Discrete vs. Continuous: Two Sides of Machine Learning

Dengyong Zhou

Department of Empirical Inference
Max Planck Institute for Biological Cybernetics
Spemannstr. 38, 72076 Tuebingen, Germany
Oct. 12, 2004

Our contributions

- Developed the discrete calculus and geometry on discrete objects
- Constructed the discrete regularization framework for learning on discrete spaces
- Proposed a family of transductive algorithms derived from the discrete framework

Outline

- Introduction
- Discrete analysis and regularization
- Related work
- Discussion and future work

Introduction

- Definition and motivations of transductive inference
- A principled approach to transductive inference

Problem setting

Consider a finite input space $\mathcal{X} = \{x_1, \ldots, x_l, x_{l+1}, \ldots, x_{l+k}\}$ and output space $\mathcal{Y} = \{-1, 1\}$.

- Observes the labels of the first l points: $(x_1, y_1), \ldots, (x_l, y_l)$.
- Predicts the labels of the given points: $x_{l+1}, \ldots, x_{l+k}$.

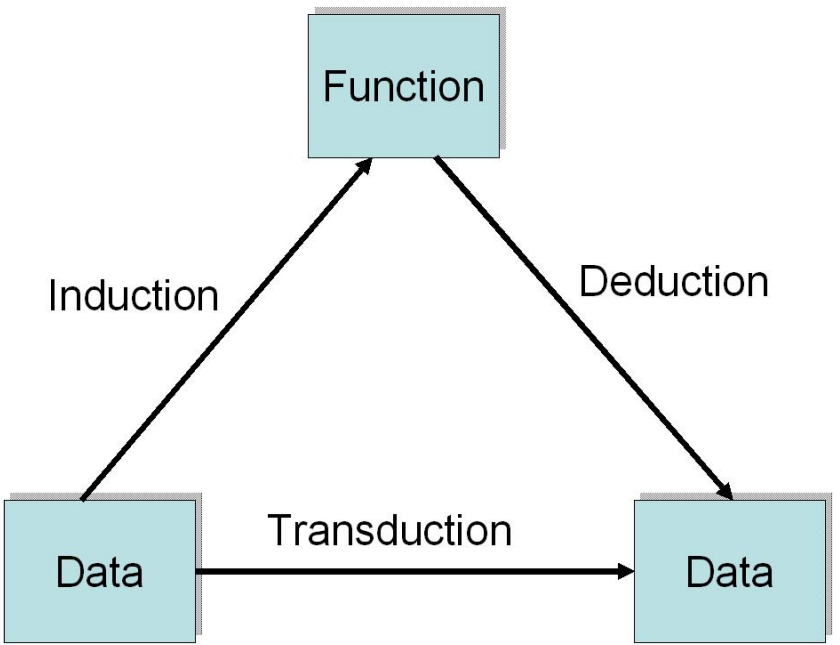
We may use any supervised learning classifiers, for instance, SVMs (`Boser et al., 1992; Cortes and Vapnik, 1995`), to estimate the labels of the given points. However, $\cdots$

Motivation in theory: Vapnik's philosophy and transduction (Vapnik, 1998)

- Do not solve a more general problem as an intermediate step. It is quite possible that you have enough information to solve a particular problem well, but have not enough information to solve a general problem.

- Transductive inference: Do not estimate a function defined on the whole space (inductive). Directly estimate the function values on the given points of interest (transductive)!

Vapnik's picture: Transduction vs. Induction



Further motivations: Learning with very few training examples

- Engineering Improving classification accuracy by using unlabeled data. Labeling needs expensive human labor, whereas unlabeled data is far easier to obtain.

- Cognitive science Understanding human inference. Humans can learn from very few labeled examples.

Introduction

- Definition and motivations of transductive inference
- A principled approach to transductive inference

A basic idea in supervised learning: Regularization

A common way to achieve smoothness (cf. `Tikhonov and Arsenin, 1977`):

$$\min_f \{ R_{\text{emp}}[f] + \lambda \Omega[f] \}$$

— $R_{\text{emp}}[f]$ is the empirical risk;
— $\Omega[f]$ is the regularization (stabilization) term;
— λ is the positive regularization parameter specifying the trade-off.

A basic idea in supervised learning: Regularization (cont.)

- Typically, the regularization term takes the form:

$$\Omega[f] = \|Df\|^2,$$

  where D is a differential operator, such as $D = \nabla$ (gradient).

- Kernel methods:

$$\Omega[f] = \|f\|_{\mathcal{H}}^2,$$

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

They are equivalent (cf. `Schölkopf and Smola, 2002`).

A principled approach to transductive inference

- Develop the discrete analysis and geometry on finite discrete spaces consisting of discrete objects to be classified

- Discretize the classical regularization framework used in the inductive inference. Then the transductive inference approaches are derived from the discrete regularization.

Transduction vs. induction: discrete vs. continuous regularizer

Outline

- Introduction
- Discrete analysis and regularization
- Related works
- Discussion and future works

Discrete analysis and regularization

- A basic differential calculus on graphs
- Discrete regularization and operators
  - ⋆ 2-smoothness $(p = 2$, discrete heat flow, linear)
  - ⋆ 1-smoothness $(p = 1$, discrete curvature flow, non-linear)
  - ⋆ ∞-smoothness $(p = \infty$, discrete large margin)

A prior assumption: pairwise relationships among points

- If there is no relation among the discrete points, then we cannot make any prediction which is statistically better than a random guess.

- Assume the pairwise relationships among points:

$$w : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+.$$

The set of points may be thought of as a weighted graph, where the weights of edges encode the pairwise relationships.

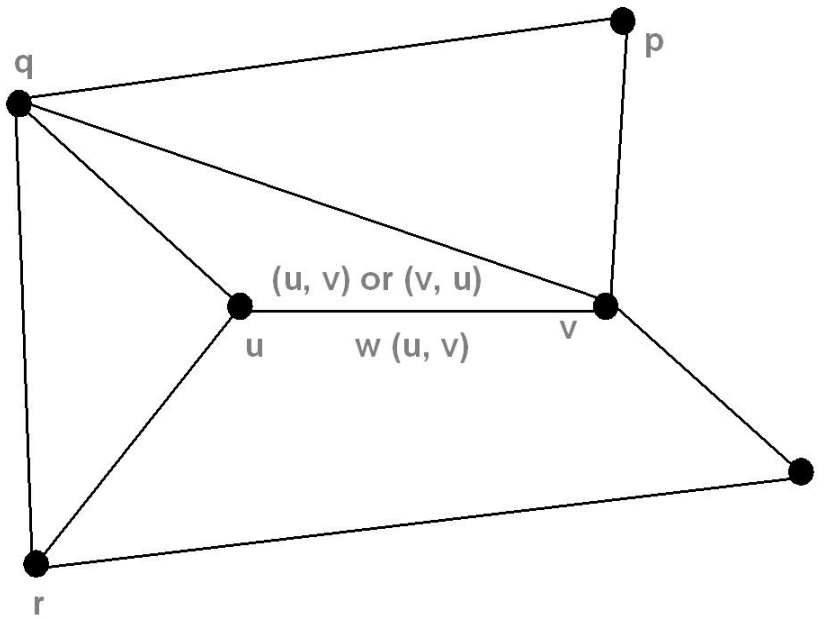
Some basic notions in graph theory

- A graph $\Gamma = (V, E)$ consists of a set V of vertices and a set of pairs of vertices $E \subseteq V \times V$ called edges.

- A graph is undirected if for each edge $(u, v) \in E$ we also have $(v, u) \in E$.

- A graph is weighted if it is associated with a function $w : E \to \mathbb{R}_+$ satisfying $w(u, v) = w(v, u)$.

Some basic notions in graph theory (cont.)



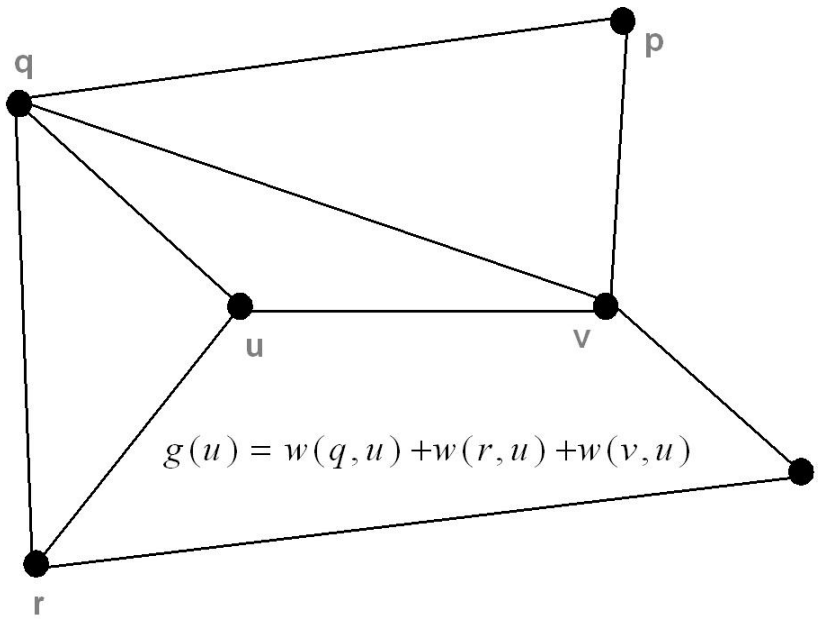
Some basic notions in graph theory (cont.)

- The degree function $g : V \to \mathbb{R}_+$ is defined to be

$$g(v) := \sum_{u \sim v} w(u, v),$$

where $u \sim v$ denote the set of vertices u connected to v via the edges (u, v). The degree can be regarded as a measure.

Some basic notions in graph theory (cont.)



The space of functions defined on graphs

- Let $\mathcal{H}(V)$ denote the Hilbert space of real-valued functions endowed with the usual inner product

$$\langle \varphi, \phi \rangle := \sum_{v} \varphi(v)\phi(v),$$

where φ and ϕ denote any two functions in $\mathcal{H}(V)$. Similarly define $\mathcal{H}(E)$. Note that function $\psi \in \mathcal{H}(E)$ need not be symmetric, i.e., we do not require $\psi(u, v) = \psi(v, u)$.

Gradient (or boundary) operator

- We define the graph gradient operator $d : \mathcal{H}(V) \rightarrow \mathcal{H}(E)$ to be (Zhou and Schölkopf, 2004)

$$(d\varphi)(u,v) := \sqrt{\frac{w(u,v)}{g(u)}}\varphi(u) - \sqrt{\frac{w(u,v)}{g(v)}}\varphi(v),$$

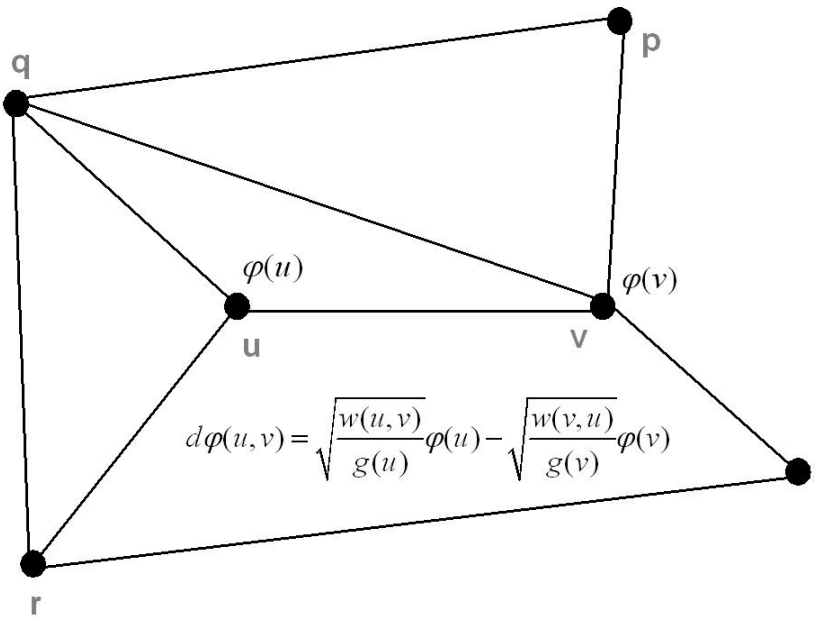
for all (u,v) in E.

Remark *In the lattice case, the gradient degrades into*

$$(d\varphi)(u,v) = \varphi(u) - \varphi(v),$$

which is the standard difference definition in numerical analysis.

Gradient (or boundary) operator



Divergence (or co-boundary) operator

- We define the adjoint $d^* : \mathcal{H}(E) \rightarrow \mathcal{H}(V)$ of d by (Zhou and Schölkopf, 2004)

$$\langle d\varphi, \psi \rangle = \langle \varphi, d^*\psi \rangle, \text{ for all } \varphi \in \mathcal{H}(V), \psi \in \mathcal{H}(E).$$

  We call d^* the graph divergence operator.

Note that the inner products are respectively in the space $\mathcal{H}(E)$ and $\mathcal{H}(V)$.

Divergence (or co-boundary) operator (cont.)

- We can show that d^* is given by

$$(d^*\psi)(v) = \sum_{u \sim v} \sqrt{\frac{w(u,v)}{g(v)}} \Big(\psi(v,u) - \psi(u,v) \Big).$$

Remark *The divergence of a vector field F in Euclidean space is defined by*

$$div(F) = \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z}.$$

The physical significance is the net flux flowing out of a point.

Edge derivative

- The edge derivative

$$\left.\frac{\partial}{\partial e}\right|_v : \mathcal{H}(V) \to \mathbb{R}$$

along edge $e = (v, u)$ at vertex v is defined by

$$\left.\frac{\partial \varphi}{\partial e}\right|_v := (d\varphi)(v, u).$$

Local smoothness measure

- Define the local variation of φ at v to be

$$\| \nabla_v \varphi \| := \left[\sum_{e \vdash v} \left(\frac{\partial \varphi}{\partial e} \Big|_v \right)^2 \right]^{1/2},$$

  where $e \vdash v$ denotes the set of edges incident on v.

Global smoothness measure

- Let $\mathcal{S}$ denote a functional on $\mathcal{H}(V)$, for any $p \in [1, \infty)$, which is defined to be

$$\mathcal{S}_p(\varphi) := \frac{1}{p} \sum_v \left\| \nabla_v \varphi \right\|^p,$$

and, especially,

$$\mathcal{S}_\infty(\varphi) := \max_v \left\| \nabla_v \varphi \right\|.$$

The functional $\mathcal{S}_p(\varphi)$ can be thought of as the measure of the smoothness of φ.

Discrete analysis and regularization

- A basic differential calculus on graphs
- Discrete regularization and operators
  - ⋆ 2-smoothness ($p = 2$, heat flow, linear);
  - ⋆ 1-smoothness ($p = 1$, curvature flow, non-linear);
  - ⋆ ∞-smoothness ($p = \infty$, large margin)
- Directed graphs

Discrete regularization on graphs

- Given a function y in $\mathcal{H}(V)$, the goal is to search for another function f in $\mathcal{H}(V)$, which is not only smooth enough on the graph but also close enough to the given function y. This idea is formalized via the following optimization problem:

$$\operatorname*{argmin}_{f \in \mathcal{H}(V)} \{\mathcal{S}_p(f) + \frac{\mu}{2} \|f - y\|^2\}.$$

- For classification problems, define $y(v) = 1$ or -1 if v is labeled as positive or negative and 0 otherwise. Each vertex v is finally classified as $\operatorname{sgn} f(v)$.

Discrete analysis and regularization

- A basic differential calculus on graphs
- Discrete regularization and operators
  - ⋆ 2-smoothness ($p = 2$, heat flow, linear);
  - ⋆ 1-smoothness ($p = 1$, curvature flow, non-linear);
  - ⋆ ∞-smoothness ($p = \infty$, large margin)
- Directed graphs

Laplacian operator

- By analogy with the Laplace-Beltrami operator on forms on Riemannian manifolds, we define the graph Laplacian $\Delta : \mathcal{H}(V) \to \mathcal{H}(V)$ by (Zhou and Schölkopf, 2004)

$$\Delta := \frac{1}{2} d^* d.$$

Remark *The Laplace in Euclidean space is defined by*

$$\Delta f = \frac{\partial^2 f_x}{\partial x^2} + \frac{\partial^2 f_y}{\partial y^2} + \frac{\partial^2 f_z}{\partial z^2}.$$

Laplacian operator (cont.)

- Laplacian is a self-adjoint and linear operator

$$\langle \Delta\varphi, \phi \rangle = \langle \frac{1}{2} d^* d\varphi, \phi \rangle = \frac{1}{2} \langle d\varphi, d\phi \rangle = \frac{1}{2} \langle \varphi, d^* d\phi \rangle = \langle \varphi, \Delta\phi \rangle.$$

- Laplacian is positive semi-definite

$$\langle \Delta\varphi, \varphi \rangle = \langle \frac{1}{2} d^* d\varphi, \varphi \rangle = \frac{1}{2} \langle d\varphi, d\varphi \rangle = \mathcal{S}_2(\varphi) \geq 0.$$

- Laplacian and smoothness

$$\Delta\varphi = \frac{\partial \mathcal{S}_2(\varphi)}{\partial \varphi}.$$

Laplacian operator (cont.)

- An equivalent definition of the graph Laplacian:

$$(\Delta\varphi)(v) := \frac{1}{2}\sum_{e\vdash v}\frac{1}{\sqrt{g}}\left(\frac{\partial}{\partial e}\sqrt{g}\frac{\partial\varphi}{\partial e}\right)\Bigg|_v .$$

  This is basically the discrete analogue of the Laplace-Beltrami operator based on the gradient.

Laplacian operator (cont.)

Computation of the graph Laplacian

- Substituting the definitions of gradient and divergence operators into that of Laplacian, we have

$$(\Delta\varphi)(v) = \varphi(v) - \sum_{u\sim v} \frac{w(u,v)}{\sqrt{g(u)g(v)}}\varphi(u).$$

Remark *In spectral graph theory* (`Chung, 1997`), *the graph Laplacian is defined as the matrix* $D^{-1/2}(D-W)D^{-1/2}$.

Solving the optimization problem $(p = 2)$

Theorem. [Zhou and Schölkopf, 2004; Zhou et al., 2003] *The solution f of the optimization problem*

$$\operatorname*{argmin}_{f \in \mathcal{H}(V)} \left\{ \frac{1}{2} \sum_{v} \left\| \nabla_v f \right\|^2 + \frac{\mu}{2} \| f - y \|^2 \right\}.$$

satisfies

$$\Delta f + \mu (f - y) = 0.$$

Corollary. $f = \mu (\mu I + \Delta)^{-1} y.$

An equivalent iterative algorithm

Isotropic information diffusion (or heat flow) (Zhou et al., 2003)

Define

$$p(u,v) = \frac{w(u,v)}{\sqrt{g(u)g(v)}}.$$

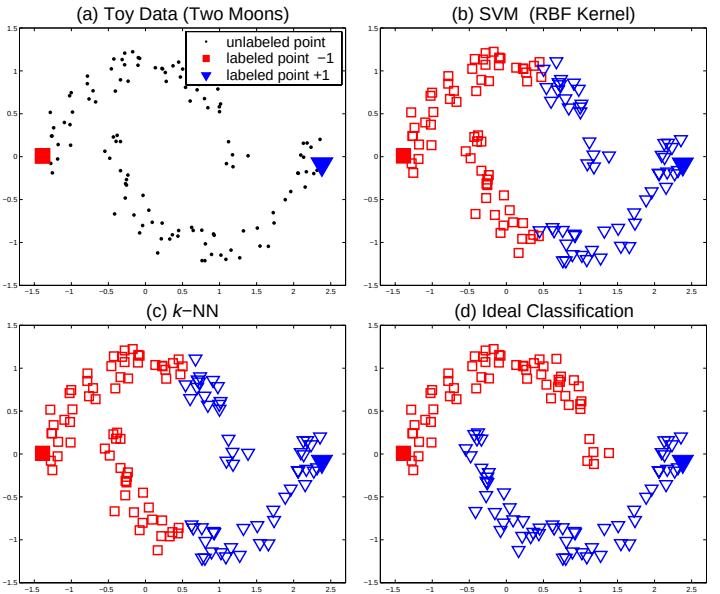
Then

$$f^{(t+1)}(v) = \sum_{u \sim v} \alpha p(u,v) f^t(u) + (1-\alpha) y(v),$$

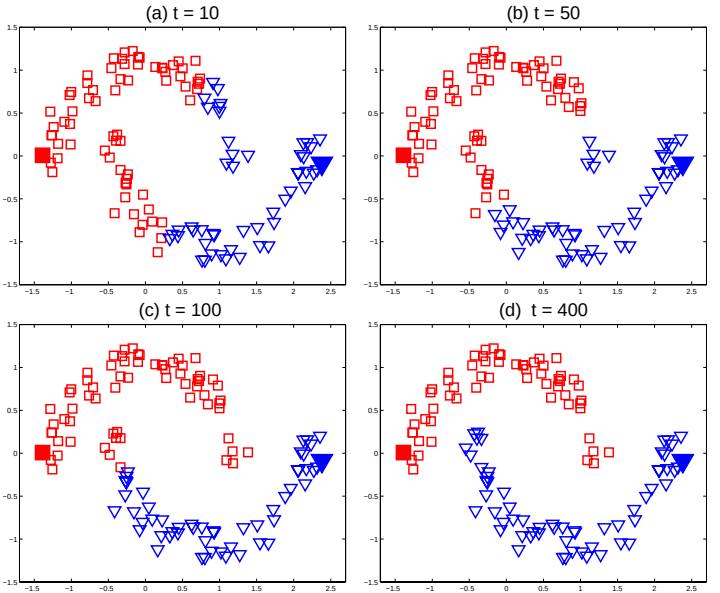
where α is a parameter in $(0,1)$.

Remark *See also* (Eells and Sampson, 1964) *for the heat diffusion on Riemannian manifolds.*

A toy classification problem

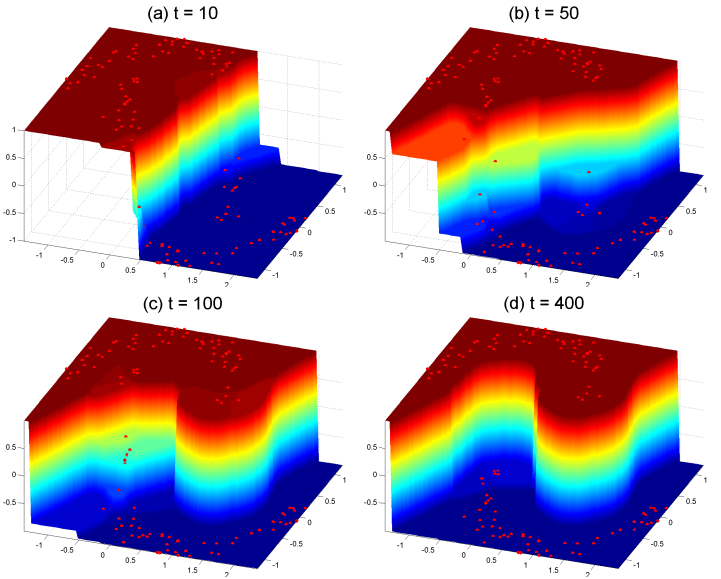


A toy classification problem (cont.)



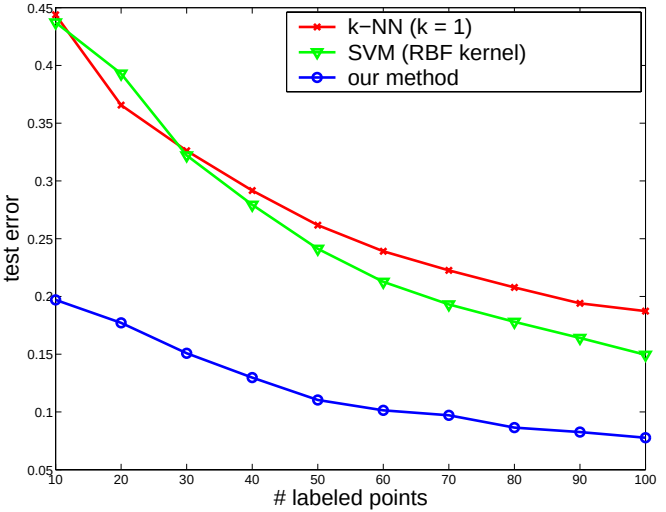
[Note: A fully connected graph: $w(u, v) = \exp(-\lambda \|u - v\|)$.]

A toy classification problem (cont.)



[Note: The function is becoming flatter and flatter.]

Handwritten digits recognition



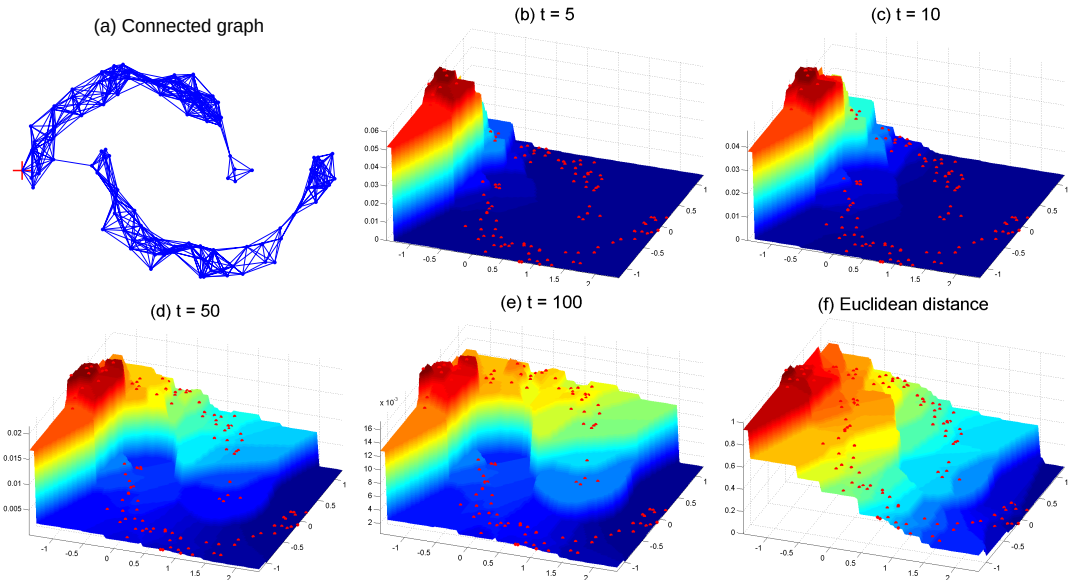
Digit recognition with USPS handwritten 16x16 digits dataset for a total of 9298. The left panel shows test errors for different algorithms with the number of labeled points increasing from 10 to 100.

Example-based ranking

- Given an input space $\mathcal{X} = \{x_0, x_1, \ldots, x_n\} \in \mathbb{R}^m$, the first point is the query. The goal is to rank the remaining points with respect to their relevances or similarities to the query. [See also (e.g., `Crammer and Singer, 2001; Freund et al., 2004`) for other ranking work in the machine learning community.]

- Define $y(v) = 1$ if vertex v is a query and 0 otherwise. Then rank each vertex v according to the corresponding function value $f(v)$ (largest ranked first).

A toy ranking problem



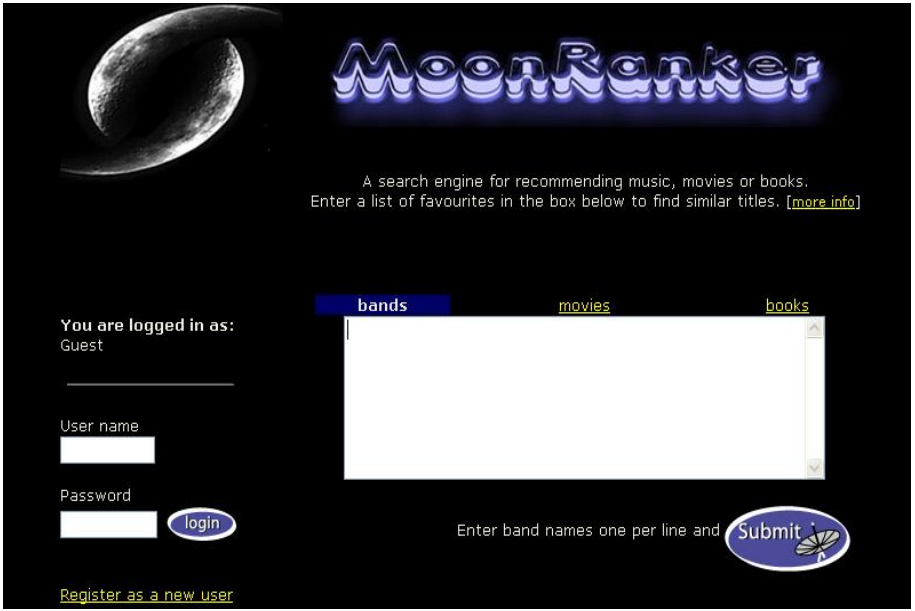
[Note: The shortest path based ranking does not work!]

Image ranking

Ranking digits in USPS. The top-left digit in each panel is the query. The left panel shows the top 99 by our method; and the right panel shows the top 99 by the Euclidean distance based ranking. Note that in addition to $\mathcal{3}$s there are many more $\mathcal{2}$s with knots in the right panel.

MoonRanker: A recommendation system

http://www.moonranker.com/



Protein ranking

(With J. Weston, A. Elisseeff, C. Leslie and W.S. Noble) **Protein ranking: from local to global structure in the protein similarity network**. PNAS 101(17) (2004).

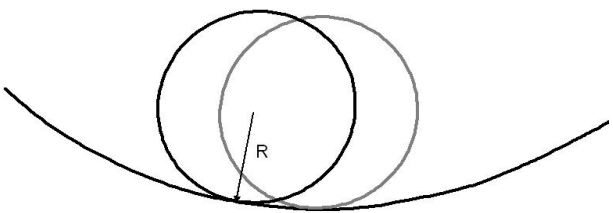
Discrete analysis and regularization

- A basic differential calculus on graphs
- Discrete regularization and operators
  - ⋆ 2-smoothness ($p = 2$, heat flow, linear);
  - ⋆ 1-smoothness ($p = 1$, curvature flow, non-linear);
  - ⋆ ∞-smoothness ($p = \infty$, large margin)
- Directed graphs

Curvature operator

- By analogy with the curvature of a curve which is measured by the change in the unit normal, we define the graph curvature $\kappa : \mathcal{H}(V) \to \mathcal{H}(V)$ by (Zhou and Schölkopf, 2004)

$$\kappa\varphi := d^* \left(\frac{d\varphi}{\|\nabla\varphi\|} \right).$$



Curvature operator (cont.)

- Computation of the graph curvature

$$(\kappa\varphi)(v) = \sum_{u\sim v} \frac{w(u,v)}{\sqrt{g(v)}}\left(\frac{1}{\|\nabla_v\varphi\|}+\frac{1}{\|\nabla_u\varphi\|}\right)\left(\frac{\varphi(v)}{\sqrt{g(v)}}-\frac{\varphi(u)}{\sqrt{g(u)}}\right)$$

  Unlike the graph Laplacian, the graph curvature is a non-linear operator.

Curvature operator (cont.)

- Another equivalent definition of the graph curvature based on the gradient:

$$(\kappa\varphi)(v) := \sum_{e \vdash v} \frac{1}{\sqrt{g}} \left(\frac{\partial}{\partial e} \frac{\sqrt{g}}{\|\nabla \varphi\|} \frac{\partial \varphi}{\partial e} \right) \Bigg|_v .$$

- An elegant property of the graph curvature

$$\kappa\varphi = \frac{\partial \mathcal{S}_1(\varphi)}{\partial \varphi}.$$

Solving the optimization problem $(p = 1)$

Theorem. [Zhou and Schölkopf, 2004] *The solution of the optimization problem*

$$\operatorname*{argmin}_{f \in \mathcal{H}(V)} \left\{ \sum_v \left\| \nabla_v f \right\| + \frac{\mu}{2} \|f - y\|^2 \right\}$$

satisfies

$$\kappa f + \mu(f - y) = 0.$$

No closed form solution.

An iterative algorithm

Anisotropic information diffusion (curvature flow)(Zhou and Schölkopf, 2004)

$$f^{(t+1)}(v) = \sum_{u \sim v} p^{(t)}(u, v) f^{(t)}(v) + p^{(t)}(v, v) y(v), \ \forall v \in V$$

Remark *The weight coefficients $p(u, v)$ are adaptively updated at each iteration, in addition to the classifying function being updated. This weight update causes the diffusion inside clusters to be enhanced, and the diffusion across clusters to be reduced.*

Compute the iteration coefficients: step 1

Compute the new weights $m : E \to \mathbb{R}$ defined by by

$$m(u,v) = w(u,v)\left(\frac{1}{\|\nabla_u f\|} + \frac{1}{\|\nabla_v f\|}\right).$$

Remark *The smoother the function f at nodes u and v, the larger the function m at edge (u,v).*

Compute the iteration coefficients: step 2

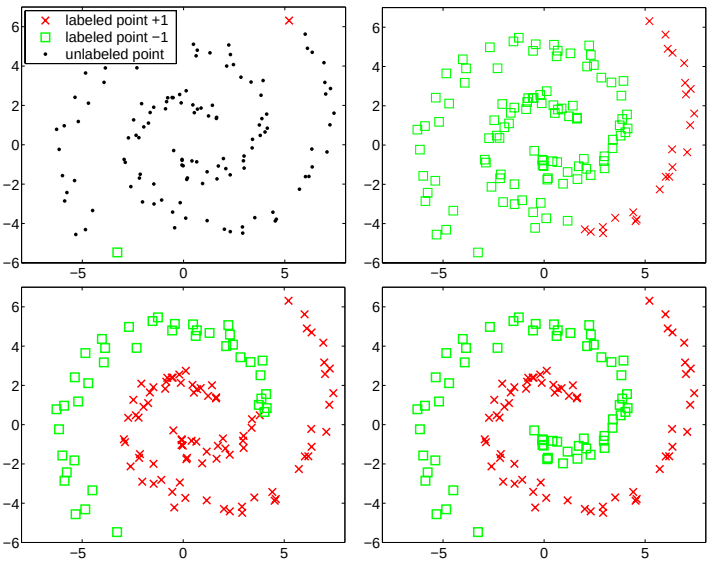
Compute the coefficients

$$p(u,v) = \frac{\frac{m(u,v)}{\sqrt{g(u)g(v)}}}{\sum_{u \sim v} \frac{m(u,v)}{g(v)} + \mu}, \text{ if } u \neq v;$$

and

$$p(v,v) = \frac{\mu}{\sum_{u \sim v} \frac{m(u,v)}{g(v)} + \mu}.$$

A toy classification problem



Classification on the spiral toy data. Top-left: toy data; top-right: spectral clustering; bottom-left: 2-smoothness; bottom-right: 1-smoothness.

Discrete analysis and regularization

- A basic differential calculus on graphs
- Discrete regularization and operators
  - ⋆ 2-smoothness ($p = 2$, heat flow, linear);
  - ⋆ 1-smoothness ($p = 1$, curvature flow, non-linear);
  - ⋆ ∞-smoothness ($p = \infty$, large margin)
- Directed graphs

Discrete large margin classification $(p = \infty)$

Discrete large margin (Zhou and Schölkopf, 2004):

$$\underset{f \in \mathcal{H}(V)}{\operatorname{argmin}} \{ \max_{v} \left\| \nabla_v f \right\| + \frac{\mu}{2} \| f - y \|^2 \}.$$

Only the worst case is considered!

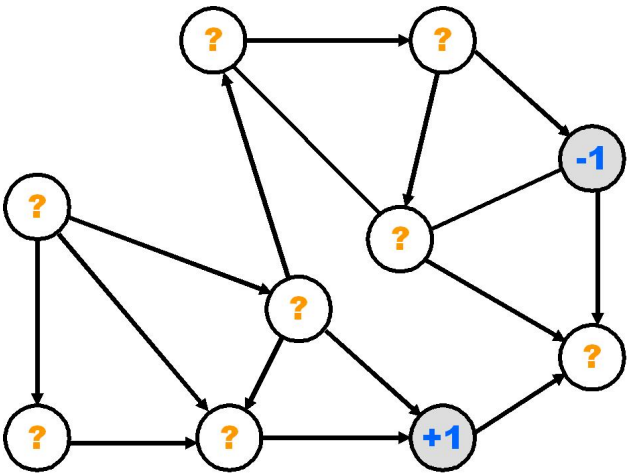
Remark *This is closely related to the classic graph bandwidth problem in combinatorial mathematics* (cf. Linial, 2002), *which is a NP-hard problem and has a polylogarithmic approximation.*

Discrete analysis and regularization

- Graph, gradient and divergence
- Discrete regularization and operators
  - ⋆ 2-smoothness ($p = 2$, heat flow, linear);
  - ⋆ 1-smoothness ($p = 1$, curvature flow, non-linear);
  - ⋆ ∞-smoothness ($p = \infty$, large margin)
- Directed graphs

Classification and ranking on directed graphs

- The differential geometry on undirected graphs can be naturally generalized to directed graphs (Zhou et al., 2004).

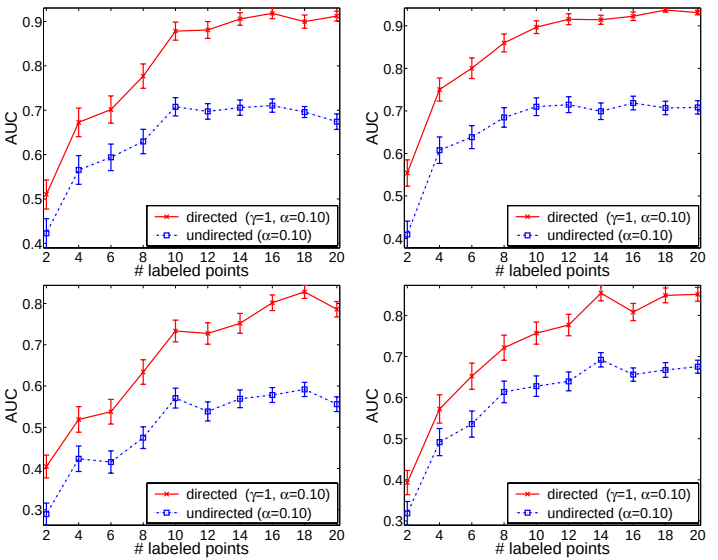


Two key observations on WWW

- The pages in a densely linked subgraph perhaps belong to a common topic. Therefore it is natural to force the classification function to change slowly on densely linked subgraphs.

- The pairwise similarity is measured based on the mutual reinforcement relationship between hub and authority (`Kleinberg, 1998`): a good hub node points to many good authorities and a good authority node is pointed to by many good hubs.

The importance of directionality



Classification on the WebKB dataset: student vs. the rest in each university. Taking the directionality of edges into account can yield substantial accuracy gains.

Outline

- Introduction
- Discrete analysis and regularization
- Related works
- Discussion and future works

A closely related regularizer

- The 2-smoothness regularizer can be rewritten as (Zhou et al, 2003)

$$\sum_{u,v} w(u,v)\left(\frac{f(u)}{\sqrt{g(u)}} - \frac{f(v)}{\sqrt{g(v)}}\right)^2$$

- A closely related one

$$\sum_{u,v} w(u,v)(f(u) - f(v))^2$$

  is proposed by (Belkin and Niyogi, 2002, 2003; Zhu et al., 2003). See also (Joachims, 2003) for a similar one.

Similarities between the two regularizers

- Both can be rewritten into the quadratic forms:

$$\sum_{u,v} w(u,v)\left(\frac{f(u)}{\sqrt{g(u)}} - \frac{f(v)}{\sqrt{g(v)}}\right)^2 = f^T D^{-\frac{1}{2}}(D-W)D^{-\frac{1}{2}} f$$

and

$$\sum_{u,v} w(u,v)(f(u)-f(v))^2 = f^T(D-W)f.$$

- Both $D^{-\frac{1}{2}}(D-W)D^{-\frac{1}{2}}$ and $D-W$ are called the graph Laplacian (unfortunate truth).

Differences between the two regularizers: limit cases

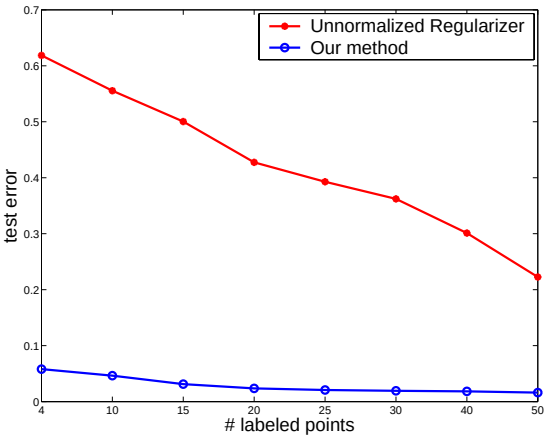
- (Bousquet et al., 2003) showed the following limit consequence:

$$\sum_{u,v} w(u,v)(f(u) - f(v))^2 \rightarrow \int \|\nabla f(x)\|^2 p^2(x) dx.$$

- A conjecture:

$$\sum_{u,v} w(u,v)\left(\frac{f(u)}{\sqrt{g(u)}} - \frac{f(v)}{\sqrt{g(v)}}\right)^2 \rightarrow \int \|\nabla f(x)\|^2 p(x) dx.$$

Difference between the two regularizers: experiments (cont.)



[Note: A subset of USPS containing the digits from 1 to 4; the same RBF kernel for all methods.]

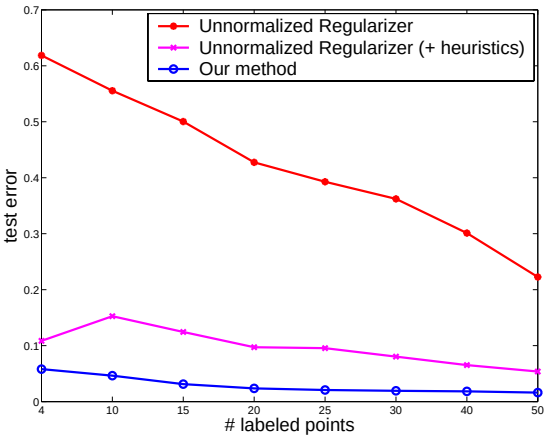
Improve the unnormalized regularizer by heuristics

- (Belkin and Niyogi, 2002, 2003) Choose a number k and construct a k-NN graph with $0/1$ weights over points. Using the weight matrix as the affinity among points.

- (Zhu et al., 2003) Estimate the proportion of different classes based on the labeled points, and then rescale the function based on the estimated proportion.

Both of them just empirically approximate to the normalization in our 2-smoothness regularizer.

Improve the unnormalized regularizer by heuristics: experiments



[Note: A subset of USPS containing the digits from 1 to 4; the same RBF kernel for all methods.]

Another related work: graph/cluster kernels

- Graph or cluster kernels (Smola and Kondor, 2003; Chapelle et al., 2002): Descompose the (normalized) graph Laplacian $K = U^T \Lambda U$ and then replace the eigenvalues λ with $\varphi(\lambda)$, where φ is a decreasing function, to obtain the so-called graph kernel:

$$\tilde{K} = U^T \text{diag}[\varphi(\lambda_1), \ldots, \varphi(\lambda_n)]U.$$

- In the closed form of the $p = 2$ case, the matrix $(\mu I + \Delta)^{-1}$ can be viewed as a graph kernel with $\varphi(\lambda) = 1/(\mu + \lambda)$.

Difference from graph/cluster kernels

- The matrix $(\mu I + \Delta)^{-1}$ is naturally derived from our regularization framework for transductive inference. In contrast, graph/cluster kernels are obtained by manipulating the eigenvalues.

- SVM combined with $(\mu I + \Delta)^{-1}$ does not work well in our transductive experiments.

- When p takes other values, e.g. $p = 1$, no corresponding kernel exists any more.

Outline

- Introduction to learn on discrete spaces
- Discrete analysis and regularization
- Related works
- Limitation and future works

Limitation: How to beat our method?

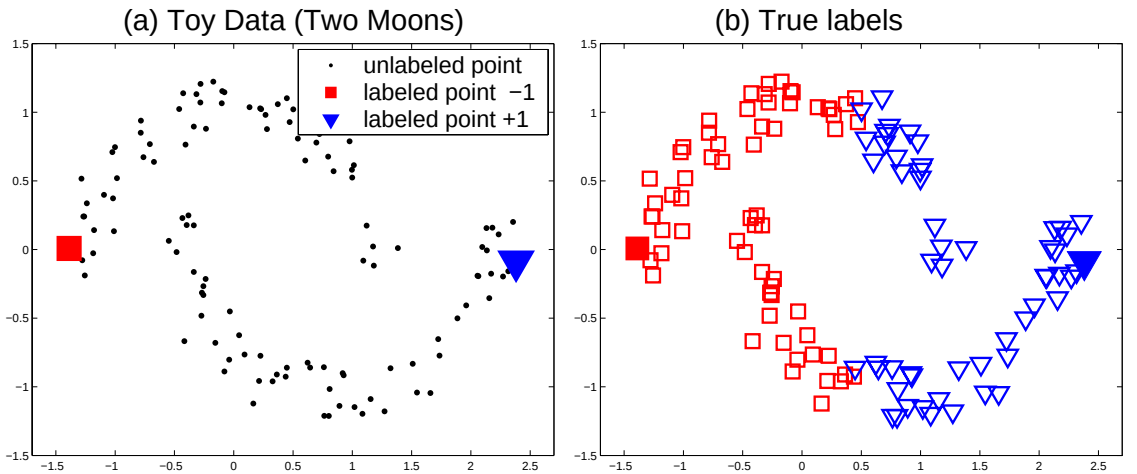
One can construct arbitrarily bad problems for a given algorithm:

Theorem. [No Free Lunch, e.g., Devroye, 1996] *For any algorithm, any n and any $\epsilon > 0$, there exists a distribution P such that $R^* = 0$ and*

$$\mathbb{P}\left[R(g_n) \geq \frac{1}{2} - \epsilon \right] = 1,$$

where g_n is the function estimated by the algorithm based on the n training examples.

Limitation: How to beat our method? (cont.)



Future work: Theory

We need a new statistical learning theory:

- How is a graph like a manifold?
  The convergence of the discrete differential operators
- How does transductive inference converge?
  Parallel to the bounds given in the context of inductive inference
- What is the transductive principle?
  Parallel to the inductive inference principle Structural Risk Minimization

Future work: Algorithms

From kernelize to transductize (or discretize):

- Learning with graph data (undirected, directed, bipartite, hypergraph), time series data, . . .
- Multi-label learning, hierarchical classification, regression, . . .
- Active learning, selection problems, . . .
- Unsupervised learning combined with partial prior knowledge, such as clustering, manifold learning, . . .

  . . .

Future work: Applications

- Computational biology
- Web information retrieval
- Natural language processing

. . .

Future work: Beyond machine learning

The discrete differential calculus and geometry over discrete objects provides the "exact" computation of the differential operators, and can be applied to more problems:

- Images processing/Computer graphics: Digital images/graphics are represented as lattices/graphs
- Structure and evolution of the real-world networks: WWW, Internet, biological networks, . . . (e.g., what do the curvatures of these networks mean?)

. . .

This talk is based on our following work:

- Differential Geometry *D. Zhou and B. Schölkopf.* Transductive Inference with Graphs. Technical Report, Max Planck Institute for Biological Cybernetics, August, 2004.
- Directed Graphs *D. Zhou, B. Schölkopf and T. Hofmann.* Semi-supervised Learning on Directed Graphs. **NIPS** 2004.
- Undirected Graphs *D. Zhou, O. Bousquet, T.N. Lal, J. Weston and B. Schölkopf.* Learning with Local and Global Consistency. **NIPS** 2003.
- Ranking *D. Zhou, J. Weston, A. Gretton, O. Bousquet and B. Schölkopf.* Ranking on Data Manifolds. **NIPS** 2003.
- Bioinformatics *J. Weston, A. Elisseeff, D. Zhou, C.Leslie and W.S. Noble.* Protein ranking: from local to global structure in the protein similarity network. **PNAS** 101(17) (2004).
- Bioinformatics *J. Weston, C. Leslie, D. Zhou, A. Elisseeff and W. S. Noble.* Semi-Supervised Protein Classification using Cluster Kernels. **NIPS** 2003.

Conclusions

- Developed the discrete analysis and geometry over discrete objects
- Constructed the discrete regularizer and the effective transductive algorithms are derived
- Validated the transductive algorithms on many real-world problems

Transduction and Induction are the two sides of machine learning: discrete vs. continuous. Two sides of the same.

. . . perhaps the success of the Heisenberg method points to a purely algebraic method of description of nature, that is, to the elimination of continuous functions from physics. Then, however, we must give up, by principle, the space-time continuum . . .

— Albert Einstein

Although there have been suggestions that space-time may have a discrete structure I see no reason to abandon the continuum theories that have been so successful.

— Stephen Hawking