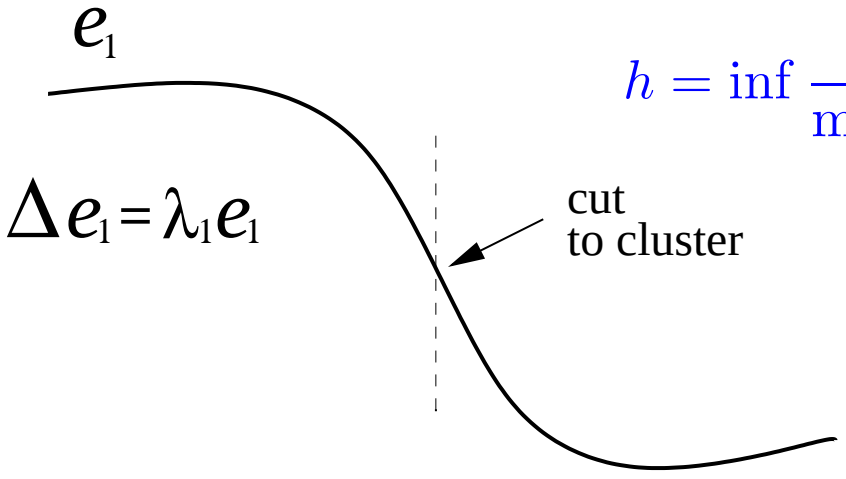
Experimental comparisons

Dataset → Algorithm ↓	g50c	Coil20	Uspst	mac-win	WebKB (link)	WebKB (page)	WebKB (page+link)
SVM (full labels)	3.82	0.0	3.35	2.32	6.3	6.5	1.0
RLS (full labels)	3.82	0.0	2.49	2.21	5.6	6.0	2.2
SVM (l labels)	8.32	24.64	23.18	18.87	25.6	22.2	15.6
RLS (l labels)	8.28	25.39	22.90	18.81	28.0	28.4	21.7
Graph-Reg	17.30	6.20	21.30	11.71	22.0	10.7	6.6
TSVM	6.87	26.26	26.46	7.44	14.5	8.6	7.8
Graph-density	8.32	6.43	16.92	10.48	-	-	-
∇TSVM	5.80	17.56	17.61	5.71	-	-	-
LDS	5.62	4.86	15.79	5.13	-	-	-
LapSVM	**5.44**	**3.66**	**12.67**	10.41	18.1	10.5	**6.4**
LapRLS	**5.18**	**3.36**	**12.69**	10.01	19.2	11.0	6.9
LapSVM$_{joint}$	-	-	-	-	**5.7**	**6.7**	**6.4**
LapRLS$_{joint}$	-	-	-	-	**5.6**	**8.0**	**5.8**

Continuous spectral clustering

Isoperimetric inequalities. Cheeger constant.



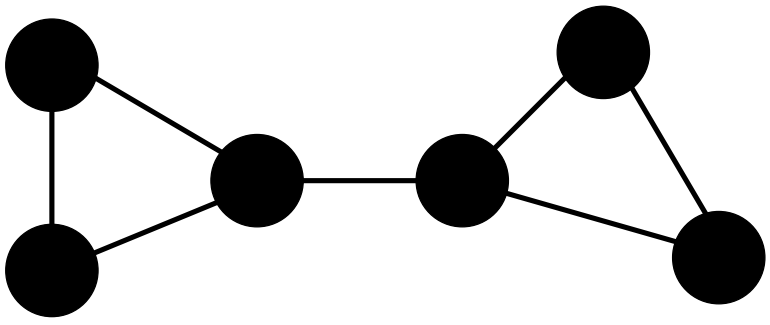


$$h = \inf \frac{\mathrm{vol}^{n-1}(\delta\mathcal{M}_1)}{\min\left(\mathrm{vol}^{n}(\mathcal{M}_1), \mathrm{vol}^{n}(\mathcal{M}-\mathcal{M}_1)\right)}$$

$$h \leq \frac{\sqrt{\lambda_1}}{2}$$

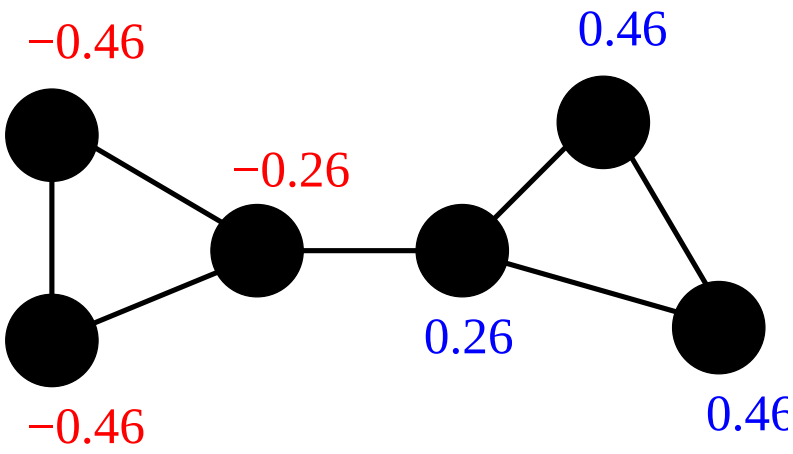
[Cheeger]

Spectral clustering



$$\mathbf{L} = \begin{pmatrix} 2 & -1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & -1 & 2 \end{pmatrix}$$

Spectral clustering

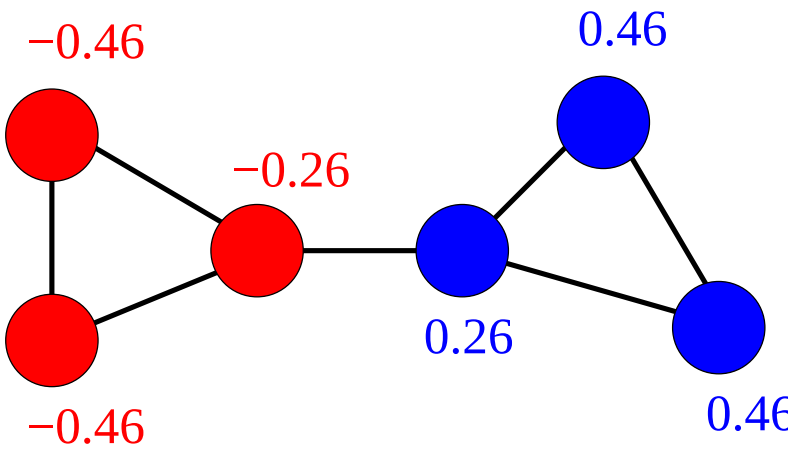


$$\mathbf{L} = \begin{pmatrix} 2 & -1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & -1 & 2 \end{pmatrix}$$

Unnormalized clustering:

$$L\mathbf{e_1} = \lambda_1 \mathbf{e_1} \qquad \mathbf{e_1} = [-0.46, -0.46, -0.26, 0.26, 0.46, 0.46]$$

Spectral clustering

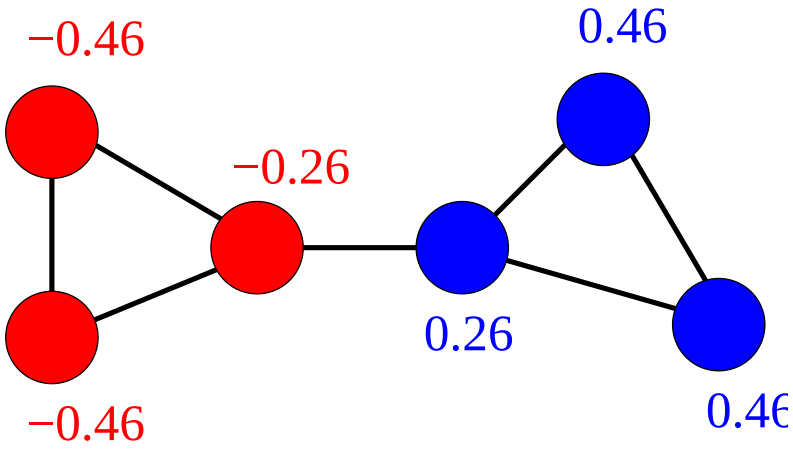


$$\mathbf{L} = \begin{pmatrix} 2 & -1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & -1 & 2 \end{pmatrix}$$

Unnormalized clustering:

$$L\mathbf{e_1} = \lambda_1 \mathbf{e_1} \qquad \mathbf{e_1} = [-0.46, -0.46, -0.26, 0.26, 0.46, 0.46]$$

Spectral clustering



$$\mathbf{L} = \begin{pmatrix} 2 & -1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & -1 & 2 \end{pmatrix}$$

Unnormalized clustering:

$$L\mathbf{e_1} = \lambda_1 \mathbf{e_1} \qquad \mathbf{e_1} = [-0.46, -0.46, -0.26, 0.26, 0.46, 0.46]$$

Normalized clustering:

$$L\mathbf{e_1} = \lambda_1 D\mathbf{e_1} \qquad \mathbf{e_1} = [-0.31, -0.31, -0.18, 0.18, 0.31, 0.31]$$

Regularized spectral clustering

$$f^* = \underset{\substack{\sum_i f(x_i)=0;\ \sum_i f(x_i)^2=1 \\ f \in \mathcal{H}_K}}{\operatorname{argmin}} \lambda \|f\|_K^2 + \sum_{i \sim j} (f(x_i) - f(x_j))^2$$

Representer theorem:

$$f^* = \sum_{i=1}^{u} \alpha_i K(x_i, \cdot)$$

$$P(\lambda K + KLK)P\mathbf{v} = \lambda P K^2 P\mathbf{v}$$

$$(\alpha_1, \ldots, \alpha_u) = P\mathbf{v}$$

Out-of-sample extension for spectral clustering.

Belkin Niyogi Sindhwani 04 Related work: Bengio, et al 04, Vert, Yamanishi 04

Regularized spectral clustering