Scalable Algorithms for Graph-Based Semi-Supervised Learning

Introduction and Challenges

Guowei Sun[1]

[1] Department of Mathematics
University of Maryland, College Park

Feb 22nd / Group Meeting

Outline

Outline

Motivation

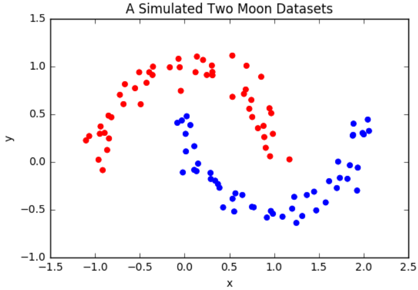


Figure: The True Data and Its Classification

- The datasets contains 100 data points
- Each data point has two feature, (x, y)
- The true classification is smooth w.r.t pair-wise distances between data points

Learning Problem on The Two Moons

Given the labeled data points, want to predict labeling for all data points?

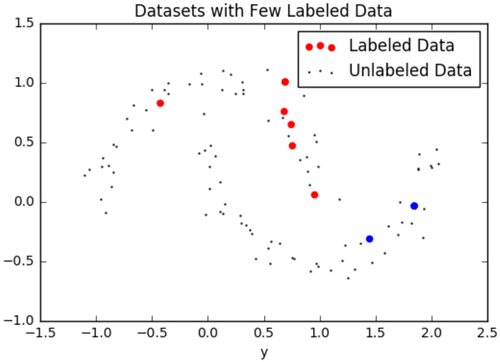


Figure: Small Number of Labeled Data Points

Supervised Learning?

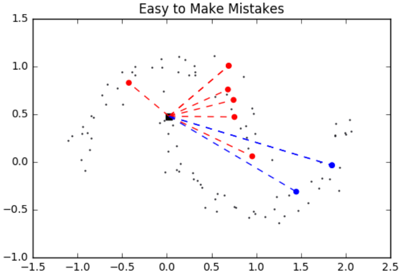


Figure: Prediction for the Selected Data Point?

- Only consider the labeled data points and the target data point
- Will almost surely make a mistake using any supervised learning method. SVM, Logistic Regression, Deep Learning etc.

Semi-supervised Learning

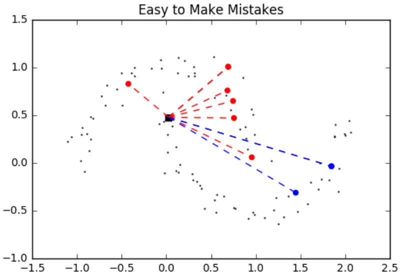


Figure: Use Unlabeled Data to Help Prediction

- Also consider the unlabeled data
- Find the "Two Moon" structure
- Make smooth predictions

Outline

Problem Formulation
General Setup

- Given $X_L = \{x_1, x_2, ..., x_l\}$, with given labels $\{y_1, y_2, ..., y_l\}$
- Given $X_U = \{x_{l+1}, x_{l+2}, ..., x_{l+u}\}$, with no given labels
- $u \gg l$ bacause labeling is expensive, and collecting data is cheap.
  e.g. image tagging is expensive compared to collecting images
- Goal: predict the labeling for X_u

Problem Formulation
Adjacency Graph Design

To represent the relationship among data points, construct adjacency matrix W:

- $w_{i,j}$ measures "similarity" among x_i and x_j
- Common Choices:
  - Gaussian Kernel: $w_{i,j} = e^{-\frac{||x_i - x_j||^2}{\sigma}}$
  - ϵ Graph: $w_{i,j} = I_{||x_i - x_j|| < \epsilon}$
  - KNN Graph: $w_{i,j} = 1$ if x_i is among the k-nearest neighbour of x_j or vice versa.

Remark: ϵ graph and KNN graph will lead to sparse matrix, therefore easier numerical computations.

Problem Formulation
Assumptions

Zhou et al, NIPS2003, 2625
Zhu et al, ICML2003, 2564
Belkin et al, JMLR2006, 2264

- Local Consistency : "similar points" should have similar labels
- Global Consistency: Points on the same strucutre (cluster or manifold) should have similar labels

Problem Formulation
Objective Function Design

Let $f : x \rightarrow \mathcal{R}$ be a prediction function we want to estimate

- zhu2003

$$min\ J(f) = \sum_{i,j} w_{i,j}(f(x_i) - f(x_J))^2$$

$$\text{subject to } f(x_p) = y_p, p = 1, .., l$$

- zhou2003

$$min\ J(f) = \sum_{i,j} w_{i,j}(f(x_i) - f(x_j))^2 + \lambda \sum_{p=1}^{l} (f(x_p) - y_p)^2$$

We only focus on local consistency. "Manifold Learning" is ignored here.

Problem Formulation

Compact Representation and Possible Modifications

Define graph Laplacian $L = D - W$, D is a diagonal matrix with $D_{ii} = \sum_j w_{i,j}$, then

$$\sum_{i,j} w_{i,j}(f(x_i) - f(x_j))^2 = f^T L f$$

Hard Constraint:

$$min\, J(f) = f^T L f$$

$$\text{s.t. } f(x_p) = y_p$$

Regularizor:

$$min\, J(f) = f^T L f + (f - y)^T I_L (f - y)$$

$$\text{where } I_L(ii) = 1_{\{i \leq l\}}$$

Problem Formulation

More Recent Standards

More recently,

- More general norm, could use l_p norm.
- Use normalied graph-laplacian: $L = I - D^{-1/2}WD^{-1/2}$

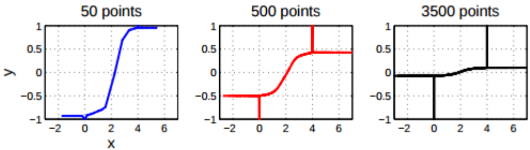
Then, the smoothness term in the objective function becomes

$$\sum_{i,j} w_{i,j} || \frac{f(x_i)}{\sqrt{D_i}} - \frac{f(x_j)}{\sqrt{D_j}} ||_{l_p}$$

Outline

Consistency and Computation
Consistency

- On a 1d problem, with data generated from $N(0,1)$ and $N(4,1)$



As $u \to \infty$, the l_2 norm forces f to be mostly constant. Therefore "informationless".

Nadler et al *The limit of Infinite Unlabeled Data*, NIPS2009

Consistency and Computation

Consistency Solved

Choose appropriate norm could solve this issue. Let d be the dimension of the problem(?). Using p norm:

- $p \leq d$ degenerate limit
- $p = \infty$ insensitive to input distribution
- $p = d + 1$ no degeneracy, and sensitive to input distribution

Alaoui et al, JMLR 2016

Outline

1. Semi-supervised Learning
   - Example on Two Moons
   - Problem Formulation
2. Challenges
   - Consistency
   - Computation
   - Two Case Studies
   - High-Level Challenges
3. Parallel Computation

- For l_2 norm regularizors, it is a standard linear optimization problem. Can be solved in $O(n^3)$ times.
- For l_p norms, algorithms are proposed (Kyng et al, *Lipshitz Learning on Graphs*,JMLR 2015). But also scales badly with data points.
- Constructing the similarity matrix W would incur $O(n^2)$ cost.
- Semi-supervised Learning is supposed to benefit from more unlabeled data??

Outline

Anchor Point Approach
Liu et al ICML 2010

- l is not too large
- u is computational infeasible for a full SSL solution
- choose $m \ll u$ unlabeled data points to serve as "anchor points"

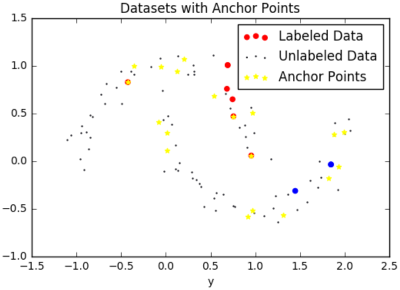


Figure: 10 Labeled Points, 20 Anchor Points, 90 Unlabeled Points

Anchor Point Approach

Non-Parametric Regression based on Anchor Points

- Given anchor points $A = \{a_1, a_2, .., a_m\}$
- Assume for $x_i \in X_U \cup X_L$, $f(x_i) = \sum_{k=1}^{m} z_{ik} f(a_k)$. The prediction function is weighted average of prediction function values on the anchor points
- Then, $f = Zf_A, f \in \mathcal{R}^{(l+u)\times 1}, Z \in \mathcal{R}^{(l+u)\times m}, f_A \in \mathcal{R}^{m\times 1}$
- They used gaussian kernel for z_{ik}.

Anchor Point Approach

Low-rank Approximation of W

- Given Z, $x_i \to z_i$ is a new representation of data points
- Therefore, treat A like a new basis for $X_U \cup X_L$. Then $w_{x_i,x_j} \approx w_{z_i,z_j}$
- They give $W = Z\Lambda^{-1}Z^T$ where $\Lambda_{kk} = \sum_{i=1}^{l+u} z_{ik}$

Anchor Point Approach
Solve The Problem

- Given Z, W, they solve the problem by finding f_A to minimize

$$J(f_A) = ||Z_l f_A - y_l||^2 + \gamma f_A^T Z^T L Z f_A$$

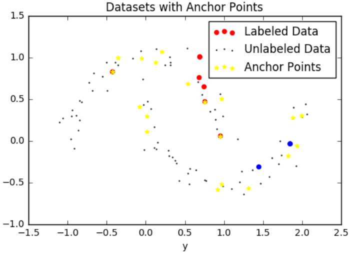
- Write $\tilde{L} = Z^T L Z = Z^T Z - (Z^T Z)\Lambda^{-1}(Z^T Z)$
- Then

$$f_A^* = (Z_l^T Z_l + \gamma \tilde{L})^{-1} Z_l^T y_l$$

Anchor Point Approach
Complexity Summary

Stage	AnchorGraph
find anchors	$O(mn)$
design Z	$O(lmn)$
graph regularization	$O(m^3 + m^2 n)$
Total Time Complexity	$O(m^2 n)$

Anchor Point Approach
Anchor Point Selection – Unsolved



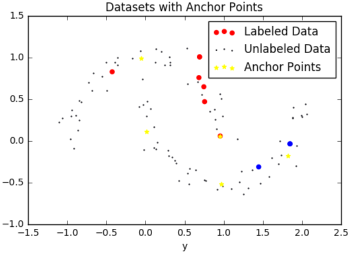


Figure: Effect of 5 and 20 Anchor Points

- If the structure of the data is complex, need larger *m*
- No guidance on choosing proper *m*

Application on Image Data Sets

Eigenfunction Approximation Approach

- 80 million images
- Used GIST descriptor to generate feature vectors to represent images
- Used PCA to find low-rank approximation of data points
- Approximate distribution of each feature using histograms generated from the entire data set

For details, see Fergus et al, *SSL in Gigantic Image Collections*, NIPS2009

Outline

High-Level Challenges

- Why isn't everyone Using SSL?
  Often labeled data is itself large enough to represent the strucutre within the data. e.g. Go games, Speech Recognition, Object Recognition etc
- Flexibility of SSL?
  Graph-based SSL, Manifold SSL etc are all very specific. There is no easy way to incorporate data into any supervised learning method. e.g. Combine SSL with Deep Learning?

Assumptions and Goals

Assumptions:

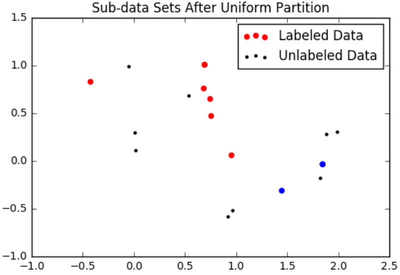
- X_L is too small to fully represent the structure within the data
- X_U is too big to be processed sequentially
- We have K nodes available

Goals:

- Minimize communication overhead for parallel processes
- Tradeoff between accuracy and speed

Equal Partition of Unlabeled Data

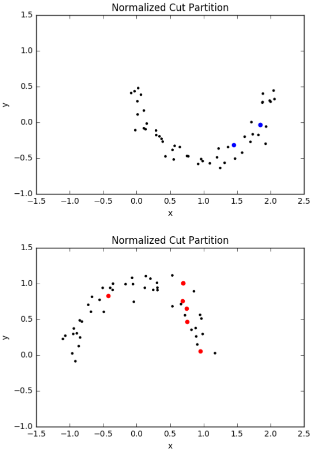
- Distribute X_L to all Nodes
- Partition X_U uniformly into K subsets $X_{u1}, .., X_{uK}$
- Treat $X_L + X_{uk}$ as a new SSL problem
- No Communication: Possible errors
- Enable Communication: lots of coding



$X_L + X_{uk}$ might be too small to represent the structure of data

Partition Through Normalized Graph Cut

- Label Propagation is Local
- Minimal Communication Overhead Between Nodes
- Cutting the Graph can be Expensive



To be Continued