ERRATA for NIPS 2008 paper "Unlabeled data: Now it helps, now it doesn't"

Aarti Singh, Robert Nowak and Xiaojin Zhu

It is virtually impossible to get anything exactly right - Carl de Boor

The authors would like to report three corrections to the original manuscript. These changes do not effect the main results of this paper, but are needed to make the arguments technically sound.

1) **Definition of margin** γ - We replace the $\|\cdot\|_\infty$ norm in the definition of the margin, with $\inf_{x\in\mathcal{X}}|\cdot|$. The correction definition is given as below.

   The collection $\mathcal{P}_{XY}$ is indexed by a margin parameter γ, which denotes the minimum width of a decision set or separation between the component support sets C_k. The margin γ is assigned a positive sign if there is no overlap between components, otherwise it is assigned a negative sign. Formally, for $j,k \in \{1,\ldots,K\}$, let

   $$d_{jk} := \min_{p,q\in\{1,2\}} \inf_{x\in\mathcal{X}} |g_j^{(p)}(x) - g_k^{(q)}(x)| \qquad j \neq k,$$

   and

   $$d_{kk} := \inf_{x\in\mathcal{X}} |g_k^{(1)}(x) - g_k^{(2)}(x)|.$$

   Then the margin is defined as

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

2) **Properties of kernel** G **used for density estimation** - The kernel G cannot be positive, integrate to one, as well as have vanishing moments of order ≥ 2. Therefore, we remove the restriction that the kernel G be positive. This does not change any of the arguments in the paper. However, since a negative kernel density estimate is not a bonafide density, one can use other techniques proposed in the literature such as estimating the log of the density [1, 2].

3) **Decision set estimation** - We modify the definition of connectedness used to estimate the decision sets. Recall that $h_m = \kappa_0((\log m)^2/m)^{1/d}$ where $\kappa_0 > 0$ is a constant. Let $g_n = (\log n)^{-2}$. Two points that are $2\sqrt{d}g_n$ close in $\mathcal{X} = [0,1]^d$ are connected if there exists a sequence of $2\sqrt{d}h_m$-dense unlabeled data points connecting the two points such that the marginal density varies smoothly

along the sequence, and the sequence lies within a distance of $g_n \log n$ of one of the points. Connectedness is then extended to points that are not $2\sqrt{d}g_n$ close by association. A formal statement is given below.

Two points $x_1, x_2 \in \mathcal{X}$ such that $\|x_1 - x_2\| \leq 2\sqrt{d}g_n$ are said to be *connected*, denoted by $x_1 \leftrightarrow x_2$, if there exists a sequence of points $x_1 = z_1, z_2, \ldots, z_{l-1}, z_l = x_2$ where $z_2, \ldots, z_{l-1} \in \mathcal{U}$, such that

(1) $\|z_j - z_{j+1}\| \leq 2\sqrt{d}h_m$, for all j,

(2) $|\widehat{p}(z_i) - \widehat{p}(z_j)| \leq \delta_n := (\log \log n)^{-1}$ for all i, j, and

(3) $\|x_1 - z_i\| = O(g_n \log n)$ for all i.

In addition, if $x_1 \leftrightarrow x_2$ and $x_2 \leftrightarrow x_3$, then $x_1 \leftrightarrow x_3$. All points that are pairwise connected specify an empirical decision set.

Lemma 1 is now modified as follows:

Lemma 1. *Let ∂D denote the boundary of a decision set D and define the set of boundary points as*

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

If $|\gamma| > 6\sqrt{d}h_m$, then for all $p \in \mathcal{P}_X$, all pairs of labeled data points $X_1, X_2 \in$ supp$(p)\backslash\mathcal{B}$ and all $D \in \mathcal{D}$, with probability $> 1 - 1/m - (\log n)^{2d} e^{-p_{\min} n/(\log n)^{2d}}$,

$$X_1, X_2 \in D \quad \text{if and only if} \quad X_1 \leftrightarrow X_2$$

for large enough $m \geq m_0$ and $n \geq n_0$, where m_0, n_0 depend only on the fixed parameters of the class $\mathcal{P}_{XY}(\gamma)$.

Proof. We will use the density estimation results (Theorem 1 and Corollary 2 in the paper) and establish the result in two steps:

1. $X_1 \in D, X_2 \notin D \Rightarrow X_1 \not\leftrightarrow X_2$: Since X_1 and X_2 belong to different decision sets, all sequences connecting X_1 and X_2 through unlabeled data points pass through a region where either (i) the density is zero and since the region is at least $|\gamma| > 6\sqrt{d}h_m$ wide, there cannot exist a sequence as defined in Section 3 such that $\|z_j - z_{j+1}\| \leq 2\sqrt{d}h_m$, or (ii) the density is positive. In the latter case, the marginal density $p(x)$ jumps by at least $p_{\min}$ one or more times along all sequences connecting X_1 and X_2. Consider the sequences connecting X_1 and X_2 through unlabeled data points such that $\|X_1 - z_i\| = O(g_n \log n)$ for all i in the sequence, and suppose the first jump occurs where decision set D ends and another decision set $D' \neq D$ begins (in the sequence). Then since D' is at least $|\gamma| > 6\sqrt{d}h_m$ wide, by Corollary 2, there exist a point z_i (possibly X_2) in the sequence that lies in $D' \setminus \mathcal{B}$. Since the density on each decision set is Hölder-α smooth and $\|X_1 - z_i\| = O(g_n \log n)$, we have $|p(X_1) - p(z_i)| \geq p_{\min} - O((g_n \log n)^{\min(1,\alpha)})$. Since $X_1, z_i \notin \mathcal{B}$, using Theorem 1, $|\widehat{p}(X_1) - \widehat{p}(z_i)| \geq |p(X_1) - p(z_i)| - 2\epsilon_m > \delta_n$ for large enough m and

n. Thus, $X_1 \not\leftrightarrow X_2$.

2. $X_1, X_2 \in D \Rightarrow X_1 \leftrightarrow X_2$: Using Proposition 1 given below (which is a stronger version of Corollary 2), it holds that with probability $> 1 - g_n^{-d} e^{-p_{\min} g_n^d n}$, the labeled data points are $2\sqrt{d}g_n$ dense. Since connectedness extends by association, it suffices to consider points X_1 and X_2 such that $\|X_1 - X_2\| \leq 2\sqrt{d}g_n$.

Next we show that if $\|X_1 - X_2\| \leq 2\sqrt{d}g_n$, then the shortest path within D that connects X_1 and X_2 is of length $O(g_n)$. If the two points are far from the boundary of D, then they may be connected by a straight line. If one and/or the other is close to the boundary, then the shortest path may need to follow the contour of the boundary to avoid passing outside of D (if the decision set is non-convex). Suppose that both X_1 and X_2 are on the boundary. Then the shortest path between them is along the boundary, and because the boundary is Lipschitz the length of the path is $O(g_n)$. Thus, it follows that, in general, the shortest path between X_1 and X_2 is $O(g_n)$. We argue that there exists a sequence of $2\sqrt{d}h_m$-dense unlabeled data points connecting X_1 and X_2 that lies close to this shortest path and is contained in $D \setminus \mathcal{B}$. Notice that since D has width at least $|\gamma| > 6\sqrt{d}h_m$, there exists a region of width $> 2\sqrt{d}h_m$ contained in $D \setminus \mathcal{B}$ and Corollary 2 implies that with probability $> 1 - 1/m$, there exists a sequence contained in $D \setminus \mathcal{B}$ connecting X_1 and X_2 through $2\sqrt{d}h_m$-dense unlabeled data points that lies within $2\sqrt{d}h_m$ of the shortest path. Therefore, this sequence of $2\sqrt{d}h_m$-dense unlabeled data points also lies within $O(g_n)$ of X_1. Further, since the sequence is contained in D and the density on D is Hölder-α smooth, we have for all points z_i, z_j in the sequence, $|p(z_i) - p(z_j)| \leq O(g_n^{\min(1,\alpha)})$. Moreover, as $z_i, z_j \notin \mathcal{B}$, using Theorem 1, $|\widehat{p}(z_i) - \widehat{p}(z_j)| \leq |p(z_i) - p(z_j)| + 2\epsilon_m \leq \delta_n$ for large enough m and n (depending only on the parameters of the class). Thus, $X_1 \leftrightarrow X_2$. □

Proposition 1. *(Empirical density of labeled data). For all $p \in \mathcal{P}_X$, with probability $> 1 - g_n^{-d} e^{-p_{\min} g_n^d n}$, for all $x \in$ supp(p), $\exists X_i \in \mathcal{L}$, the set of labeled data points, s.t. $\|X_i - x\| \leq \sqrt{d}g_n$.*

Proof. Consider a point $x \in \text{supp}(p)$. Let $\bar{x}$ denotes the point closest to x on a uniform grid over the domain $\mathcal{X} = [0,1]^d$ with spacing g_n. Let $B_{\bar{x}}$ denote a square cell of sidelength g_n centered at $\bar{x}$. If we can show that $B_{\bar{x}}$ contains at least one labeled data point for all $\bar{x}$ with high probability, then it follows that $\exists X_i \in \mathcal{L}$ s.t. $\|X_i - x\| \leq \sqrt{d}g_n$ for all $x \in \text{supp}(p)$. We now bound the probability that no labeled point lies in the ball $B_{\bar{x}}$ for some $\bar{x}$ as follows:

$$
\begin{aligned}
P(\cup_{\bar{x}}\{B_{\bar{x}} \text{ contains no labeled data}\}) &\leq \sum_{\bar{x}} P(B_{\bar{x}} \text{ contains no labeled data}) \\
&\leq \sum_{\bar{x}} \Pi_{i=1}^{n} P(X_i \notin B_{\bar{x}}) \\
&\leq \sum_{\bar{x}} (1 - p_{\min} g_n^d)^n \\
&\leq g_n^{-d} e^{-p_{\min} g_n^d n}
\end{aligned}
$$

Third step follows since $p(x) \geq p_{\min}$ for all $x \in \text{supp}(p)$ and volume of $B_{\bar{x}}$ is g_n^d. The last step follows since $1 - z \leq e^{-z}$ and the number of $\bar{x}$ points are g_n^{-d}. □

This introduces an additional term of $O((\log n)^{2d} e^{-p_{\min} n/(\log n)^{2d}})$ in Corollary 1 bound on the performance of the semi-supervised learner, due to the probability with which Lemma 1 now holds. However, the additional term is $O(\epsilon_2(n))$ (at least up to log factors) in both the regression and classification settings, and thus the conclusions remain the same.

Acknowledgement

The authors would like to thank David Eis for initiating the discussion that led to the corrections mentioned in this document.

References

[1] Hjort, N. L. and Jones, M. C. (1996). Locally parametric nonparametric density estimation. The Annals of Statistics 24 16191647.

[2] Loader, C. (1999a). Local Regression and Likelihood. Springer-Verlag. New York, NY.