Supplementary Material: Local Supervised Learning

Joseph Wang
Dept. of Electrical and Computer Engineering
Boston University
Boston, MA 02116
joewang@bu.edu

Venkatesh Saligrama
Dept. of Electrical and Computer Engineering
Boston University
Boston, MA 02116
srv@bu.edu

1 Proof of Lemma 2.2

The global loss surrogate is given by:

$$\hat{R}(g, f_0, f_1) = \frac{1}{n}\sum_{i=1}^{n} \phi\left(g(x_i)\right)\phi\left(y_i f_0(x_i)\right) + \frac{1}{n}\sum_{i=1}^{n} \phi\left(-g(x_i)\right)\phi\left(y_i f_1(x_i)\right). \tag{1}$$

The data can be viewed as training data of twice the size. Each point is represented twice,once with label 1 and weight are $\phi\left(y_i f_0(x_i)\right)$ and once with label -1 with weight $\phi\left(y_i f_1(x_i)\right)$.

Note that if the loss of both f_0 and f_1 is the same for both classifiers, the point does not effect the learning of the partitioning classifier. Otherwise, even if both f_0 and f_1 produce the same sign for the output, the partitioning classifier will attempt to place the observation in the region with the largest positive margin.

2 Proof of Theorem 2.4

The local linear classifier, F, is composed of the rejection classifiers, $g_1, g_2, \ldots, g_{r-1}$, and the region classifiers, $f_1, f_2, \ldots, f_r$. As the output F can be expressed as a boolean function of $2r-1$ linear functions, each with a VC-dimension of $d+1$, from Lemma 2 of [1], the VC-dimension of the local linear classifier can be bounded:

$$VC\left(F\right) \leq 2(2r-1)\log(e(2r-1))(d+1). \tag{2}$$

3 Proof of Theorem 2.5

Consider the 4 clusters of points, C_1, C_2, C_3, and C_4, located at $(1,0)$, $(0,1)$, $(-1,0)$, and $(0,-1)$, each with an identical number of points ($|C_1| = |C_2| = |C_3| = |C_4|$). For the XOR, without loss of generality, let the points in cluster C_1 and C_3 have a label of -1, and the points in the other cluster have a label of 1. Consider the initial random sampling of the reject region, r_i, and define the rejected set in each cluster as R_1, R_2, R_3, and R_4, where $|R_j| = \sum_{i \in C_j} \mathbb{1}_{r_i=1}$. Suppose we sample equally from each label, ssuch that $|R_1| + |R_3| = |R_2| + |R_4|$. Then with high probability one of the two clusters for each label will have more points. Without loss of generality, assume C_3 has a larger rejected set than the other clusters, i.e. $|R_3| > |R_1|$, $|R_3| > |R_2|$, and $|R_3| > |R_4|$, and therefore C_1 has a smaller rejected set than all other clusters. The class prior probabilities can be expressed:

$$\pi_{-1} = \pi_1 = \frac{1}{2}.$$

The mean of the points labeled -1 that are not rejected can be expressed:

$$[\mu_{-1},\ 0] = \left[\frac{|R_3| - |R_1|}{|C_1| + |C_3| - |R_1| - |R_3|},\ 0\right]^T.$$

Given that more points are rejected from cluster C_3, $\mu_{-1} > 0$. Similarly, the mean of the points labeled 1 that are reject can be written:

$$[0,\ \mu_1] = \left[0,\ \frac{|R_4| - |R_2|}{|C_2| + |C_4| - |R_2| - |R_4|}\right]^T.$$

Given that $|R_3| > |R_2|$, $|R_3| > |R_1|$, $|R_1| < |R_2|$, and $|R_1| < |R_4|$, $\mu_{-1} > |\mu_1|$. Additionally, the covariance can be expressed:

$$\Sigma = \begin{bmatrix} \frac{1-\mu_{-1}^2}{2} & 0 \\ 0 & \frac{1-\mu_1^2}{2} \end{bmatrix}.$$

Points in cluster C_3 have linear discriminant functions:

$$\delta_{-1} = \frac{\mu_{-1}}{(\mu_{-1} - 1)},$$

and

$$\delta_1 = \frac{\mu_1^2}{(\mu_1^2 - 1)}.$$

C_3 will be classified incorrectly if the following inequality holds:

$$\delta_{-1} < \delta_1.$$

This inequality can be rewritten:

$$\begin{aligned} \frac{\mu_{-1}}{\mu_{-1} - 1} &< \frac{\mu_1^2}{\mu_1^2 - 1} \\ \mu_{-1}\left(\mu_1^2 - 1\right) &< \mu_1^2\left(\mu_{-1} - 1\right) \\ \mu_1^2 &< \mu_{-1}, \end{aligned}$$

As $-1 \leq \mu_1 \leq 1$, $0 < \mu_{-1} < 1$, and $\mu_{-1} > |\mu_1|$, the above inequality must hold, and therefore the points in cluster C_3 will be classified incorrectly. Similarly, for the points in C_1, the discriminant functions can be expressed:

$$\delta_{-1} = \frac{\mu_{-1}}{(1 + \mu_{-1})},$$

and

$$\delta_1 = \frac{-\mu_1^2}{(1 - \mu_1^2)}.$$

C_1 will be classified correctly if:

$$\delta_{-1} > \delta_1.$$

This is equivalent to:

$$\frac{\mu_{-1}}{(1 + \mu_{-1})} > \frac{-\mu_1^2}{(1 - \mu_1^2)},$$

which must hold, as $-1 \leq \mu_1 \leq 1$ and $0 < \mu_{-1} < 1$.

For the rejected points, the points in C_1 will be classified incorrectly and the points in C_3 will be classified correctly by f_1. Therefore, C_1 and C_3 will have different rejection labels and will be partitioned into region 0 and 1, respectively. Given these clusters are in separate regions, the data in each region is linearly separable regardless of the partitioning of C_2 and C_4 and will be classified correctly.

Therefore if we reject data points equally with respect to labels, local linear classification using LDA will only fail to converge to the correct answer if $|R_1| = |R_2| = |R_3| = |R_4|$.

References

[1] Eduardo D. Sontag. Vc dimension of neural networks. In *Neural Networks and Machine Learning*, pages 69–95. Springer, 1998.