Unlabeled Data: Now It Helps, Now It Doesn't

A. Singh, R. D. Nowak, and X. Zhu. In NIPS, 2008. 1

Mark Andrew Ward and Max Kuang

Courant Institute, NYU

April 21, 2015

Outline

Outline

Conflicting Views in Semi-supervised Learning

- Given n iid labeled sample $\{(x_i, y_i)\}_{i=1,\ldots,n}$ and m iid unlabeled sample $\{x_i'\}_{i=1,\ldots,m}$, can we do better than supervised learning from merely n labeled points $\{(x_i, y_i)\}_{i=1,\ldots,n}$.
- Not always better: Only when there exists a **link** between the **marginal data distribution** $P(x)$ and the **target function to be learned** $y = f(x)$.
- **Links**: **cluster assumption** and **manifold assumption**.
- Does unlabeled data help in error convergence rate under different assumptions?

	SSL helps	SSL does not help
Cluster assumption	Castelli and Cover[1, 2]	Rigollet[5]
Manifold assumption	Lafferey and Wasserman[3]	Niyogi[4]

Conflicting Views in Semi-supervised Learning

- This work focuses on learning under the **cluster assumption** and provides **finite sample bounds** to identify situations in which unlabeled data will help to improve learning.

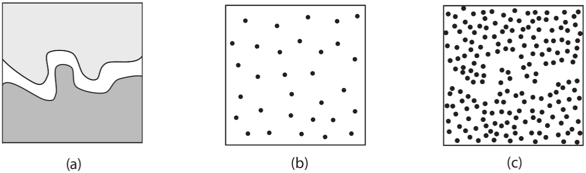


Figure 1: (a) Two separated high density sets with different labels that (b) cannot be discerned if the sample size is too small, but (c) can be estimated if sample density is high enough.

The Cluster Assumption: Marginal Distributions

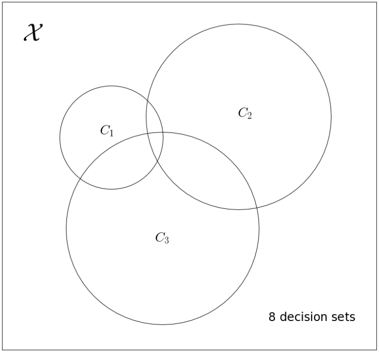
- The marginal distribution $p(x) = \sum_{k=1}^{K} a_k p_k(x)$ is the mixture of a finite, but unknown, number of component densities $\{p_k\}_{k=1}^{K}$.
- Restrictions on p_k:
  1. p_k is supported on a compact connected set $C_k \in \mathcal{X}$ with Lipschitz boundaries. Specifically:

$$C_k = \{ \quad x \equiv (x_1, x_2, \ldots, x_d) \in \mathcal{X} : \\ g_k^{(1)}(x_1, x_2, \ldots, x_{d-1}) \leq x_d \leq g_k^{(2)}(x_1, x_2, \ldots, x_{d-1})\} \quad (1)$$

  2. p_k is bounded from above and below, $0 < b \leq p_k \leq B$.
  3. p_k is Holder-α smooth on C_k with Holder constant K_1.

The Cluster Assumption: Dicision Sets

- Let $\mathcal{D}$ denote the collection of all non-empty sets obtained as intersections of $\{C_k\}_{k=1}^{K}$.
- **Cluster assumption: the target function $y = f(x)$ to be learnt is smooth on each set $D \in \mathcal{D}$.**



The Cluster Assumption: Margin γ

- The margin γ of a distribution is defined to be the minimal width of a decision set.
- The margin γ is assigned a positive sign if there is no overlap between components,otherwise it is assigned a negative sign.

$$
\begin{aligned}
d_{jk} &:= \min_{p,q\in\{1,2\}} \|g_j^{(p)} - g_k^{(q)}\|_\infty & j \neq k,\\
d_{kk} &:= \|g_k^{(1)} - g_k^{(2)}\|_\infty,
\end{aligned}
$$

where $\|\cdot\|_\infty$ denotes the sup-norm, and

$$
\sigma = \begin{cases} 1 & \text{if } C_j \cap C_k = \emptyset\ \forall j \neq k, \text{ where } j,k \in \{1,\ldots,K\} \\ -1 & \text{otherwise} \end{cases}
$$

Then the margin is defined as

$$
\gamma = \sigma \cdot \min_{j,k\in\{1,\ldots,K\}} d_{jk}.
$$

Outline

Summary

- Under cluster assumption, we are trying to figure out for what (n, m, γ)(and possibly other constraints) SSL surpasses SL for all general learners.

$$\text{SSL} > \text{SL}$$
$$\Uparrow$$
$$\text{SSL Learner} \approx \text{Clairvoyant Learner} > \text{General SL Learner}$$
$$\Uparrow(\text{definition})$$

Supervised learners with perfect knowledge of the decision sets $\mathcal{D}$

Learning the Decision Sets: SSL Learner $\approx$ Clairvoyant Learner

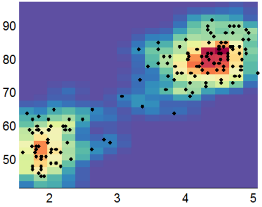
- Decision sets are learnable using unlabeled data: marginal density p is smooth within each decision set but exhibits jumps at the decision set boundaries.
- Main learning procedure:
  1. Marginal Density Estimation: From unlabeled sample $\{x_i\}_{i=1,\ldots,m}$ to density estimator $\hat{p}(x)$
  2. Decision Set Estimation: From $\hat{p}(x)$ to decision set estimator $\hat{\mathcal{D}}$

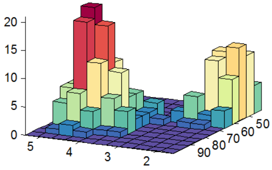
Learning the Decision Sets: Marginal Density Estimation

- Sup-norm kernel density estimator[6]: Consider a uniform grid over the feature space $\mathcal{X} = [0,1]^d$ with spacing $2h_m$, where $h_m = \kappa_0((\log m)^2/m)^{1/d}$.
- Make a histogram-style density estimation on the grid using kernel density estimation:

$$\hat{p}(x) = \frac{1}{mh_m^d} \sum_{i=1}^{m} G(H_m^{-1}(X_i - \bar{x})) \tag{2}$$

Where $\bar{x}$ is the closest point to x on the grid, G is the kernel and $H_m = h_m I$





Learning the Decision Sets: Decision Set Estimation

- Locating the jumps in $\hat{p}(x)$: **p-connectivity** of data points.
- Two point $x_1, x_2 \in \mathcal{X}$ are said to be **connected**, if there exists a sequence of points $x_1 = z_1, z_2, \ldots, z_l = x_2$ such that $z_2, ..., z_{l-1} \in U, \|z_j - z_{j+1} \leq 2\sqrt{d}h_m\|$.
- Two point $x_1, x_2 \in \mathcal{X}$ are said to be **p-connected**, if in addition to being **connected**, we have $|\hat{p}(z_i) - \hat{p}(z_j)| \leq (\log m)^{-1/3}$ for all z_i, z_j satisfying $\|z_i - z_j\| \leq h_m \log m$.
- All points that are pairwise **p-connected** specify an empirical decision set $\hat{D}$ and we derived $\hat{\mathcal{D}}$.

Learning the Decision Sets: Guarantee

Lemma 1. *Denote the set of boundary points as*

$$\mathcal{B} := \{z : z_d = g_k^{(p)}(z_1, \ldots, z_{d-1}), k \in \{1, \ldots, K\}, p \in \{1, 2\}\}$$

and define the boundary set as

$$\mathcal{R}_{\mathcal{B}} := \{x : \inf_{z \in \mathcal{B}} \|x - z\| \leq 2\sqrt{d} h_m\}.$$

If $|\gamma| > C_o(m/(\log m)^2)^{-1/d}$, *where* $C_o = 6\sqrt{d}\kappa_0$, *then for all* $p \in \mathcal{P}_X$, *all pairs of points* $x_1, x_2 \in supp(p) \setminus \mathcal{R}_{\mathcal{B}}$ *and all* $D \in \mathcal{D}$, *with probability* $> 1 - 1/m$,

$$x_1 \overset{p}{\leftrightarrow} x_2 \quad \text{if and only if} \quad x_1, x_2 \in D,$$

for large enough $m \geq m_0 \equiv m_0(p_{\min}, K, \kappa_1, d, \alpha_1, B, G, \kappa_0)$.[1]

- Exclude decision boundaries, p-connectivity is equivalent to decision set with high probability.
- Margin γ plays an important role: theorem works only when $|\gamma|$ is in the order of the spacing h_m.

Learning the Decision Sets: Proof

1. Uniform bound for density estimation: Given certain assumptions on Kernels, we have with probability at least $1 - \frac{1}{m}$:

$$\sup_{x \in supp(p) \setminus \mathcal{R}_{\mathcal{B}}} |p(x) - \hat{p}(x)| < \left| h_m^{\min(1,\alpha_1)} + \sqrt{\frac{\log m}{m_m^d}} \right| \tag{3}$$

2. Conectivity: For all $x \in supp(p) \setminus \mathcal{R}_{\mathcal{B}}$, with probability $1 - \frac{1}{m}$, there exsits an unlabeled sample X_i that $\|X_i - x\| < \sqrt{d} h_m$
3. Bound $|\hat{p}(x) - \hat{p}(x')|$ using nearby unlabeled points z, z': $|\hat{p}(x) - \hat{p}(z)|$,$|\hat{p}(z) - \hat{p}(z')|$, $|\hat{p}(z) - p(z)|$ and $|p(z) - p(z')|$

SSL Performance Analysis

- Let $\mathcal{R}(f)$ denote the risk of interest for a given target function f and excess risk $\mathcal{E}(f) = \mathcal{R}(f) - \mathcal{R}^*$, where $\mathcal{R}^*$ is the infimum risk over all possible learners.
- SSL learner$\approx$ clairvoyant learner:

Corollary 1. *Assume that the excess risk $\mathcal{E}$ is bounded. Suppose there exists a clairvoyant supervised learner $\widehat{f}_{\mathcal{D},n}$, with perfect knowledge of the decision sets $\mathcal{D}$, for which the following finite sample upper bound holds*

$$\sup_{\mathcal{P}_{XY}(\gamma)} \mathbb{E}[\mathcal{E}(\widehat{f}_{\mathcal{D},n})] \leq \epsilon_2(n).$$

Then there exists a semi-supervised learner $\widehat{f}_{m,n}$ such that if $|\gamma| > C_o(m/(\log m)^2)^{-1/d}$,

$$\sup_{\mathcal{P}_{XY}(\gamma)} \mathbb{E}[\mathcal{E}(\widehat{f}_{m,n})] \leq \epsilon_2(n) + O\left(\frac{1}{m} + n\left(\frac{m}{(\log m)^2}\right)^{-1/d}\right).$$

SSL Performance Analysis: Proof and Remarks

- To prove the theorem, one only need to use the fact that $\hat{\mathcal{D}}$ is very close to $\mathcal{D}$ in a probability sense. Using condition probability, $\mathbb{E}\,\mathcal{E}(\hat{f}_{m,n})$ is close to $\mathbb{E}\,\mathcal{E}(\hat{f}_{D,n})$.
- Conditions to make SSL Learner be close to clairvoyant learner are:
  1. The margin γ is large enough: $|\gamma| > C_0(m/(\log m)^2)^{-1/d}$
  2. The error term is smaller than $\varepsilon_2(n)$: $(n/\varepsilon_2(n))^d = O(m/(\log m)^2)$
- If the clairvoyant learner outperforms general SL learners:

$$\inf_{f_n} \sup_{P_{XY}} \mathbb{E}[\mathcal{E}(f_n)] \geq \varepsilon_1(n) > \varepsilon_2(n) \tag{4}$$

  We have that there exists a SSL Learner that outperforms general SL learners.

Outline

1. Introduction
   - Conflicting Views in Semi-supervised Learning
   - The Cluster Assumption
2. Finite Sample Analysis of Semi-supervised Learning
   - Summary
   - Learning the Decision Sets
   - SSL Performance Analysis
3. Density-adaptive Regression
   - Optimal Decision Rule
   - SSL Algorithm
   - Error Bounds
4. Concluding Remarks

Optimal Decision Rule: Definition

- Y continuous and bounded random variable
- $f^*(x) = \mathbb{E}[Y|X = X]$, under the squared error loss
- Let $\mathbb{E}_k$ denote expectation with respect to $p_k(Y|X = x)$ and define $f_k(x) = \mathbb{E}_k[Y|X = x]$ then

$$f^*(x) = \sum_{k=1}^{K} \frac{a_k p_k(x)}{\sum_{j=1}^{K} a_j p_j(x)} f_k(x) \qquad (5)$$

- Assumptions:
  1. f_k is uniformly bounded, $|f_k| \leq M$
  2. f_k is Holder-α smooth on C_k

SSL Algorithm

- Since f^* is smooth on each $D \in \mathcal{D}$, perform local polynomial fits within each empirical decision set, using labeled training data that are p-connected
- Use spatially adaptive estimator, optimal for piecewise-smooth functions
- Guarantee SSL still achieves an error bound that is no worse than lower bound for SL when components are indiscernible even with unlabeled data.

SSL Algorithm

- Semi-supervised learner:

$$\hat{f}_{m,n,x}(\cdot) = \arg\min_{f' \in \Gamma} \sum_{i=1}^{n} (Y_i - f'(X_i))^2 \mathbf{1}_{x \overset{p}{\longleftrightarrow} X_i} + \text{pen}(f') \qquad (6)$$

$$\hat{f}_{m,n}(x) \equiv \hat{f}_{m,n,x}(\cdot)$$

- Γ: collection of piecewise polynomials, defined over a recursive dyadic partitioning of the domain $\mathcal{X} = [0,1]^d$
- $\text{pen}(f') \propto log(\sum_{i=1}^{n} \mathbf{1}_{x \overset{p}{\longleftrightarrow} X_i}) \cdot \#f'$, where $\#f'$ is the number cells over which f' is defined

Error Bounds: Overview

- For piecewise Holder-α smooth functions, finite sample error bound of $\max(n^{-2\alpha/(2\alpha+d)}, n^{-1/d})$
- Assume $m \gg n^{2d}$ so that $\sup_{P_{XY}} \mathbb{E}[\mathcal{E}(\hat{f}_{m,n})]$ scales as $\varepsilon_2(n)$
- Assume $d \geq 2\alpha/(2\alpha - 1)$, since when $d < 2\alpha/(2\alpha - 1)$ learning decision sets does not simplify supervised learning task

Margin range γ	SSL upper bound $\epsilon_2(n)$	SL lower bound $\epsilon_1(n)$	SSL helps
$\gamma \geq \gamma_0$	$n^{-2\alpha/(2\alpha+d)}$	$n^{-2\alpha/(2\alpha+d)}$	No
$\gamma \geq c_o n^{-1/d}$	$n^{-2\alpha/(2\alpha+d)}$	$n^{-2\alpha/(2\alpha+d)}$	No
$c_o n^{-1/d} > \gamma \geq C_o(\frac{m}{(\log m)^2})^{-1/d}$	$n^{-2\alpha/(2\alpha+d)}$	$n^{-1/d}$	Yes
$C_o(\frac{m}{(\log m)^2})^{-1/d} > \gamma \geq -C_o(\frac{m}{(\log m)^2})^{-1/d}$	$n^{-1/d}$	$n^{-1/d}$	No
$-C_o(\frac{m}{(\log m)^2})^{-1/d} > \gamma$	$n^{-2\alpha/(2\alpha+d)}$	$n^{-1/d}$	Yes
$-\gamma_0 > \gamma$	$n^{-2\alpha/(2\alpha+d)}$	$n^{-1/d}$	Yes

Error Bounds: Overview

Margin range γ	SSL upper bound $\epsilon_2(n)$	SL lower bound $\epsilon_1(n)$	SSL helps
$\gamma \geq \gamma_0$	$n^{-2\alpha/(2\alpha+d)}$	$n^{-2\alpha/(2\alpha+d)}$	No
$\gamma \geq c_o n^{-1/d}$	$n^{-2\alpha/(2\alpha+d)}$	$n^{-2\alpha/(2\alpha+d)}$	No
$c_o n^{-1/d} > \gamma \geq C_o(\frac{m}{(\log m)^2})^{-1/d}$	$n^{-2\alpha/(2\alpha+d)}$	$n^{-1/d}$	Yes
$C_o(\frac{m}{(\log m)^2})^{-1/d} > \gamma \geq -C_o(\frac{m}{(\log m)^2})^{-1/d}$	$n^{-1/d}$	$n^{-1/d}$	No
$-C_o(\frac{m}{(\log m)^2})^{-1/d} > \gamma$	$n^{-2\alpha/(2\alpha+d)}$	$n^{-1/d}$	Yes
$-\gamma_0 > \gamma$	$n^{-2\alpha/(2\alpha+d)}$	$n^{-1/d}$	Yes

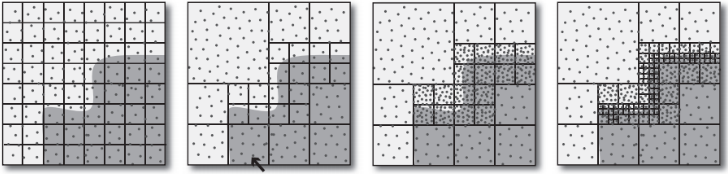
- γ_0 fixed constant, corresponds to considering a fixed collection of distributions whose complexity does not change with the amount of data
- Constants C_0 and c_0 characterize margin and only depend on fixed parameters of the class $P_{XY}(\gamma)$

Error Bounds: Proof

- Based on theorem from Castro 2005. Let $n_D = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{x\in D}$

$$\mathbb{E}[(f^*(X) - \widehat{f}_{D,n}(X))^2 \mathbf{1}_{X\in D} | n_D] \leq C \left(\frac{n_D}{\log n_D}\right)^{-\frac{2\alpha}{d+2\alpha}}.$$

- Decompose the error of the estimator in to three different cases



Outline

- Under the cluster assumption, there exist general situations which SSL can be significantly better than SL in terms of achieving smaller finite sample error bounds than any SL
- Likely that similar conclusion may be drawn under the manifold assumption where the curvature of the manifold will play a similar role to the margin under the cluster assumption
- Showed SSL simplifies learning when there is a link between the marginal and conditional distributions holds
- Interested in SSL whose performance does not deteriorate when the link or margin is not discernible using unlabeled data or does not hold
- Ensure SSL performance is no worse than what SL would achieve such as in Density-adaptive Regression

Further Reading I

Castelli, V., Cover, T.M.
On the exponential value of labeled samples
Pattern Recognition Letters 16(1) (1995) 105-111

Castelli, V., Cover, T.M.
The relative value of labeled and unlabeled samples in pattern recognition
IEEE Transactions on Information Theory 42(6) (1996) 2102-2117

Lafferty, J., Wasserman, L.
Statistical analysis of semi-supervised regression
Advances in Neural Information Processing Systems 20, NIPS. (2008) 801-808

Niyogi, P.
Manifold regularization and semi-supervised learning: Some theoretical analyses
Technical Report TR-2008-01, Computer Science Department, University of Chicago.

Rigollet, P.
Generalization error bounds in semi-supervised classification under the cluster assumption
Journal of Machine Learning Research 8 (2007) 1369-1392

Korostelev, A., Nussbaum, M.
The asymptotic minimax constant for sup-norm loss in nonparametric density estimation
Bernoulli 5(6) (1999) 1099-1118

Castro, R., Willett, R., Nowak, R.
Faster rates in regression via active learning
Advances in Neural Information Processing Systems, NIPS. (2005) 179-186