An Impossibility Result for High Dimensional Supervised Learning

Mohammad H. Rohban, Prakash Ishwar, Birant Orten, William Clem Karl and Venkatesh Saligrama

Boston University, Department of Electrical and Computer Engineering, Information Sciences and Systems Group

Goal

Establishing a strong notion of impossibility of learning

- Binary supervised classification : $\mathcal{T}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$.
- Class conditional densities are Gaussian : $\mathcal{N}(\mu_\pm, \Sigma)$.
- High Dimensional Setting : $d/n \to \infty$ as $n \to \infty$.
- The optimal Bayes error does not scale with n and d.
- Can we lower bound worst case probability of error for any classifier asymptotically?

Preliminaries

- Training set $\mathcal{T}_n$, with n iid samples from $p(\mathbf{x}, y|\theta)$.
- Learning rule $\widehat{y}_{\mathcal{T}_n} : \mathbb{R}^d \to \{-1, 1\}$.
- Conditional Probability of error : $P_{e|\mathcal{T}_n,\theta} = \Pr_{(\mathbf{x},y)}(\widehat{y}_{\mathcal{T}_n}(\mathbf{x}) \neq y|\mathcal{T}_n, \theta)$.
- Probability of error : $P_{e|\theta} = \mathbb{E}_{\mathcal{T}_n}(P_{e|\mathcal{T}_n,\theta})$.
- Bayes optimal error : $P^*_{e|\theta} = \min_{\widehat{y}} P_{e|\theta}$.
- Set of feasible parameters to keep $P^*_{e|\theta}$ fixed :

$$\Theta(\alpha) = \{\theta : P^*_{e|\theta} = Q(\alpha/2)\}$$

- Worst case probability of error :

$$P_e(n, d, \Theta(\alpha), \widehat{y}) = \sup_{\theta \in \Theta(\alpha)} P_{e|\theta}$$

- Goal : Evaluation of $\liminf_{(d,n/d)\to(\infty,0)} P_e(n, d, \Theta(\alpha), \widehat{y}_{\mathcal{T}_n})$.
- It is generally known that if the number of samples n is much less than the VC-dimension V, the learning is impossible in the *distribution free* setting.

Theorem (Anthony and Bartlett, 1999)

If $P^*_e = c$ and $n < V/(320\epsilon^2)$, then for all supervised learning rules $\widehat{y}_{\mathcal{T}_n}$, and all $\epsilon, \delta \in (0, 1/64)$,

$$\sup_{p(\mathbf{x},y):P^*_e=c} \Pr(P_{e|\mathcal{T}_n,\theta} - c \geq \epsilon) \geq \delta$$

Major Differences in our Results

- The set of distributions is not as large as all distributions and is not as small as only a single distribution.
- The worst case is evaluated only on a subset of Gaussian distributions in our results.

Feasible Sets for Gaussian Distribution

- $\theta = (\mu_+, \mu_-, \Sigma)$ and $P^*_{e|\theta} = Q\left(\frac{1}{2}\|\Sigma^{-\frac{1}{2}}(\mu_+ - \mu_-)\|_2\right)$
- $\Theta_0(\alpha) = \left\{(\mu_+, \mu_-, \Sigma) : \|\Sigma^{-\frac{1}{2}}(\mu_+ - \mu_-)\|_2 = \alpha\right\}$
- A canonical subset of $\Theta_0(\alpha)$, which is easier to analyze

$$\Theta_{\text{Sphere}}(\alpha) := \left\{\left(\mathbf{h}, -\mathbf{h}, \beta^2\mathbf{I}\right) : \|\mathbf{h}\|_2 = 1, \beta = 2/\alpha\right\}.$$

Theorem (Main Result)

For any sequence of classifiers $\widehat{y}_{\mathcal{T}_n}$, and $\alpha \geq 0$, we have

$$\liminf_{(d,n/d)\to(\infty,0)} P_e(n, d, \Theta_{\textit{Sphere}}(\alpha), \widehat{y}_{\mathcal{T}_n}) \geq \frac{1}{2}$$

Discussion

- Even in the easier scenario of Θ_{Sphere}, there is no hope to get meaningful classification.
- Trivial to show that the same result holds for $\Theta_0(\alpha) \supseteq \Theta_{\text{Sphere}}(\alpha)$.
- It is consistent with the previous results :
  - If ML estimation of $\mu_\pm$ and Σ are plugged into the optimal Bayes classifier, error is asymptotically not less than half [Bickel and Levina, 2004].
  - Suppose that $\Theta_1(\alpha) = \{(\mathbf{h}, -\mathbf{h}, \mathbf{h}\mathbf{h}^\top + \sigma^2\mathbf{I}) : P^*_{e|\mathbf{h}} = Q(\alpha/2)\}$. If ML estimation of $\mathbf{h}$ is plugged into the optimal Bayes classifier, the performance is no better than a random coin toss asymptotically [Orten et al, 2011].

Necessary Condition on Θ for Learnability

- $\Theta_{\text{subset}} := \{(\mathbf{h}, -\mathbf{h}, \beta^2\mathbf{I}) \in \Theta_{\text{Sphere}}, \mathbf{h} \in \mathcal{H} \subseteq \mathcal{S}^{d-1}\}$.
- Let $\text{vol}(\mathcal{H}) \triangleq \Pr_{H\sim U(\mathcal{S}^{d-1})}(H \in \mathcal{H})$.

Corollary

Suppose that $\lim_{d\to\infty} \text{vol}(\mathcal{H})$ exists. If for a sequence of classifiers $\widehat{y}_{\mathcal{T}_n}$,

$$\limsup_{(d,n/d)\to(\infty,0)} P_e(n, d, \Theta_{\textit{Sphere}}, \widehat{y}_{\mathcal{T}_n}) = \frac{1}{2}$$

and

$$\limsup_{(d,n/d)\to(\infty,0)} P_e(n, d, \Theta_{\textit{subset}}, \widehat{y}_{\mathcal{T}_n}) < \frac{1}{2}$$

then

$$\lim_{d\to\infty} \text{vol}(\mathcal{H}) = 0.$$

Discussion

- Consistent classifiers exist if $\mathbf{h}$ belongs to either of these sparsity classes [Orten et al, 2011]:
  - Assume that sorted absolute values of components of $\mathbf{h}$ $(h_{(1)}, \ldots, h_{(d)})$ decay exponentially or polynomially fast :

$$\mathcal{H}_{exp} = \left\{\mathbf{h} : \left|h_{(k)}\right| = M_1(d)\alpha^k, 0 < \alpha < 1\right\}$$
$$\mathcal{H}_{poly} = \left\{\mathbf{h} : \left|h_{(k)}\right| = M_2(d)k^{-\beta}, \beta > 0.5\right\}$$

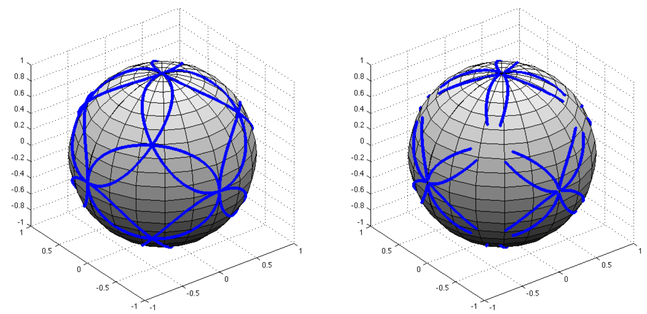

Figure: Left : $\mathcal{H}_{exp}$ - Right : $\mathcal{H}_{poly}$ for $d = 3$.

Boston University, Boston, MA email : mhrohban@bu.edu Homepage: http://blogs.bu.edu/mhrohban