Beyond the Point Cloud: From Transductive to Semi-Supervised Learning

Vikas Sindhwani, Partha Niyogi, Mikhail Belkin

Andrew B. Goldberg
goldberg@cs.wisc.edu

Department of Computer Sciences
University of Wisconsin, Madison

Stat 860 November 27, 2007

Outline

Outline

Introduction to Semi-Supervised Learning (SSL)

- In many real world classification tasks, labeled data is expensive.
- Unlabeled data, however, is often freely and readily available.
  - Examples: crawled Web pages, image search results, speech recordings
- Semi-supervised learning tries to use unlabeled data to learn better classifiers.

Typical SSL Setup

Given

- l labeled data points $\{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\}$, where each $x_i \in \mathcal{X}$ and $y_i \in \{-1, +1\}$.
- u unlabeled data points $\{x_{l+1}, x_{l+2}, \ldots, x_{l+u}\}$.
- (for future reference, $n = l + u$)

Do

- (Transduction) Predict labels $\{y_{l+1}, y_{l+2}, \ldots, y_{l+u}\}$.
- (True SSL) Learn $f : \mathcal{X} \mapsto \mathbb{R}$

Transductive vs. Semi-Supervised

Transductive

- Labeled and unlabeled data form point cloud.
- Simply learn a function over the point cloud.
- Classic example (on board)

Semi-supervised

- Also uses both labeled and unlabeled data during training.
- But learns function defined over *entire* space.
- Can make predictions for *unseen test data*.

Why SSL Algorithms (May) Work

Key Assumptions

Most SSL methods make one or both of the following assumptions:

- *Manifold assumption*: classification function is smooth with respect to the underlying marginal data distribution (estimated by unlabeled data).
- *Cluster assumption*: classes form distinct clusters that are separated by low density regions (i.e., areas where there is no unlabeled data).

A Few Classes of SSL Methods

SSL Methods

- Self-training
- Expectation maximization for Gaussian Mixture Models
- Cluster-then-label
- Co-training or multi-view methods
- Graph-based methods*
- Manifold regularization*

* closely related to today's talk

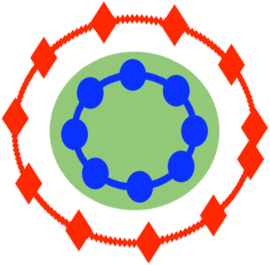
See Also

http://pages.cs.wisc.edu/~jerryzhu/research/ssl/semireview.html

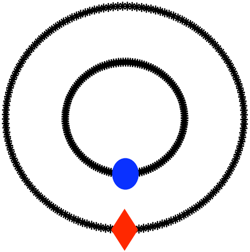
Outline

Two Concentric Circles Example

(a) two classes on concentric circles



(b) two labeled points



Typical Kernel-Based Approach

- Use Gaussian (RBF) kernel $k(x,z) = e^{-\frac{||x-z||^2}{2\sigma^2}}$. k defines RKHS $\mathcal{H}$.
- Learning involves solving a regularization problem:

$$f = \arg\min_{h \in \mathcal{H}} \frac{1}{l} \sum_{i=1}^{l} V(h, x_i, y_i) + \gamma ||h||_{\mathcal{H}}^2$$

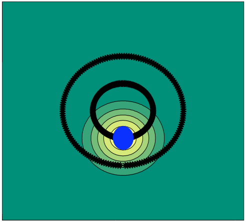
  where $||h||_{\mathcal{H}}$ is the RKHS norm, and V is a loss function (square loss for RLS, hinge loss for SVM)
- Representer theorem tells us solution has the form:

$$f(x) = \sum_{i=1}^{l} \alpha_i k(x, x_i)$$

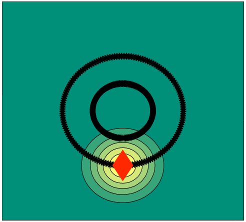
Result Using a Gaussian Kernel

- For two labeled points, the learned function is a linear combination of two Gaussians: (a) and (b).
- Gaussian kernel has spherical symmetry, so the end result is a linear decision boundary: (c).

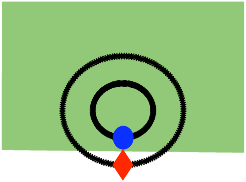
(a) gaussian kernel centered on labeled point 1



(b) gaussian kernel centered on labeled point 2



(c) classifier learnt in the RKHS



Graph-Based Semi-Supervised Learning

- Graph-based SSL creates nearest-neighbor graph of all data (edge weights W_{ij} or 0, if not neighbors). Then solve:

$$\arg\min_{\mathbf{f}} \frac{1}{l}\sum_{i=1}^{l}(f_i - y_i)^2 + \frac{\gamma}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(f_i - f_j)^2$$

- Manifold regularization (MR) solves a related RKHS problem:

$$\arg\min_{h\in\mathcal{H}} \frac{1}{l}\sum_{i=1}^{l} V(h, x_i, y_i) + \gamma_A ||h||_{\mathcal{H}}^2 + \frac{\gamma_I}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(h(x_i) - h(x_j))^2$$

- This paper solves MR problem using a special kernel.

Outline

Properties of New Kernel

Question

Can we define a kernel $\tilde{k}$ that is adapted to the geometry of the underlying data distribution?

Key properties

- $\tilde{k}$ should be valid kernel and define a new RKHS $\tilde{\mathcal{H}}$.
- $\tilde{k}$ should implement geometric intuitions (separate the two circles)
- Want to solve problem in new RKHS $\tilde{\mathcal{H}}$:

$$g = \underset{h \in \tilde{\mathcal{H}}}{\arg\min} \frac{1}{l} \sum_{i=1}^{l} V(h, x_i, y_i) + \gamma ||h||_{\tilde{\mathcal{H}}}^2$$

with a solution that's still a kernel expansion using *only the labeled points*: $g(x) = \sum_{i=1}^{l} \alpha_i \tilde{k}(x, x_i)$.

Two Circles Example: Desired Decision Surface

- Want solution that's a kernel expansion using only labeled points.

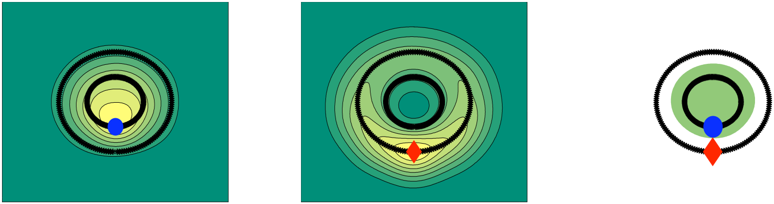
$$g(x) = \sum_{i=1}^{l} \alpha_i \tilde{k}(x, x_i)$$

but produces a circular decision boundary.

(a) deformed kernel centered on labeled point 1

(b) deformed kernel centered on labeled point 2

(c) classifier learnt in the deformed RKHS



- How is this possible? Stay tuned...

Properties of New Kernel

General strategy for getting intuitive decision surface with *supervised* kernel methods:

- Deform the original RKHS to obtain $\tilde{\mathcal{H}}$.
- Use unlabeled data to estimate marginal distribution.
- Derive explicit expression for $\tilde{k}$ in terms of unlabeled data.
- Solve regularization problem with only labeled data in $\tilde{\mathcal{H}}$.

Novel contributions:

- First truly data-dependent non-parametric kernel defined over all data points (true semi-supervised learning).
- General class of algorithms that can be customized with different base kernels, loss functions, etc.

Outline

Review of RKHS Background

- X is compact domain in Euclidean space or a manifold
- $\mathcal{H}$ is a complete Hilbert space of functions $X \mapsto \mathbb{R}$, with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$
- $\mathcal{H}$ is an RKHS if point evaluation functionals are bounded:
  - For any $x \in X, f \in \mathcal{H}, \exists C$, s.t. $|f(x)| \leq C||f||_{\mathcal{H}}$
- By Riesz representation theorem, can construct symmetric positive semi-definite kernel $k(x, z)$ s.t.

$$f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}} \quad k(x, z) = \langle k(x, \cdot), k(z, \cdot) \rangle_{\mathcal{H}}$$

Game plan

Show how general procedure to "deform" norm $|| \cdot ||_{\mathcal{H}}$ creates new RKHS $\tilde{\mathcal{H}}$ with $\tilde{k}(x, z)$

Defining the warped RKHS $\tilde{\mathcal{H}}$

- Let $\mathcal{V}$ be a linear space with positive semi-definite inner product (i.e., quadratic form)
- Let $S : \mathcal{H} \mapsto \mathcal{V}$ be a bounded linear operator
- *Define* $\tilde{\mathcal{H}}$ to be space of the *same* functions as $\mathcal{H}$ but modified inner product:

$$\langle f, g \rangle_{\tilde{\mathcal{H}}} = \langle f, g \rangle_{\mathcal{H}} + \langle Sf, Sg \rangle_{\mathcal{V}}$$

- Proposition: $\tilde{\mathcal{H}}$ is an RKHS (i.e., complete with bounded point evaluations)
- Proof: Straightforward result due to $\tilde{\mathcal{H}}$ and $\mathcal{H}$ containing the same elements. (details in paper)

Point-Cloud Norms

- Recall:

$$\langle f, g \rangle_{\tilde{\mathcal{H}}} = \langle f, g \rangle_{\mathcal{H}} + \langle Sf, Sg \rangle_{\mathcal{V}}$$

- In general case, difficult to connect k and $\tilde{k}$.
- We care only about the case when S and $\mathcal{V}$ depend on the data.
- For "point-cloud norms," we can express the relation explicitly (next few slides).
- Goal is to find modification to standard kernel that relies on the geometry of unlabeled data.

Incorporating the Data in the Kernel

- Recall:
$$\langle f, g\rangle_{\tilde{\mathcal{H}}} = \langle f, g\rangle_{\mathcal{H}} + \langle Sf, Sg\rangle_{\mathcal{V}}$$
- Given data $x_1, x_2, \ldots, x_n$, and let $\mathcal{V} = \mathbb{R}^n$
- Let $S : \mathcal{H} \mapsto \mathbb{R}^n$ be the evaluation map $S(f) = \mathbf{f} = (f(x_1), \ldots, f(x_n))$.
- Thus, we can write semi-norm on $\mathbb{R}^n$ using some s.p.d. matrix M:

$$||Sf||^2_{\mathcal{V}} = \mathbf{f}^\top M \mathbf{f}$$

Modified Regularization Problem

- Recall: $\langle f, g\rangle_{\tilde{\mathcal{H}}} = \langle f, g\rangle_{\mathcal{H}} + \langle Sf, Sg\rangle_{\mathcal{V}}, \qquad ||Sf||_{\mathcal{V}}^2 = \mathbf{f}^\top M \mathbf{f}$
- The regularization problem:

$$f = \arg\min_{h \in \tilde{\mathcal{H}}} \frac{1}{l} \sum_{i=1}^{l} V(h, x_i, y_i) + \gamma ||h||_{\tilde{\mathcal{H}}}^2$$

thus becomes

$$f = \arg\min_{h \in \mathcal{H}} \frac{1}{l} \sum_{i=1}^{l} V(h, x_i, y_i) + \gamma(||h||_{\mathcal{H}}^2 + \mathbf{h}^\top M \mathbf{h})$$

- Note that $\mathbf{h}$ is based on labeled and unlabeled data. M can encode smoothness w.r.t. graph/manifold.
- We'll now show how to solve the first problem directly using an explicit form for $\tilde{k}$.

Deriving an Explicit form for $\tilde{k}(x, z)$

Outline for deriving $\tilde{k}(x, z)$:

- Show that
$$span\{k(x_i, \cdot)\}_{i=1}^{n} = span\{\tilde{k}(x_i, \cdot)\}_{i=1}^{n}$$
- This leads to
$$\tilde{k}(x, \cdot) = k(x, \cdot) + \sum_{j=1}^{l+u} \beta_j(x) k(x_j, \cdot)$$
- Solve linear system involving all data to find $\beta_j(x)$ coefficients.
- Then we can compute $\tilde{k}(x, z)$ explicitly.

Deriving an Explicit form for $\tilde{k}(x, z)$

- Decompose $\tilde{\mathcal{H}}$ orthogonally as:

$$\tilde{\mathcal{H}} = span\{\tilde{k}(x_1, \cdot), \ldots, \tilde{k}(x_n, \cdot)\} \oplus \tilde{\mathcal{H}}^\perp$$

where $\tilde{\mathcal{H}}^\perp$ contains functions equal 0 at all data points.

- Thus, for $f \in \tilde{\mathcal{H}}^\perp, Sf = 0$, and $\langle f, g, \rangle_{\tilde{\mathcal{H}}} = \langle f, g \rangle_{\mathcal{H}}$ for any g.
- As a result, for any $f \in \tilde{\mathcal{H}}^\perp$,

$$f(x) = \langle f, \tilde{k}(x, \cdot) \rangle_{\tilde{\mathcal{H}}} = \langle f, k(x, \cdot) \rangle_{\mathcal{H}} = \langle f, k(x, \cdot) \rangle_{\tilde{\mathcal{H}}}$$

- Thus, $\langle f, k(x, \cdot) - \tilde{k}(x, \cdot) \rangle_{\tilde{\mathcal{H}}} = 0$ or $k(x, \cdot) - \tilde{k}(x, \cdot) \in (\tilde{\mathcal{H}}^\perp)^\perp$.

Deriving the Warped Kernel

- Because $k(x,\cdot) - \tilde{k}(x,\cdot) \in (\tilde{\mathcal{H}}^{\perp})^{\perp}$, we can write

$$k(x,\cdot) - \tilde{k}(x,\cdot) \in span\{\tilde{k}(x_1,\cdot),\ldots,\tilde{k}(x_n,\cdot)\}$$

- But, for any $x_i \in X$ and $f \in \tilde{\mathcal{H}}^{\perp}$, $\langle k(x_i,\cdot), f\rangle_{\tilde{\mathcal{H}}} = \langle k(x_i,\cdot), f\rangle_{\mathcal{H}} = f(x_i) = 0$.
- Thus, $k(x_i,\cdot) \in (\tilde{\mathcal{H}}^{\perp})^{\perp}$. Combining these results, we see

$$span\{k(x_i,\cdot)\}_{i=1}^n \subseteq span\{\tilde{k}(x_i,\cdot)\}_{i=1}^n$$

- Also possible to show that $\tilde{k}(x_i,\cdot) \in (\tilde{\mathcal{H}}^{\perp})^{\perp}$, so

$$span\{\tilde{k}(x_i,\cdot)\}_{i=1}^n \subseteq span\{k(x_i,\cdot)\}_{i=1}^n$$

- Therefore, the two spans are the same.

Deriving the Warped Kernel

- If $span\{\tilde{k}(x_i, \cdot)\}_{i=1}^n$ is the same as $span\{k(x_i, \cdot)\}_{i=1}^n$, we can use the result that $k(x, \cdot) - \tilde{k}(x, \cdot) \in span\{\tilde{k}(x_1, \cdot), \dots, \tilde{k}(x_n, \cdot)\}$ to write

$$\tilde{k}(x, \cdot) = k(x, \cdot) + \sum_j \beta_j(x) k(x_j, \cdot)$$

  where the β_j coefficients depend on the data x.
- Warped kernel $\tilde{k}$ is simply k modified by some linear combination of data points.
- If we can find an explicit expression for $\beta_j(x)$, then we'll have an explicit form for $\tilde{k}$!

Deriving the Warped Kernel

- We now find $\beta_j(x)$.
- System of linear equations formed by evaluating $k(x_i, x)$ or $k_{x_i}(x)$:

$$
\begin{aligned}
k_{x_i}(x) &= \langle k(x_i,\cdot), \tilde{k}(x,\cdot)\rangle_{\tilde{\mathcal{H}}} \quad \text{(repro. prop. of } \tilde{\mathcal{H}}\text{)} \\
&= \langle k(x_i,\cdot), k(x,\cdot) + \sum_j \beta_j(x) k(x_j,\cdot)\rangle_{\tilde{\mathcal{H}}} \\
&= \langle k(x_i,\cdot), k(x,\cdot) + \sum_j \beta_j(x) k(x_j,\cdot)\rangle_{\mathcal{H}} + \mathbf{k_{x_i}}^\top M\mathbf{g}
\end{aligned}
$$

where $\mathbf{k_{x_i}} = (k(x_i, x_1), \ldots, k(x_i, x_n))^\top$ and $\mathbf{g}$ has n components $g_m = k(x, x_m) + \sum_j \beta_j(x) k(x_j, x_m), m = 1, \ldots, n$.

Deriving the Warped Kernel

- Final linear system for coefficients $\beta(x) = (\beta_1(x), \ldots, \beta_n(x))^\top$:

$$(I + MK)\beta(x) = -M\mathbf{k_x}$$

  where K is the kernel matrix on all n data points, and $\mathbf{k_x} = (k(x_1, x), \ldots, k(x_n, x))^\top$.
- Now solving for $\beta(x)$ gives us an expression for $\tilde{k}$.

Reproducing Kernel of $\tilde{\mathcal{H}}$

$$\tilde{k}(x, z) = k(x, z) - \mathbf{k_x}^\top (I + MK)^{-1} M\mathbf{k_z}$$

Choosing Deformation Matrix M

- Now that we can express $\tilde{k}$ explicitly, we need to choose M.
  - Want M to encode our intuition about the data geometry.
- Choose *graph Laplacian* associated with the point cloud.
  - Implements smoothness assumption w.r.t. graph over data.
- Let W be the graph edge weight matrix with $W_{ij} = e^{-\frac{||x_i - x_j||^2}{2\sigma^2}}$, if x_i and x_j are nearest neighbors, and 0 otherwise.
- Let D be the diagonal degree matrix with $D_{ii} = \sum_j W_{ij}$.
- Laplacian $L = D - W$.
- Note: $\mathbf{f}^\top L \mathbf{f} = \sum_{i,j=1}^{n} W_{ij}(f(x_i) - f(x_j))^2$

Using Laplacian as Deformation Matrix

- Thus, using $M = \frac{\gamma_I}{\gamma_A} L$, the problem in modified RKHS $\tilde{\mathcal{H}}$:

$$f = \arg\min_{h \in \tilde{\mathcal{H}}} \frac{1}{l} \sum_{i=1}^{l} V(h, x_i, y_i) + \gamma_A ||h||^2_{\tilde{\mathcal{H}}} \quad (1)$$

is equivalent to the MR problem in original RKHS $\mathcal{H}$:

$$f = \arg\min_{h \in \mathcal{H}} \frac{1}{l} \sum_{i=1}^{l} V(h, x_i, y_i) + \gamma_A ||h||^2_{\mathcal{H}} + \gamma_I \sum_{i,j=1}^{n} W_{ij} (h(x_i) - h(x_j))^2 \quad (2)$$

- Thus, solving (1) using $\tilde{k}(x, z)$ in a standard kernel method achieves MR result from (2).

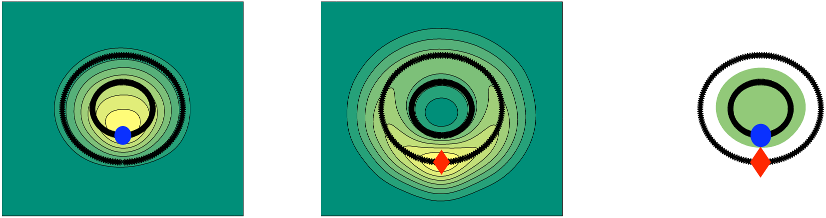
Revisiting the Two Circles Example

- Using the Laplacian warped kernel, the result is a combination of kernels that adhere to the geometry of the space.
- Now the decision boundary correctly separates the circles.

(a) deformed kernel centered on labeled point 1

(b) deformed kernel centered on labeled point 2

(c) classifier learnt in the deformed RKHS



Outline

Summary of Methods

- Use $M = \frac{\gamma_I}{\gamma_A} L^p$ for some integer p
- Methods: Laplacian SVM, Laplacian RLS
- Using the warped kernel, solve both using standard solvers.
- Some parameters fixed to reduce complexity, others chosen by grid search using 5-fold CV.
- Compared against standard SVM and RLS without data-dependent kernel, and to other transductive methods.

Summary of Datasets

- Range of tasks:
  - artificial two 50-dim Gaussian data
  - image and digit recognition data
  - text classification data (i.e., classify newsgroup posts by topic)
  - Web page classification using page text and/or hyperlink text
- Properties:
  - 2–20 classes
  - 50–7000 dimensions
  - 12–50 labeled points
  - 500–2000 unlabeled points

Summary of Results

- Transductive (in-sample unlabeled data) results
  - Significant gains for LapSVM and LapRLS
- Semi-supervised (out-of-sample generalization) results
  - As good as transductive results
- Study of parameters
  - Larger γ_I leads to much better in-sample performance
  - Need to increase γ_A to maintain out-of-sample performance

Conclusions

- Showed how to derive a “warped kernel” that adapts to the underlying data geometry.
- Allows semi-supervised learning beyond transduction.
- Permits simple training using standard supervised methods.
- General framework: changing deformation matrix M allows other forms of unlabeled-data-based regularization.

Demos:

- http://people.cs.uchicago.edu/~mrainey/jlapvis/JLapVis.html
- http://people.cs.uchicago.edu/~vikass/manifoldregularization.html