Total Variation Clustering

Xavier Bresson (CityU)

Joint work with Tony F. Chan (HKUST), Thomas Laurent (UCR), Xue-Cheng Tai (UIB), David Uminsky (USF), Arthur Szlam (CCNY), James von Brecht (UCLA), Ruiliang Zhang (CityU)

Workshop on "Convex Relaxation Methods for Geometric Problems in Scientific Computing"

IPAM, UCLA

February 12, 2013

Machine Learning

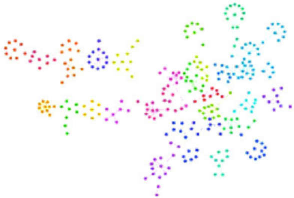
- Definition: Machine Learning is the branch of Artificial Intelligence which is devoted to the design and study of algorithms that learn patterns from data sets in order to make intelligent decisions.

- Applications: Machine vision, Finance, Web Search Engine, social network analysis, etc.

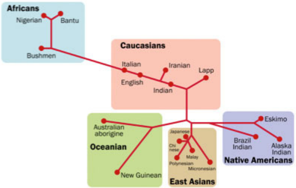
- Emergence of new powerful techniques combining the key mathematical tools of sparsity (compressive sensing), (non-smooth) convex optimization and relaxation methods.

Data Clustering

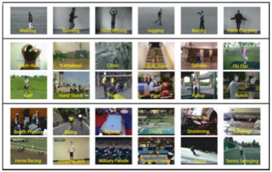
- Objective: Data clustering aims at partitioning data points into sensible groups.
- Some applications:



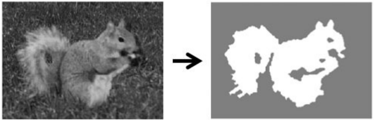
Social network analysis
Community detection



Human genetic clustering



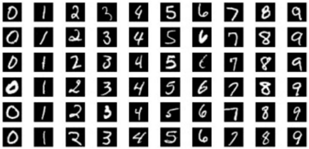
Multimedia organization
Video retrieval

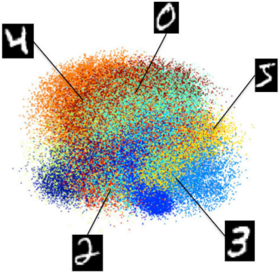


Object identification
Image segmentation

Unsupervised Data Clustering

- Objective: Unsupervised data clustering aims at partitioning data points into sensible groups without any prior information.
- An example: the popular MNIST dataset (Yan LeCun, NYU): 70,000 28×28 images (images are in $\mathbb{R}^{28^2=784}$) of handwritten digits, 0 through 9. There are 6,824 handwritten 0's, 7,233 handwritten 1's, 7,054 handwritten 2's, etc.

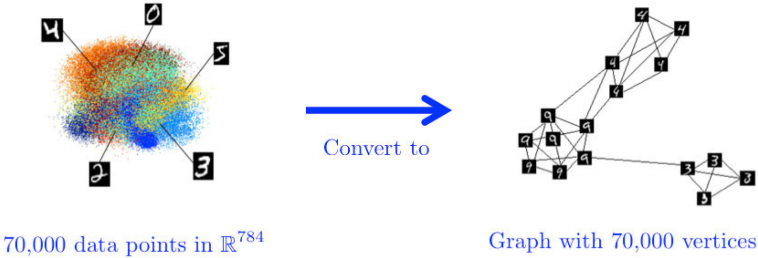




$\Rightarrow$ The goal is to design an efficient algorithm that will break the data set into 10 groups: the 0's, the 1's, etc.

Clustering as Balanced Cut Problem 1/3

- Construct a graph from a set of data points $V = \{x_1, ..., x_N\}$:



- k-NN graph: Connect each data point with its k (=5,10) nearest neighbors.

- Graph weight: $w_{ij} = e^{-\|x_i - x_j\|_2^2/\sigma}$
  ($w_{ij} \approx 1$ if x_i and x_j are similar and $w_{ij} \approx 0$ if x_i and x_j are dissimilar).

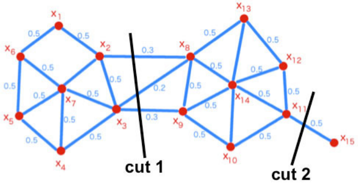
Clustering as Balanced Cut Problem 2/3

- Min cut clustering: cut a graph into two disjoint sets (A, A^c), $A^c = V \setminus A$ while cutting as few links as possible $\Leftrightarrow$ minimize the cut:

$$\min_{A \subsetneq V} \ \mathrm{cut}(A, A^c)$$

where $\mathrm{cut}(A, A^c) = \sum_{i \in A, j \in A^c} w_{ij}$.

- Example:



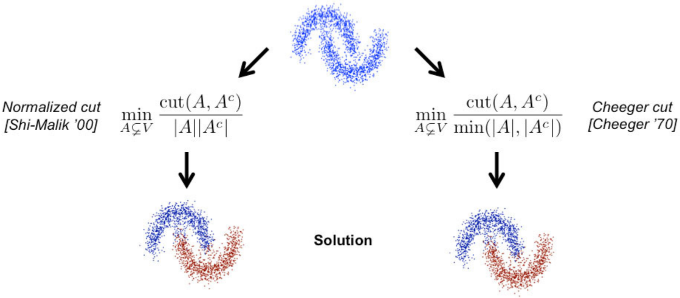
Value of cut1: $\mathrm{cut}(A, A^c) = 0.3 + 0.2 + 0.3 = 0.8$
Value of cut2: $\mathrm{cut}(A, A^c) = 0.5$

- The min cut is biased - it favors cutting small sets of isolated nodes in the graph. Need to separate the data set into two groups of roughly equal size while cutting as few links as possible $\Rightarrow$ balanced cut.

Clustering as Balanced Cut Problem 3/3

- Popular balanced cuts: Cheeger cut [Cheeger '70], Normalized cut [Shi-Malik '00].

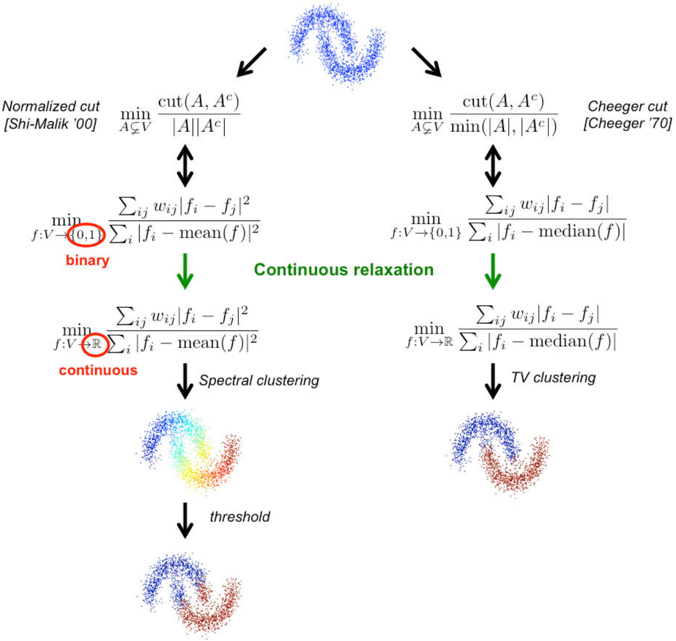
- Example: Two-moon dataset: 2,000 data points in $\mathbb{R}^{100}$.



where $|A|$ is the number of data points in A.

- Balanced cut problems are proved to be NP-hard [Papadimitriou '97].

Continuous Relaxation



Outline

- Spectral clustering
- Total Variation clustering
- Why Total Variation?
- Algorithm
- Experiments
- Current extensions
  - Transductive TV Clustering
  - Semi-Supervised TV Classification

Spectral clustering

Spectral clustering - ℓ^2 relaxation of balanced cut

- ℓ^2 relaxation of the combinatorial problem:

$$\min_{A \subsetneq V} \frac{\mathrm{cut}(A, A^c)}{|A||A^c|} \Leftrightarrow \min_{f:V\to\{0,1\}} \frac{\sum_{ij} w_{ij}|f_i - f_j|^2}{\sum_i |f_i - \mathrm{mean}(f)|^2} \to \min_{f:V\to\mathbb{R}} \frac{\sum_{ij} w_{ij}|f_i - f_j|^2}{\sum_i |f_i - \mathrm{mean}(f)|^2} \tag{1}$$

- (1) is the discrete Rayleigh quotient:

$$\min_{f:\Omega\to\mathbb{R}\perp\mathbf{1}} \frac{\langle \Delta f, f\rangle}{\|f\|_{L^2}^2} = \min_{f\perp\mathbf{1}} \frac{\|\nabla f\|_{L^2}^2}{\|f\|_{L^2}^2}$$

where $f \perp \mathbf{1} \Leftrightarrow \int f(x)\,\mathbf{1}\,dx = 0 \Leftrightarrow f$ has mean zero
($f = \mathbf{1}$ is the first eigenvector because $\Delta\mathbf{1} = 0 \times \mathbf{1}$)

$$\min_{f\perp\mathbf{1}} \frac{\|\nabla f\|_{L^2}^2}{\|f\|_{L^2}^2} = \min_{f:\Omega\to\mathbb{R}} \frac{\|\nabla(f - \mathrm{mean}(f))\|_{L^2}^2}{\|f - \mathrm{mean}(f)\|_{L^2}^2} = \min_{f:\Omega\to\mathbb{R}} \frac{\|\nabla f\|_{L^2}^2}{\|f - \mathrm{mean}(f)\|_{L^2}^2}$$

$$= \min_f \frac{\text{measure of flatness}}{\text{variance}}$$

Laplacian eigenvectors

- Solution of optimization problem (1) is well-known: it is the second eigenvector of the graph Laplacian.
- Standard spectral clustering algorithm:

  Given a set of data points,

  1- Compute the graph similarity matrix w_{ij},

  2- Compute the normalized graph Laplacian [Shi-Malik '00, Coifman-et.al. '05, Belkin-Niyogi '05] s.a.

  $$L = I - D^{-1/2} W D^{-1/2}$$

  where D is a diagonal matrix with entries $d_i = \sum_j w_{ij}$,

  3- Compute the second eigenvector of L:

  $$L\Psi_2 = \lambda_2 \Psi_2$$

  4- Threshold Ψ_2 to define two clusters.

- Advantages: unique minimizer, fast to compute with any good linear algebra software.

Limitation of Spectral Clustering

- Spectral clustering provides satisfying solutions as long as the data geometry/structure is not too "complex".

- This limitation comes from the relaxation that is loose:

$$\frac{1}{2\max_i d_i} h_\star^2 \leq \lambda_2 \leq 2h_\star$$

where

$$\lambda_2 = \min_{f:V\to\mathbb{R}} \frac{\sum_{ij} w_{ij}|f_i - f_j|^2}{\sum_i |f_i - \mathrm{mean}(f)|^2} \quad \text{(spectral solution)}$$

and

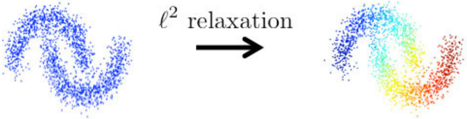
$$h_\star = \min_{A\subsetneq V} \frac{\mathrm{cut}(A, A^c)}{|A||A^c|} \quad \text{(combinatorial solution)}$$

Total Variation clustering

Total Variation Clustering [Szlam-B '09]

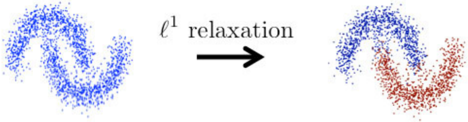
- ℓ^2 relaxation of the combinatorial problem:

$$\min_{A \subsetneq V} \frac{\mathrm{cut}(A, A^c)}{|A||A^c|} \Leftrightarrow \min_{f: V \to \{0,1\}} \frac{\sum_{ij} w_{ij} |f_i - f_j|^2}{\sum_i |f_i - \mathrm{mean}(f)|^2} \to \min_{f: V \to \mathbb{R}} \frac{\sum_{ij} w_{ij} |f_i - f_j|^2}{\sum_i |f_i - \mathrm{mean}(f)|^2}$$



- ℓ^1 relaxation of the combinatorial problem:

$$\min_{A \subsetneq V} \frac{\mathrm{cut}(A, A^c)}{\min(|A|, |A^c|)} \Leftrightarrow \min_{f: V \to \{0,1\}} \frac{\sum_{ij} w_{ij} |f_i - f_j|}{\sum_i |f_i - \mathrm{median}(f)|} \to \min_{f: V \to \mathbb{R}} \frac{\sum_{ij} w_{ij} |f_i - f_j|}{\sum_i |f_i - \mathrm{median}(f)|}$$



Why Total Variation?

Imaging Science

- Image denoising = remove noise in images

f_{original}

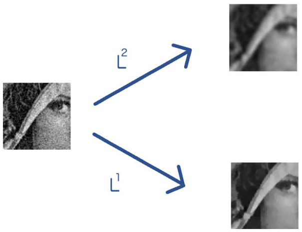
f_0

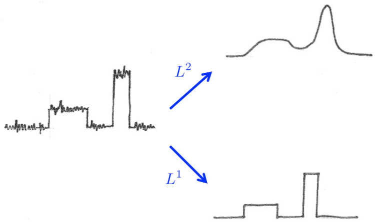
- Denoising cast as optimization problem:

$$f_{\text{denoized}} = \arg\min_f \left\{ \|\nabla f\|_{L^2}^2 + \lambda \|f - f_0\|_{L^2}^2 \right\} \qquad \text{(Heat equation)}$$

$$f_{\text{denoized}} = \arg\min_f \left\{ \|\nabla f\|_{L^1} + \lambda \|f - f_0\|_{L^2}^2 \right\} \qquad \text{(ROF model)}$$

L^2/Dirichlet v.s. L^1/TV





- Total variation preserves sharp edges (Dirichlet does not).

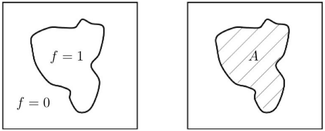
TV makes the connection between geometry and function

- Coarea formula [Federer '69]:

$$TV(f) = \int_{\mathbb{R}^d} |\nabla f| = \int_{\mathbb{R}} \mathrm{Per}(A_\mu) d\mu \quad \text{where} \quad A_\mu = \{x : f(x) \geq \mu\}$$

- When $f = 1_A$ is an indicator function of a set A:

$$TV(f = 1_A) = \mathrm{Per}(A)$$



f A

- Equivalence between geometrical problem and functional problem is essential as the balanced cut is basically the discrete version of a geometric problem:

Balanced cut

	Continuous		Discrete
Geometry	$\min_{A \subset \mathbb{R}^d} \frac{\mathrm{Per}(A)}{\min(\mathrm{Area}(A), \mathrm{Area}(A^c))}$	$\rightarrow$	$\min_{A \subsetneq V} \frac{\mathrm{cut}(A, A^c)}{\min(\lvert A\rvert, \lvert A^c\rvert)}$
	$\Updownarrow$		$\Updownarrow$
Function	$\min_{f:\mathbb{R}^d \to \{0,1\}} \frac{\int_{\mathbb{R}^d} \lvert\nabla f\rvert}{\int_{\mathbb{R}^d} \lvert f(x) - \mathrm{median}(f)\rvert dx}$	$\rightarrow$	$\min_{f:V \to \{0,1\}} \frac{\sum_{ij} w_{ij} \lvert f_i - f_j\rvert}{\sum_i \lvert f_i - \mathrm{median}(f)\rvert}$

TV Clustering

TV Clustering is an Exact Relaxation of Balanced Cut !

- Equivalence between the combinatorial problem and the continuous problem:

$$\min_{A \subsetneq V} \frac{\mathrm{cut}(A, A^c)}{\min(|A|, |A^c|)}$$

NP-hard Combinatorial Problem
Cheeger Balanced Cut

Tight Relaxation $\longrightarrow$

$\longleftrightarrow$ **Exact Relaxation**

$$\min_{f:V \to [0,1]} \frac{\sum_{ij} w_{ij} |f_i - f_j|}{\sum_i |f_i - \mathrm{median}(f)|}$$

Continuous Optimization Problem
TV Clustering

- Note that the continuous optimization problem is non-convex $\Rightarrow$ no algorithm can guarantee to find a global minimizer.

TV Clustering is a Tight Relaxation of Balanced Cut [Szlam-B '09, Chung '97]

- **Theorem.** Let A^* be a minimizer of $\min_{A \subsetneq V} \left\{ CC(A) = \frac{\mathrm{cut}(A,A^c)}{\min(|A|,|A^c|)} \right\}$, then any indicator (i.e. binary) function of median zero

$$f^*(x_i) = \begin{cases} a & \text{if } x_i \in A^* \\ b & \text{if } x_i \in (A^*)^c \end{cases}$$

is a minimizer of $\min_{f:V\to\mathbb{R}} \left\{ E(f) = \frac{\sum_{ij} w_{ij}|f_i - f_j|}{\sum_i |f_i - \mathrm{median}(f)|} = \frac{\|f\|_{TV}}{\|f - \mathrm{median}(f)\|_1} \right\}$.

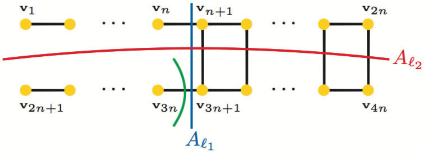
Proof.
1) Extreme points of the TV-unit ball

$$B_{TV} = \{f \in \mathbb{R}^n : \|f\|_{TV} \leq 1, \mathrm{median}(f) = 0\}$$

are binary functions [Strang '83].
2) Rescale: Fix $\|f\|_{TV} = 1$. This leads to $\max_{f \in B_{TV}} \|f\|_1$.
3) Function $\|f\|_1$ is convex and takes its maximum at extreme points of B_{TV} (which are binary functions).
4) Observe that $E(f = 1_A) = 2.CC(A)$. So if A^* is a solution of $\min_A CC(A)$ then f^* is a solution of the continuous relaxation $\min_f E(f)$.

TV Clustering is an **Exact** Relaxation of Balanced Cut [B-Laurent-Uminski-von Brecht '12]

- The cockroach graph [Guattery-Miller '98].
  The ℓ^2-relaxation of Normalized Cut, i.e. spectral clustering, gives the partition in red. The resulting Normalized Cut energy exhibits arbitrarily large deviations from the optimal solution of the Normalized Cut in green.
  The optimal solutions of the ℓ^1-relaxation and the Cheeger cut, shown in blue, coincide.



$$\min_{A \subsetneq V} \frac{\mathrm{cut}(A, A^c)}{|A||A^c|} = \frac{1}{3k^2} << \frac{1}{4k} = \frac{\mathrm{cut}(A_{\ell_2}, A_{\ell_2}^c)}{|A_{\ell_2}||A_{\ell_2}^c|}$$

$$\min_{A \subsetneq V} \frac{\mathrm{cut}(A, A^c)}{\min\{|A|, |A^c|\}} = \frac{\mathrm{cut}(A_{\ell_1}, A_{\ell_1}^c)}{\min\{|A_{\ell_1}|, |A_{\ell_1}^c|\}}$$

Characterization of Local Minimizers

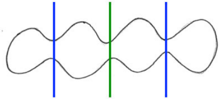
- Theorem (Explicit Correspondence of Local Minima)

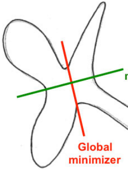
  1. *Suppose $S_1 \subsetneq S_2 \subsetneq \cdots \subsetneq S_k$ is a k-local minimum of the combinatorial problem and let $f \in$ strict convex hull $\{f_{S_1}, \cdots, f_{S_k}\}$. Then any function of the form $g = \alpha f + \beta \mathbf{1}$ defines a $(k+1)$-valued local minimum of the continuous problem and with $E(g) = \mathcal{C}(S_1)$.*
  2. *Suppose that f is a $(k+1)$-valued local minimum and let $c_1 > c_2 > \cdots > c_{k+1}$ denote its range. For $1 \leq i \leq k$ set $\Omega_i = \{f = c_i\}$. Then the increasing collection of sets $S_1 \subsetneq \cdots \subsetneq S_k$ given by*

$$S_1 = \Omega_1, \quad S_2 = \Omega_1 \cup \Omega_2 \qquad \cdots \qquad S_k = \Omega_1 \cup \cdots \cup \Omega_k$$

  is a k-local minimum of the combinatorial problem with $\mathcal{C}(S_i) = E(f)$.

- Interpretation: A cut is a local minima of the non-convex energy E iff it has a smaller energy than all the non-intersecting cuts.





Algorithm

Algorithmic challenges

- Optimization problem:

$$\min_{f:V\to\mathbb{R}} \frac{T(f)}{B(f-\mathrm{median}(f))}$$

  where

$$T(f) = \sum_{ij} w_{ij}|f_i - f_j| = \|f\|_{\mathrm{TV}}$$

$$B(f) = \sum_i |f_i| = \|f\|_{\ell^1}$$

- This problem is non-differentiable and non-convex $\Rightarrow$ existence of local minimizers.

- But the main challenge is actually to design a fast algorithm that is guaranteed to converge to a local minimizer:

  • In [Szlam-B '09], the algorithm was heuristic (no proof of convergence).

  • In [B-Laurent-Uminski-von Brecht '12], we introduced an explicit-implicit gradient flow algorithm and prove the convergence.

Explicit-implicit gradient flow [B-Laurent-Uminski-von Brecht '12]

- It is equivalent to minimize

$$\left\{ \frac{T(f)}{B(f-\mathrm{median}(f))} \right\} \text{ or } \left\{ E(f) = \frac{T(f)}{B(f)} \text{ s.t. } \mathrm{median}(f) = 0 \right\}.$$

- The explicit-implicit gradient flow for $E(f)$ is

$$\frac{f^{k+1}-f^k}{\tau^k} = -\frac{\partial T(f^{k+1}) - E(f^k)\partial B(f^k)}{B(f^k)}$$

where τ^k is the time step, ∂T and ∂B are subgradients of T and B.

This leads to

$$g^k = f^k + \frac{\tau^k}{B(f^k)} E(f^k)\partial B(f^k)$$

$$f^{k+1} = \arg\min_f \left\{ T(f) + B(f^k)\frac{\|f-g^k\|_2^2}{2\tau^k} \right\}.$$

- Notes:

1) To remove the scaling effect we project each iterate onto the sphere $\mathcal{S}^{n-1} = \{u \in \mathbb{R}^n : \|u\|_2 = 1\}$ at the end of each iteration.

2) Numerical experiments suggest fast convergence speed for time step:

$$\tau^k = c\frac{B(f^k)}{E(f^k)}, \quad c > 0.$$

Algorithm

- **Algorithm**. Starting from $f^0 \in \mathcal{S}^{n-1}$ with $\mathrm{median}(f^0) = 0$, define the sequence of iterates:

$$
\begin{aligned}
g^k &= f^k + c\partial B(f^k) \\
\hat{h}^k &= \arg\min_f \left\{ T(f) + E(f^k) \frac{\|f - g^k\|_2^2}{2c} \right\} \qquad (2) \\
h^k &= \hat{h}^k - \mathrm{median}(\hat{h}^k)\mathbf{1} \\
f^{k+1} &= \frac{h^k}{\|h^k\|_2}
\end{aligned}
$$

Notes:

1) The non-smooth optimization problem (2) is the standard ROF problem [Rudin-Osher-Fatemi '92] that can be solved efficiently using approaches borrowed from Compressive Sensing such as Alternating Direction Method of Multipliers (ADMM) [Goldstein-Osher '09], Iterative Shrinkage-Thresholding [Beck-Teboulle '09], Primal-Dual methods [Chambolle-Pock '11], etc

2) We use $c = 1$ in all experiments.

Properties of the algorithm

- Monotonicity:
  $E(f^k) \geq E(f^{k+1}) + \frac{E(f^k)}{B(h^k)} \|h^k - f^k\|_2^2$
- Compactness:
  All the iterates $f^k, g^k, \hat{h}^k, h^k$ belong to a (compact) annulus.
- "Continuity"
  This allows to prove that all the accumulation points of the sequence $\{f^k\}$ are critical points of the Energy.
- $\|f^{k+1} - f^k\|_{L^2} \to 0$
  Either the sequence converges,
  or the set of accumulation points is a connected subset of $\mathcal{S}_0^{n-1}$
- Full convergence near local min:
  If f^* is an isolated local minima of the energy,
  then $f^k \to f^*$ if the initial iterate is close enough to f^*.

Speed up [B-Laurent-Uminski-von Brecht '13]

- Approximate ROF [B-Laurent-Uminski-von Brecht '12]: the stopping criterion for the inner ROF problem is chosen to be

$$\|h_{i+1}^k - h_i^k\|_2 < \varepsilon,$$

but what values for ε? For large ε, no guarantee of monotonicity and for small ε, slow algorithm.

- Adaptable stopping condition:

$$E(f^k) \geq E(h_i^k) + \theta \frac{E(f^k)}{B(h_i^k)} \|h_i^k - f^k\|_2^2, \quad \theta \in (0,1)$$

This stopping condition guarantees monotonicity and convergence!

- We use $\theta = 0.99$ in all experiments.
- Experiments showed a speed improvement of **2x** compared to the non-adaptable algorithm.

Experiments

Experiments

- We use the benchmark MNIST (digit numbers), USPS (digit numbers) and COIL (rotating objects), CURET (textures) datasets.

  We preprocessed the MNIST, USPS and COIL data by projecting onto the first 50 principal components, and take $k = 10$ nearest neighbors for the MNIST, USPS and CURET datasets and $k = 5$ nearest neighbors for the COIL dataset.

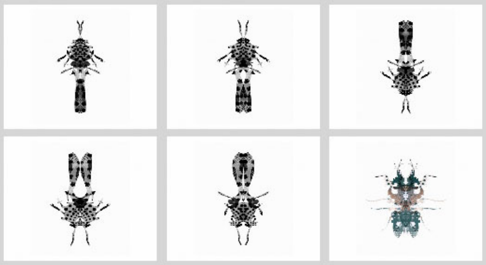
  The table summarizes the results of these tests.

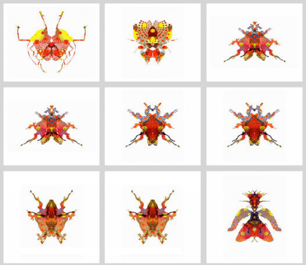
	Spectral clustering		TV clustering	
	Err. (%)	Time (min.)	Err. (%)	Time (min.)
MNIST (10 classes)	24.32	**2.85**	**11.76**	21.85
USPS (10 classes)	26.33	**0.46**	**4.11**	3.08
COIL (20 classes)	40.34	**0.15**	**1.58**	2.12
CURET (7 classes)	26.3	**0.09**	**17.1**	0.97

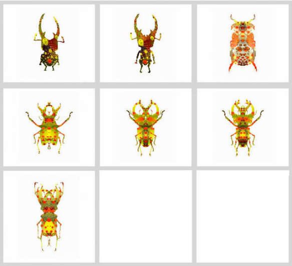
For fun: organizing artist photos 1/2

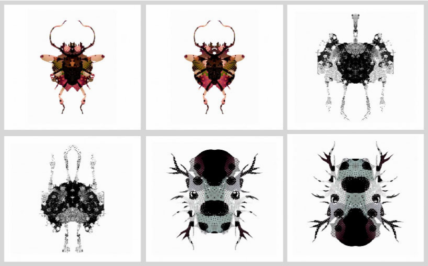
- Diego Porcel (http://www.diegoporcel.com) is a photograph who creates fractal-insects.

For fun: organizing artist photos 2/2

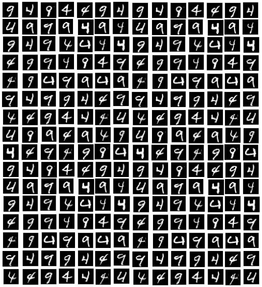


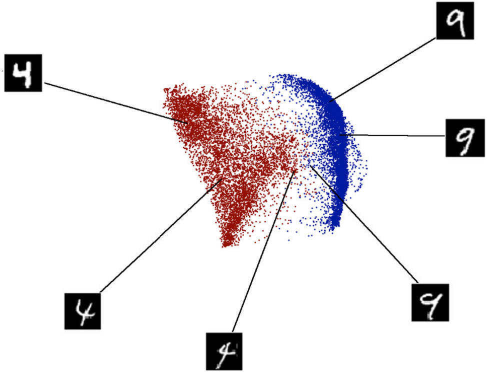






Matlab Demo





Take Home Message

Normalized cut [Shi-Malik '00]

$$\min_{A \subsetneq V} \frac{\mathrm{cut}(A, A^c)}{|A||A^c|}$$

Cheeger cut [Cheeger '70]

$$\min_{A \subsetneq V} \frac{\mathrm{cut}(A, A^c)}{\min(|A|, |A^c|)}$$

$$\min_{f:V \to \{0,1\}} \frac{\sum_{ij} w_{ij} |f_i - f_j|^2}{\sum_i |f_i - \mathrm{mean}(f)|^2}$$

binary

$$\min_{f:V \to \{0,1\}} \frac{\sum_{ij} w_{ij} |f_i - f_j|}{\sum_i |f_i - \mathrm{median}(f)|}$$

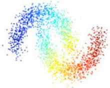
Continuous relaxation

$$\min_{f:V \to \mathbb{R}} \frac{\sum_{ij} w_{ij} |f_i - f_j|^2}{\sum_i |f_i - \mathrm{mean}(f)|^2}$$

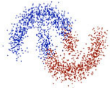
continuous

$$\min_{f:V \to \mathbb{R}} \frac{\sum_{ij} w_{ij} |f_i - f_j|}{\sum_i |f_i - \mathrm{median}(f)|}$$

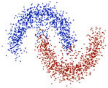
Spectral clustering



threshold



TV clustering



Future work

- New applications: spectral clustering is a building block for data analysis products.
  $\Rightarrow$ TV clustering is promising to improve solutions of existing problems.

  Example: Unsupervised document retrieval (Google Search Engine)
  The objective is to find the most relevant documents related to a query.
  Joint work with T. Laurent (UCR) and J. Wu (CityU).

	TREC-6		TREC-7		TREC-8	
	P@20	MAP	P@20	MAP	P@20	MAP
Baseline	44.60	39.62	45.80	31.22	49.40	33.63
Language model '11	45.40	39.87	**46.20**	31.41	49.60	33.95
TV clustering '12	**45.80**	**40.22**	**46.20**	**31.59**	**49.80**	**34.16**

Table: Retrieval success rate (%)

Current extensions

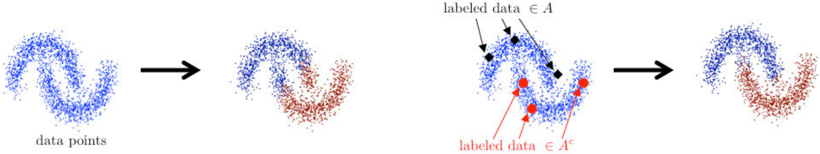
From Unsupervised Clustering to Transductive Clustering

Unsupervised and Transductive Clustering

- Unsupervised clustering: partitioning data points into sensible groups with no prior knowledge about data points.

- Transductive clustering [Vapnik-et.al. '98]: partitioning data points into sensible groups given a small set of labeled data (labeled data are data assigned to a specific group).

  Property: a few labeled data can significantly improve unsupervised clustering results.



Unsupervised clustering (no prior information)

Transductive clustering (with labeled data)

Mathematical formulation of Transductive Clustering

- Unsupervised clustering: given N unlabeled data points $\{x_i\}_{1\leq i\leq N}$, solve the optimization problem:

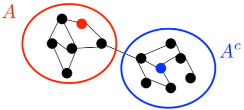
$$\min_{A \subsetneq V} \frac{\mathrm{cut}(A,A^c)}{\min(|A|,|A^c|)}$$

- Transductive clustering: given N data points

$$\{x_i\}_{1\leq i\leq N} = \{ \underbrace{x_1, \dots, x_{N-n}}_{N-n \text{ unlabeled data}}, \underbrace{x_1^\ell, \dots, x_n^\ell}_{n \text{ labeled data}} \},$$

where labeled data $\{x_i^\ell\}_{1\leq i\leq n}$ are assigned either to A or A^c, solve the optimization problem:

$$\min_{A \subsetneq V} \frac{\mathrm{cut}(A,A^c)}{\min(|A|,|A^c|)} \quad \text{given the labeled data } \{x_i^\ell\}_{1\leq i\leq n} \tag{3}$$

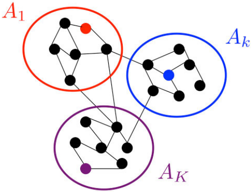


The combinatorial problem (3) is still a NP-hard problem $\Rightarrow$ relaxation needed.

Multiclass transductive clustering

- Formulation:

$$\min_{\{A_k\}} \sum_{k=1}^{K} \frac{\mathrm{cut}(A_k, A_k^c)}{\min(|A_k|, |A_k^c|)} \quad s.t. \quad \begin{cases} \cup_{k=1}^{K} A_k = V \\ A_i \cap A_j = \emptyset \quad \forall i \neq j \\ \text{and labels } \{x_i^\ell\}_{1 \leq i \leq n} \end{cases}$$



Multiclass Transductive Clustering

Total Variation Transductive Clustering [B-Tai-Chan-Szlam '12]

- Continuous relaxation: Multiclass TV transductive clustering

$$\min_{\{A_k\}} \sum_{k=1}^{K} \frac{\mathrm{cut}(A_k, A_k^c)}{\min(|A_k|, |A_k^c|)} \quad s.t. \quad \begin{cases} \cup_{k=1}^{K} A_k = V \\ A_i \cap A_j = \emptyset \quad \forall i \neq j \\ \text{and labels } \{x_i^\ell\}_{1 \leq i \leq n} \end{cases}$$

$\Updownarrow$ equivalence

$$\min_{f_k : V \to \{0,1\}} \frac{\sum_{ij} w_{ij} |f_k(i) - f_k(j)|}{\sum_i |f_k(i) - \mathrm{median}(f_k)|} \quad s.t. \quad \begin{cases} \sum_{k=1}^{K} f_k(i) = 1 \quad \forall i \in V \text{ (simplex constraint)} \\ f_k(i) = \begin{cases} 1 & \forall x_i^\ell \in A_k \\ 0 & \forall x_i^\ell \notin A_k \end{cases} \end{cases}$$

$\downarrow$ ℓ^1 relaxation

$$\min_{f_k : V \to [0,1]} \frac{\sum_{ij} w_{ij} |f_k(i) - f_k(j)|}{\sum_i |f_k(i) - \mathrm{median}(f_k)|} \quad s.t. \quad \begin{cases} \sum_{k=1}^{K} f_k(i) = 1 \quad \forall i \in V \\ f_k(i) = \begin{cases} 1 & \forall x_i^\ell \in A_k \\ 0 & \forall x_i^\ell \notin A_k \end{cases} \end{cases}$$

- Exact relaxation? Probably no.

Optimization

- Algorithm

  For n=1,2,... until convergence:

  For $k = 1$ to K

$$
\begin{aligned}
g_k^n &= f_k^n + \partial B(f_k^n) \\
\hat{h}_k^n &= \arg\min_f \left\{ T(f_k) + E(f_k^n) \frac{\|f_k - g_k^n\|_2^2}{2} \right\} \\
h_k^n &= \hat{h}_k^n - \text{median}(\hat{h}_k^n)\mathbf{1} \\
\hat{f}_k^{n+1} &= \frac{h_k^n}{\|h_k^n\|_2} \\
\hat{f}_k^{n+1}(i) &= \begin{cases} 1 & \text{for data points } i \text{ in class } A^k \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

  Simplex projection step [Michelot '86]:

$$
\mathbf{f}^{n+1} = \Pi_{\sum_{k=1}^{K} \hat{f}_k(i)=1 \text{ for } i} (\hat{\mathbf{f}}^{n+1}) \quad \text{where} \quad \mathbf{f} = (f_1, ..., f_K)
$$

- Note: Heuristic optimization → how to define a monotonic algorithm for multi-class ?

Spectral Transductive Clustering [Belkin-Niyogi '02]

- Two-class transductive clustering

  Step 1: Laplacian eigenvectors
  Compute the graph Laplacian and M eigenvectors: $\Phi = [\phi_1, ..., \phi_L]$.

  Step 2: Build linear classifier
  Find coefficient vector $\mathbf{a}$

$$\min_{\mathbf{a}} \underbrace{\sum_{i=1}^{n} (y_i^\ell - \sum_{j=1}^{M} a_j \phi_j(i))^2,}_{\|y^\ell - \Phi_\ell \mathbf{a}\|_2^2}$$

  Solution is given by

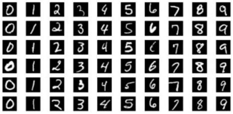
$$\mathbf{a} = (\Phi_\ell^T \Phi_\ell) \Phi_\ell^T y^\ell$$

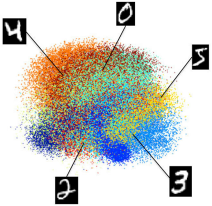
  Step 3: Estimate class of unlabeled data points

$$y_i = \begin{cases} 1 & \text{if} \quad \sum_{j=1}^{M} a_j \phi_j(i)) \geq 0 \\ -1 & \text{if} \quad \sum_{j=1}^{M} a_j \phi_j(i)) < 0 \end{cases}, \quad \text{for} \quad i = 1, ..., N - n$$

- Natural extension to multi-class

Experiment: MNIST





- MNIST has $N = 70,000$ data points and $K = 10$ classes. We cluster 60,000 unlabeled data points using n labeled data points.

labeled data per cluster (n/K)	1	5	10	50	100
Spectral Transductive Clustering	24.78%	8.08%	4.48%	2.82%	2.47%
TV Transductive Clustering	**2.43%**	**2.45%**	**2.45%**	**2.41%**	**2.38%**

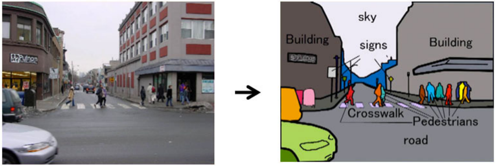
Table: Clustering error. First row indicates the number of labeled data, second row presents the spectral method and the last row the TV method. We have made 10 experiments and computed the mean percentage of misclassification. For each experiment, labeled points are randomly selected.

- The TV algorithm is more accurate and robust than the spectral algorithm, particularly when considering a few labeled data points.

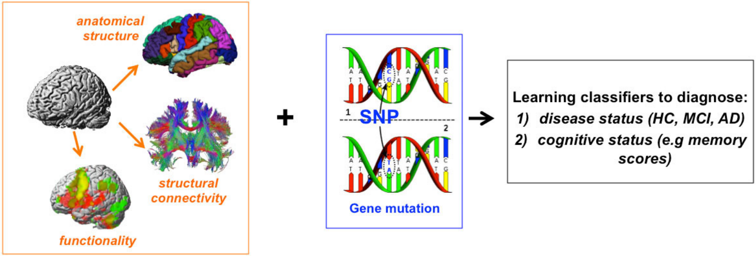
From Transductive Clustering to Semi-Supervised Classification

Data Classification

- Objective: assigning new data points to a class defined by some labeled data.
- Some applications:



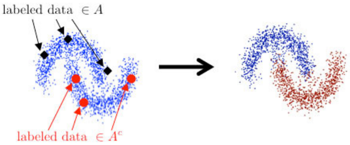
Machine vision
Object recognition



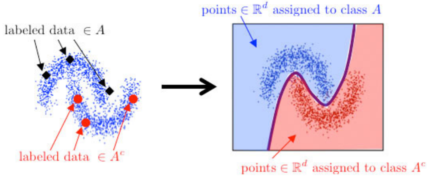
Neuroimaging
Alzheimer disease detection and analysis

Transductive Clustering and Classification

- Transductive clustering: partitioning data points into sensible groups given a small set of labeled data.
- Supervised classification: learning a classification function from labeled data points.
- Semi-Supervised classification: learning a classification function from labeled and unlabeled data points.



Transductive clustering
find sensible groups of data points



Semi-supervised classification
find classifier for new points

Mathematical formulation of Supervised Classification

- Objective: Given a set of N labeled data points $\{(x_i, y_i)\}$ with $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$, find a "good" classifier for new points.

- Optimization problem: Find function $f : \mathbb{R}^d \to \mathbb{R}$ as

$$\min_{f \in \mathcal{H}_K} \|f\|^2_{\mathcal{H}_K} + \gamma \sum_i \underbrace{V(f, x_i, y_i)}_{\text{loss function}}, \tag{4}$$

  where $\mathcal{H}_K$ is a Reproducing Kernel Hilbert Space (RKHS).

- Representer theorem: solution of (4) can be expressed as

$$f(x) = \sum_i \alpha_i K(x, x_i), \tag{5}$$

  where K is a positive definite kernel.

- Regularized Least Squares (RLS): $V = (y_i - f(x_i))^2$

  Plugging (5) into (4), we have [Smale-Poggio '00]:

$$\min_{\alpha \in \mathbb{R}^N} \alpha^T K \alpha + \gamma \|y - K\alpha\|_2^2$$

  which solution is given by solving a linear system of equations:

$$\alpha = (I_N + \gamma K)^{-1}(\gamma y)$$

Mathematical formulation of Semi-Supervised Classification [Belkin-Niyogi-Sindhwani '06]

- Objective: Given a (small) set of labeled data points $\{(x_i, y_i)\}_{1 \leq i \leq N-n}$ and a (large) set of unlabeled data points $\{x_i\}_{N-n+1 \leq i \leq N}$, find a "good" classifier that can adapt to the geometry of all (labeled and unlabeled) data points.

- Optimization problem: Find function $f : \mathbb{R}^d \to \mathbb{R}$ as

$$\min_{f \in \mathcal{H}_K} \|f\|^2_{\mathcal{H}_K} + \gamma \sum_i V(f, x_i, y_i) + \beta \underbrace{\|f\|_{\mathcal{M}}}_{\text{manifold regularization}}, \qquad (6)$$

such as $\|f\|_{\mathcal{M}} \approx \sum_{i,j} w_{ij} |f_i - f_j|^2 = f^T L f$ where L is the graph Laplacian.

- Representer theorem: if V is the quadratic loss, the solution of (6) can be expressed as $f(x) = \sum_i \alpha_i K(x, x_i)$ which leads to

$$\min_{\alpha \in \mathbb{R}^N} \ \alpha^T K \alpha + \gamma \|y - K\alpha\|_2^2 + \beta (K\alpha)^T L (K\alpha)$$

which solution is

$$\alpha = (I_N + \gamma K + \beta L K)^{-1} (\gamma y)$$

Note: α depend on the geometry of the manifold where the data points are sampled.

Support Vector Machine (SVM) Classification 1/4

- Linear SVM classification [Cortes-Vapnik '95]: given a set of N labeled data points $\{(x_i, y_i)\}$, $x_i \in \mathbb{R}^d, y_i \in \{-1, +1\}$, find a classifier f solution of:

$$\min_{f \in \mathcal{H}_K} \|f\|^2_{\mathcal{H}_K} \quad \text{s.t.} \quad \begin{cases} f_i - b \geq 1 \ \forall x_i \in A \text{ with } y_i = 1 \\ f_i - b \leq -1 \ \forall x_i \in A^c \text{ with } y_i = -1 \end{cases} \tag{7}$$

$$\Updownarrow$$

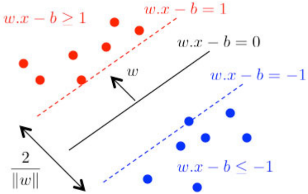
$$y_i(f_i - b) \geq 1 \quad \forall i \in V$$

where b is an offset vector. Assume the data points are linearly separable. The Representer Theorem guarantees existence of a classifier $f(x) = \sum_i \alpha_i K(x_i, x)$. If $K(x, y) = x.y$ then

$$f(x) = \sum_i \alpha_i x_i . x = \underbrace{\sum_i \alpha_i x_i}_{w} . x = w.x \tag{8}$$

Plugging (8) into (7) gives the standard linear SVM problem:

$$\min_{w \in \mathbb{R}^N} \|w\|^2_2 \quad \text{s.t.} \quad y_i(w.x_i - b) \geq 1 \quad \forall i,$$

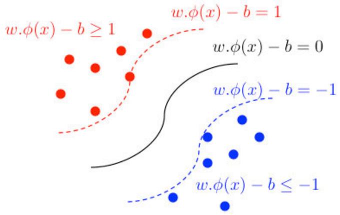


Support Vector Machine Classification 2/4

- Non-linear SVM classification [Cortes-Vapnik '95] with kernel trick [Aizerman-et.al. '64]

  Objective: learn a non-linear classifier using non-linear kernel functions.
  Examples of kernel:

$$K(x,y) = \begin{cases} e^{-\|x-y\|_2^2/\sigma} & \text{Gaussian kernel} \\ (1 + x.y/\sigma)^{\eta} & \text{Polynomial kernel} \end{cases}$$



  The kernel is related to the transform ϕ by the equation:

$$K(x_i, x_j) = \phi(x_i)\phi(x_j)$$

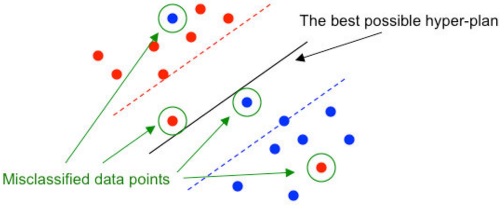
Support Vector Machine Classification 3/4

- Soft margin SVM [Cortes-Vapnik '95]: when data points are not separable, the problem changes to find a (non-linear) hyper-plane that separates the data points as clean as possible, while minimizing the number of misclassified data points.

  Optimization problem:

$$\min_{f \in \mathcal{H}_K} \|f\|^2_{\mathcal{H}_K} + \mu \sum_i \xi_i \quad \text{s.t.} \quad \begin{cases} y_i(f_i - b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases} \quad \forall i$$

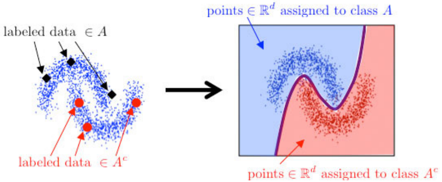
  where ξ_i are slack variables which measure the degree of misclassification of data x_i.



Support Vector Machine Classification 4/4

- Semi-supervised SVM/Laplacian SVM [Belkin-Niyogi-Sindhwani '06]: given labeled data points $\{(x_i, y_i)\}_{1 \le i \le N-n}$ and unlabeled data points $\{x_i\}_{N-n+1 \le i \le N}$, the SVM classifier is solution of the optimization problem:

$$\min_{f \in \mathcal{H}_K} \|f\|^2_{\mathcal{H}_K} + \mu \sum_i \xi_i + \beta \underbrace{\sum_{i,j} w_{ij} |f_i - f_j|^2}_{\text{Dirichlet energy}} \quad \text{s.t.} \quad \begin{cases} y_i(f_i - b) \ge 1 - \xi_i \\ \xi_i \ge 0 \end{cases} \quad \forall i$$



Semi-supervised classification
find classifier for new points

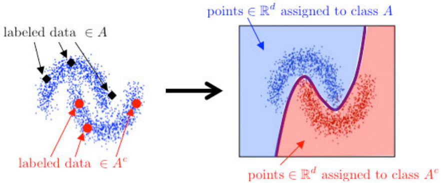
- The Dirichlet regularization is equivalent to a heat diffusion process $\Rightarrow$ classifier f cannot be a binary function !

Total Variation SVM Classification [B-Zhang '12]

- Objective: find tight/exact approximation of binary classification function.

- TV-SVM optimization:

$$\min_{f \in \mathcal{H}_K} \|f\|^2_{\mathcal{H}_K} + \mu \sum_i \xi_i + \beta \underbrace{\frac{\sum_{ij} w_{ij} |f_i - f_j|}{\sum_i |f_i - \text{median}(f)|}}_{\text{TV Clustering}} \quad \text{s.t.} \quad \begin{cases} y_i(f_i - b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases} \quad \forall i$$



Semi-supervised classification
find classifier for new points

- Exact relaxation? Probably no.

Optimization

- Algorithm

For n=1,2,... until convergence

$$\begin{aligned}
g^{n+1} &= f^n + sign(f^n) \\
e^{n+1} &= \mathrm{SVM}(g^{n+1}) \\
h^{n+1} &= \mathrm{argmin}_h \; TV(h) + \frac{E^n}{2}\|h - e^{n+1}\|_2^2 \\
t^{n+1} &= h^{n+1} - median(h^{n+1}) \\
t^{n+1} &= \begin{cases} 1 & \text{for data points } i \text{ in class } A \\ -1 & \text{for data points } i \text{ in class } A^c \end{cases} \\
f^{n+1} &= \frac{t^{n+1}}{\|t^{n+1}\|_2}
\end{aligned}$$

where SVM(g) is a standard SVM problem with a quadratic term defined as

$$\min_{e,\xi,b} \frac{\lambda}{2}||e||^2_{\mathcal{H}_K} + \mu \sum_{i \in N} \xi_i + \frac{r}{2}\|e - g\|_2^2 \quad \text{s.t.}$$

$$\begin{cases} y_i(e_i - b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases} \quad \forall i$$

- Note: Heuristic optimization.

Experiments

- Binary classification:

# labels per class	1	5	10	50
Lap-SVM (classification error %)	13.79	9.84	7.61	4.77
TV-SVM (classification error %)	**3.87**	**3.74**	**4.00**	**2.73**

Table: Binary semi-supervised classification algorithms tested on the sets of 4's and 9's from USPS dataset. The 4's has 652 training points and 200 test points and the 9's has 644 training points and 177 test points. Error is averaged over 10 runs with randomly selected labels.

- Multi-class classification:

# labels per class	1	5	10	50
Lap-SVM (classification error %)	49.95	14.21	6.27	2.82
TV-SVM (classification error %)	**2.94**	**2.08**	**1.72**	**1.74**

Table: Multi-class semi-supervised classification algorithms tested on four classes (0's, 1's, 4's and 9's) from USPS dataset. The 0's has 1194 training points and 359 test points and the 1's has 1005 training points and 264 test points. Error is averaged over 10 runs with randomly selected labels.

Related work

- *Nonlinear Eigenproblem approach:*

  • T. Buhler, M. Hein, "Spectral clustering based on the graph p-Laplacian", 2009

  • M. Hein, T. Buhler, "An inverse power method for nonlinear eigenproblems with applications in 1-spectral clustering and sparse PCA", 2010

  • M. Hein, S. Setzer, "Beyond Spectral Clustering-Tight Relaxations of Balanced Graph Cuts", 2011

  • S. Rangapuram and M. Hein, "Constrained 1-Spectral Clustering", 2012

- *TV Phase-field approach:*

  • A. Bertozzi and Arjuna Flenner, "Diffuse Interface Models on Graphs for Classification of High Dimensional Data", 2011

  • Y. van Gennip and A. Bertozzi, "Γ-Convergence of Graph Ginzburg-Landau Functionals", 2012

  • E. Merkurjev, T. Kostic and A. Bertozzi, "An MBO Scheme on Graphs for Segmentation and Image Processing", 2012

- *TV Transductive and supervised classification in imaging:*
  • S.H. Kang, B. Shafei, and G. Steidl, "Supervised and Transductive Multi-Class Segmentation Using p-Laplacians and RKHS methods", 2012.

Conclusion

- Total variation offers a powerful tool to find tight/exact relaxation of fundamental (NP-)hard data analysis problems: unsupervised, transductive clustering and semi-supervised classification.
- TV-based learning algorithms are more challenging to design because these optimization problems are non-differentiable and non-convex.
- Promising to improve functionalities of real-world products like search engines, social network analysis, neuroscience, etc.

Papers and Codes

- Papers:

  • A. Szlam and X. Bresson, "A Total Variation-based Graph Clustering Algorithm for Cheeger Ratio Cuts", CAM report 09-68, August 2009

  • A. Szlam and X. Bresson, "Total Variation and Cheeger Cuts", International Conference on Machine Learning (ICML), 1039-1046, 2010

  • X. Bresson, X.-C. Tai, T. F. Chan and A. Szlam, "Multi-Class Transductive Learning based on L1 Relaxations of Cheeger Cut and Mumford-Shah-Potts Model", 2012

  • X. Bresson and T. Laurent, "Asymmetric Cheeger Cut and Application to Multi-Class Unsupervised Clustering", CAM Report 12-27, April 2012

  • X. Bresson, T. Laurent, D. Uminsky and J.H. von Brecht, "Convergence and Energy Landscape for Cheeger Cut Clustering", Annual Conference on Neural Information Processing Systems (NIPS), 2012

  • X. Bresson and R. Zhang, "TV-SVM: Total Variation Support Vector Machine for Semi-Supervised Data Classification", 2012

  • X. Bresson, T. Laurent, D. Uminsky and J.H. von Brecht, "An Adaptive Total Variation Algorithm for Computing the Balanced Cut of a Graph", 2013

  • J. Wu, T. Laurent and X. Bresson, "TV Clustering for Unsupervised Document Retrieval", in preparation

- Codes are available on my website:
  http://www.cs.cityu.edu.hk/∼xbresson

Thank you !