Nonparametric Maximum Margin Similarity for Semi-Supervised Learning

Yingzhen Yang, Xinqi Chu, Zhangyang Wang, Thomas S. Huang
Beckman Institute for Advanced Science and Technology,
University of Illinois at Urbana-Champaign
Urbana, IL 61801
{yyang58,chu36,zwang119,t-huang1}@illinois.edu

Abstract

Nonparametric Label Propagation (LP) has been proven to be effective for semi-supervised learning problems, and it predicts the labels for unlabeled data by a harmonic solution of an energy minimization problem which encourages local smoothness of the labels in accordance with the similarity graph. We relate LP to a novel nonparametric maximum margin similarity framework, where LP becomes a special case when the separation parameter is sufficiently large. We provide theoretical result that the hinge loss of a linear classifier at a specific margin in the transformed space is bounded by the hinge loss of the similarity margins in the original space with a high probability. A new Maximum Margin Similarity Graph (MMSG) is presented under this framework for semi-supervised learning, which minimizes the hinge loss of the similarity margins. The experimental results demonstrate its superiority compared to LP.

1 Introduction

Nonparametric Label Propagation (LP) algorithms [1, 2] are widely used semi-supervised learning algorithms. With a predefined nonparametric similarity graph, LP determines the labels of unlabeled data by the minimization of the objective function which encourages local smoothness of the labels according to the graph weight of the similarity graph. The typical LP algorithm [1] renders a harmonic solution which can also be interpreted by random walks from unlabeled data to the labeled data.

On the other hand, the success of LP algorithms highly depends on the underlying similarity graph. Most similarity graphs for LP are constructed empirically and the objective function over the similarity graphs is defined as sum of the product of pairwise similarity and the squared label difference. In this paper, we relate LP to a novel nonparametric maximum margin similarity framework with the concept of similarity margin, and present a new semi-supervised learning algorithm called Maximum Margin Similarity Graph (MMSG). The conventional LP algorithm [1] can be interpreted as a special case of our MMSG algorithm when the separation parameter is sufficiently large. In our framework, a similarity margin is defined for each datum x as the difference of its similarity to the data with the same label, and its similarity to the data with different label. The objective function of MMSG is the sum of hinge loss in terms of the similarity margins. We present theoretical result on the hinge loss of the similarity margins, which shows that the hinge loss of a linear classifier at a specific margin in the transformed space is bounded by the hinge loss of the similarity margins in the original space with a high probability. The original space is mapped to the transformed space by the similarity function. The experimental results evidence the effectiveness of MMSG over LP on real data set.

Our maximum margin similarity framework is inspired by the similarity learning method [3], which suggests a new viewpoint of kernel functions as a similarity measures between the data. Their method also naturally handles general non-kernel similarity functions, e.g. those do not have implicit feature mappings and the property of positive semi-definiteness. By the sample-based similarity margin rather than the expectation based margin in [3], our framework leads to an tractable optimization problem which is solved by the projected subgradient method. The minimization of the objective function is justified by our new learning theorem for maximum margin similarity. Similar to [3], our framework applies to any bounded similarity functions.

2 Maximum Margin Similarity

We introduce the maximum margin similarity framework, and the theorem under this framework which shows the existence of a good classifier that has bounded hinge loss in the transformed space with a high probability.

2.1 Notations

The similarity function over $\mathbb{R}^d \times \mathbb{R}^d$ is defined as any bounded pairwise function $S\colon \mathbb{R}^d \times \mathbb{R}^d \to [-1,1]$. The labeled data $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$ are drawn i.i.d. from some distribution P on $\mathbb{R}^d \times \{-1,1\}$. We consider binary classification in this paper, and the multi-class case can be handled in the one-vs-all manner. The similarity margin of the datum $\mathbf{x} \in \mathbb{R}^d$ is defined as the difference of sum of $\mathbf{x}$'s similarity to the data with the same label as $\mathbf{x}$, and the sum of $\mathbf{x}$'s similarity to the data with different label:

$$\gamma_{\mathbf{x}} = \frac{1}{n}\Big(\sum_{j:y_j=y(\mathbf{x})} S(\mathbf{x},\mathbf{x}_j) - \sum_{j:y_j\neq y(\mathbf{x})} S(\mathbf{x},\mathbf{x}_j)\Big) \tag{1}$$

where $y(\mathbf{x})$ is the label of $\mathbf{x}$.

2.2 Theoretical Guarantee

Intuitively, the similarity margin for each datum should be large so as to separate different classes. To facilitate optimization algorithms, the small similarity margin is penalized by hinge loss. For the separation parameter $\gamma > 0$, the hinge loss of the similarity margins for the data $\mathcal{D}$ is defined as

$$H_{\gamma,\mathcal{D}} = \frac{1}{n}\sum_{i=1}^{n} \max\{0, 1 - \frac{\gamma_i}{\gamma}\} \tag{2}$$

where $\gamma_i = \gamma_{\mathbf{x}_i}$ is the similarity margin of $\mathbf{x}_i$. Theorem 1 shows that with a high probability, there exists a linear classifier in the transformed space with hinge loss bounded by $H_{\gamma,\mathcal{D}}$.

Theorem 1. *Let S be the similarity function defined in Section 2.1. Given the data $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$, define the mapping $F_{\mathcal{D}}\colon \mathbb{R}^d \to \mathbb{R}^n$ as $F_{\mathcal{D}}(\mathbf{x}) = \frac{1}{\sqrt{n}}\big(S(\mathbf{x},\mathbf{x}_1), S(\mathbf{x},\mathbf{x}_2), \dots, S(\mathbf{x},\mathbf{x}_n)\big)$. For $\delta_1, \delta_2, \delta_3 > 0$, with probability at least $1 - \delta_1 - \delta_2 - \delta_3$ over the data $\mathcal{D}$, there exists a linear classifier in the transformed space induced by $F_{\mathcal{D}}$ such that this classifier has hinge loss at most $H_0 = H_{\gamma,\mathcal{D}} + \frac{\sqrt{\frac{2}{n}\log\frac{n}{\delta_1}}}{\gamma} + \sqrt{\frac{2}{n\gamma^2}\log\frac{1}{\delta_2}} + \delta_3(1+\frac{1}{\gamma})$ with respect to the margin γ. Namely, there exists a vector $\boldsymbol{\beta} \in \mathbb{R}^n$ such that $\mathbb{E}_{(\mathbf{x},y)\sim P}\Big[\max\{0, 1 - \frac{y\langle\boldsymbol{\beta}, F_{\mathcal{D}}(\mathbf{x})\rangle}{\gamma}\}\Big] \le H_0$.*

The proof is in the Appendix. Theorem 1 justifies the minimization of the hinge loss of the similarity margins (2) for semi-supervised learning, which is to minimize the upper bound for the hinge loss of a classifier at margin γ in the transformed space.

2.3 Maximum Margin Similarity Graph for Semi-Supervised Learning

We now introduce Maximum Margin Similarity Graph for semi-supervised learning, and MMSG minimizes the hinge loss of the similarity margins (2) by projected subgradient method. The bounded similarity function $S(u,v) = \exp\big(-\frac{1}{0.03}(1 - \frac{u^T v}{\|u\|_2\|v\|_2})\big)$, which is suggested in [1], is used throughout our experiments. The data $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_l, \mathbf{x}_{l+1}, \ldots, \mathbf{x}_n\} \subseteq \mathbb{R}^d$ are comprised of labeled and unlabeled set, and the first l points have labels $y_i \in \{-1, 1\}$ for $1 \le i \le l$. In the following text, ℓ_i is the label of $\mathbf{x}_i$ for $i \in \{1, \ldots, n\}$, and $\ell_i = y_i$ for $i \in \{1, \ldots, l\}$. The optimization problem of MMSG is presented below, where the discreteness condition is relaxed so that $\boldsymbol{\ell}$ takes real values for unlabeled data:

$$\min_{\boldsymbol{\ell}} \quad H_{\boldsymbol{\ell}} = \frac{1}{n}\sum_{i=1}^{n} \max\{0, 1 - \frac{\gamma_i}{\gamma}\} \tag{3}$$

$$s.t. \quad \ell_i \in [-1, 1],\ i \in \{1+1, \ldots, n\} \tag{4}$$

$$\ell_i = y_i,\ i \in \{1, \ldots, l\} \tag{5}$$

where γ_i is rewritten as $\gamma_i = \frac{1}{n}\sum\limits_{j=1}^{n} S(\mathbf{x}_i, \mathbf{x}_j)\big(1 - \frac{1}{2}(\ell_i - \ell_j)^2\big)$ and it is convex function of $\boldsymbol{\ell}$. Note that when the separation parameter is sufficiently large ($\gamma \ge \max_i\{\gamma_i\}$), $H_{\boldsymbol{\ell}} = 1 - \frac{\sum_{i=1}^{n} \gamma_i}{n\gamma}$, and the optimization problem (3) is reduced to that of Label Propagation [1]. In the iteration $k \ge 0$ of the projected subgradient method, we set $\boldsymbol{\ell}^{(k+1)} = P_C(\boldsymbol{\ell}^{(k)} - \eta_k \mathbf{g}^{(k)})$ where η_k is the learning rate, $\mathbf{g}^{(k)} \in \mathbb{R}^n$ is the subgradient of $H_{\boldsymbol{\ell}}$ at $\boldsymbol{\ell}^{(k)}$:

$$\mathbf{g}_t^{(k)} = \frac{1}{n^2\gamma}\Big(\big(\sum_{j=1}^{n} S(\mathbf{x}_t, \mathbf{x}_j)(\ell_t^{(k)} - \ell_j^{(k)})\big)\mathbb{1}_{\gamma_t<\gamma} + \sum_{i \ne t} S(\mathbf{x}_i, \mathbf{x}_t)(\ell_t^{(k)} - \ell_i^{(k)})\mathbb{1}_{\gamma_i<\gamma}\Big),\ 1 \le t \le n$$

P_C is a projection operator and C indicates the feasible set specified by the constraint (4) and (5).

3 Experimental Results

In this section we conduct experiments on the UCI Ionosphere data set, and the accuracy of LP and MMSG with respect to different labeled set size are shown in Figure 1. For each labeled set size, we perform 5 trials randomly. We set the the separation parameter γ as the average of $\{\gamma_i\}$ using the labels produced by Label Propagation [1]. It is observed that the accuracy curves of MMSG and LP share the same tendency with respect to the size of labeled set, and MMSG always achieves better accuracy by minimizing the hinge loss of the similarity margins

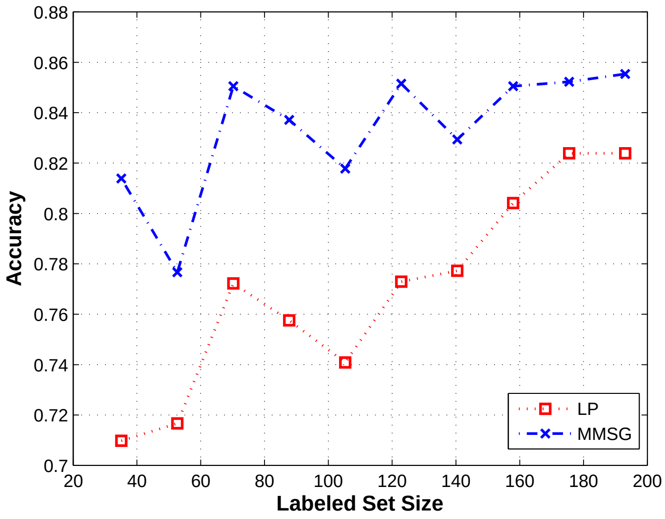

Figure 1: Semi-supervised learning results on the UCI Ionosphere data set.

4 Appendix

Proof of Theorem 1. First, we show that with a high probability, $\mathbb{E}_{(\mathbf{x},y)\sim P}\Big[\max\{0, 1 - \frac{y\mathbb{E}_{\mathbf{x}',y'\sim P}\big[S(\mathbf{x},\mathbf{x}')y'\big]}{\gamma}\}\Big]$ is upper bounded. By the McDiarmids inequality, for $\delta_1, \delta_2 > 0$,

$$\Pr\Big[\frac{1}{n}\sum_{i=1}^{n}\max\{0, 1 - \frac{E_i}{\gamma}\} - H_{\gamma,\mathcal{D}} \geq \frac{\sqrt{\frac{2}{n}\log\frac{n}{\delta_1}}}{\gamma}\Big] \leq \delta_1 \tag{6}$$

$$\Pr\Bigg[\mathbb{E}_{(\mathbf{x},y)\sim P}\Big[\max\{0, 1 - \frac{y\mathbb{E}_{\mathbf{x}',y'\sim P}\big[S(\mathbf{x},\mathbf{x}')y'\big]}{\gamma}\}\Big] - \frac{1}{n}\sum_{i=1}^{n}\max\{0, 1 - \frac{E_i}{\gamma}\} \geq \sqrt{\frac{2}{n\gamma^2}\log\frac{1}{\delta_2}}\Bigg] \leq \delta_2 \tag{7}$$

where $E_i = y_i\mathbb{E}_{(\mathbf{x},y)\sim P}\big[S(\mathbf{x}_i,\mathbf{x})y\big]$. Therefore, with probability at least $1 - \delta_1 - \delta_2$ over the data $\mathcal{D}$,

$$\mathbb{E}_{(\mathbf{x},y)\sim P}\Big[\max\{0, 1 - \frac{y\mathbb{E}_{\mathbf{x}',y'\sim P}\big[S(\mathbf{x},\mathbf{x}')y'\big]}{\gamma}\}\Big] \leq H_{\gamma,\mathcal{D}} + \sqrt{\frac{2}{n\gamma^2}\log\frac{1}{\delta_2}} + \frac{\sqrt{\frac{2}{n}\log\frac{n}{\delta_1}}}{\gamma} \tag{8}$$

Let $\boldsymbol{\beta} = (\beta_1, \dots, \beta_n) \in \mathbb{R}^n$ with $\beta_i = \frac{y_i}{\sqrt{n}}$. Again, by the McDiarmids inequality we bound $y\langle\boldsymbol{\beta}, F_{\mathcal{D}}(\mathbf{x})\rangle$ by $y\mathbb{E}_{(\mathbf{x}',y')\sim P}\big[S(\mathbf{x},\mathbf{x}')y'\big]$:

$$\Pr\Big[y\langle\boldsymbol{\beta}, F_{\mathcal{D}}(\mathbf{x})\rangle \leq y\mathbb{E}_{(\mathbf{x}',y')\sim P}\big[S(\mathbf{x},\mathbf{x}')y'\big] - \sqrt{\frac{2}{n}\log\frac{1}{\delta_3^2}}\Big] \leq \delta_3^2 \tag{9}$$

And (9) holds for any $(\mathbf{x}, y)$. Let $\mathcal{A}$ denote the event in (9), it follows from Fubini's Theorem that $\mathbb{E}_{\mathcal{D}\sim P^n}\Big[\mathbb{E}_{(\mathbf{x},y)\sim P}\big[\mathbb{1}_{\mathcal{A}}\big]\Big] \leq \delta_3^2$. According to the Markov's inequality, $\Pr_{\mathcal{D}\sim P^n}\Big[\Pr_{(\mathbf{x},y)\sim P}\big[\mathcal{A}\big] \geq \delta_3\Big] \leq \delta_3$ which means that with probability at least $1-\delta_3$ over $\mathcal{D}$, the probability measure of the set $(\mathbf{x}, y)$ that makes $\mathcal{A}$ happen is at most δ. Note that for $(\mathbf{x}, y)$ that $\mathcal{A}$ does not hold,

$$\max\{0, 1 - \frac{y\langle\boldsymbol{\beta}, F_{\mathcal{D}}(\mathbf{x})\rangle}{\gamma}\} \leq \max\{0, 1 - \frac{y\mathbb{E}_{(\mathbf{x}',y')\sim P}\big[S(\mathbf{x},\mathbf{x}')y'\big]}{\gamma}\} + \sqrt{\frac{2}{n}\log\frac{1}{\delta_3^2}} \tag{10}$$

And for $(\mathbf{x}, y)$ that $\mathcal{A}$ holds, $\max\{0, 1 - \frac{y\langle\boldsymbol{\beta}, F_{\mathcal{D}}(\mathbf{x})\rangle}{\gamma}\} \leq 1 + \frac{\|\boldsymbol{\beta}\|\|F_{\mathcal{D}}(\mathbf{x})\|}{\gamma} = 1 + \frac{1}{\gamma}$.

According this result, (8) and (9), with probability at least $1 - \delta_1 - \delta_2 - \delta_3$ over $\mathcal{D}$,

$$\begin{aligned}
&\mathbb{E}_{(\mathbf{x},y)\sim P}\Big[\max\{0, 1 - \frac{y\langle\boldsymbol{\beta}, F_{\mathcal{D}}(\mathbf{x})\rangle}{\gamma}\}\Big] \\
&= \mathbb{E}_{(\mathbf{x},y)\sim P,\mathcal{A}^c}\Big[\max\{0, 1 - \frac{y\langle\boldsymbol{\beta}, F_{\mathcal{D}}(\mathbf{x})\rangle}{\gamma}\}\Big] + \mathbb{E}_{(\mathbf{x},y)\sim P,\mathcal{A}}\Big[\max\{0, 1 - \frac{y\langle\boldsymbol{\beta}, F_{\mathcal{D}}(\mathbf{x})\rangle}{\gamma}\}\Big] \\
&\leq \mathbb{E}_{(\mathbf{x},y)\sim P}\Big[\max\{0, 1 - \frac{y\mathbb{E}_{(\mathbf{x}',y')\sim P}\big[S(\mathbf{x},\mathbf{x}')y'\big]}{\gamma}\}\Big] + \delta_3(1 + \frac{1}{\gamma}) \\
&\leq H_{\gamma,\mathcal{D}} + \sqrt{\frac{2}{n\gamma^2}\log\frac{1}{\delta_2}} + \frac{\sqrt{\frac{2}{n}\log\frac{n}{\delta_1}}}{\gamma} + \delta_3(1 + \frac{1}{\gamma})
\end{aligned} \tag{11}$$

□

Acknowledgements. This material is based upon work supported by the National Science Foundation under Grant No. 1318971. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] Xiaojin Zhu, Zoubin Ghahramani, and John D. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, pages 912–919, 2003.

[2] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16 [Neural Information Processing Systems, NIPS 2003, December 8-13, 2003, Vancouver and Whistler, British Columbia, Canada]*, 2003.

[3] Maria-Florina Balcan and Avrim Blum. On a theory of learning with similarity functions. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 73–80, New York, NY, USA, 2006. ACM.