An Impossibility Result for High Dimensional Supervised Learning

Mohammad Hossein Rohban
(joint work with P. Ishwar, B. Orten, W. C. Karl, and V. Saligrama)

ISS Group, ECE Dept. Boston University

12 Apr. 2013

Table of Contents

1. Our Problem Settings
2. The Goal
3. VC Theory ?!
4. Main Results and Discussion
5. Proof Techniques
6. Conclusion

Problem Setting

- We are given iid samples $\mathcal{T}_n = \{(\mathbf{x}_1, y_1) \ldots, (\mathbf{x}_n, y_n)\}$ as a training set.
- $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$.
- We have the same a priori class probabilities : $\Pr(y = 1) = \Pr(y = -1)$.
- $(\mathbf{x}_i, y_i) \sim p(\mathbf{x}, y|\theta)$.
- Class conditional densities are Gaussians with the same covariance matrix. $\theta = (\mu_+, \mu_-, \Sigma)$.
- It is known that
  - Defining $\Delta = (\mu_+ - \mu_-)$ and $\mu = \frac{\mu_+ + \mu_-}{2}$:

$$\begin{aligned} \widehat{y}^*(\mathbf{x}) &= \underset{y \in \{-1,+1\}}{\arg\max} \; p_{Y|X,\theta}(y|\mathbf{x}, \theta) \\ &= \operatorname{sign}\left(\Delta^\top \Sigma^+ (\mathbf{x} - \mu)\right) \end{aligned} \tag{1}$$

  - $P_e^* \triangleq \min_{\widehat{y}} \Pr(\widehat{y}(\mathbf{x}) \neq y(\mathbf{x})) = Q\left(\frac{1}{2}\|\Sigma^{-\frac{1}{2}}(\mu_+ - \mu_-)\|_2\right)$.

Problem Setting (cont.)

- Scaling Regime :
  - High Dimensional Setting : as $n \to \infty$, $n/d \to 0$.
  - P_e^* is kept the same as n changes :
    $\theta \in \Theta(\alpha) = \left\{(\mu_+, \mu_-, \Sigma) : \|\Sigma^{-\frac{1}{2}}(\mu_+ - \mu_-)\|_2 = \alpha\right\}$
- The goal is to prove some asymptotic lower bound for error probability of *every classifier* in the worst case.
- Differences from previous work
  - $n/d \to c > 0$ as $(n, d) \to \infty$ [Donoho, Jin 2004] and [Singh et al 2010].
  - Analyzing asymptotic behavior of *specific classifiers* : plug-in rules [Shao et al 2012] and [Orten et al 2011] or Fisher Linear Discriminant, Naive Bayes rule [Bickel and Levina 2004].

The Goal

- Supervised Classification Rule : $\widehat{y}_{\mathcal{T}_n} : \mathbb{R}^d \to \{-1, 1\}$.
- Defining error probability of $\widehat{y}_{\mathcal{T}_n}$ conditioned on θ as

$$P_{e|\theta}(\widehat{y}_{\mathcal{T}_n}) = \Pr(\widehat{y}_{\mathcal{T}_n}(\mathbf{x}) \neq y|\theta) \tag{2}$$

- Defining *worst case* error probability of $\widehat{y}_{\mathcal{T}_n}$ as $P_e(n, d, \Theta, \widehat{y}_{\mathcal{T}_n}) \triangleq \sup_{\theta \in \Theta} P_{e|\theta}(\widehat{y}_{\mathcal{T}_n})$.
- **Goal** : Find a lower bound on $\liminf_{(d, n/d) \to (\infty, 0)} P_e(n, d, \Theta(\alpha), \widehat{y}_{\mathcal{T}_n})$ for all learning rules $\widehat{y}$ (possibly aware of the *Gaussianity* and *structure of* Θ) and problem difficulties α.

Goal (cont.)

- $\Theta_{\text{Sphere}}(\alpha)$ is a canonical subset of $\Theta(\alpha)$ which is of special interest

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

- Consider the case that $\mathbf{x} = y\mathbf{h} + \mathbf{z}$, where $\mathbf{z}$ is the WGN.
  - Can be considered as a model with latent variables y and $\mathbf{h}$.
  - $\Theta_{\text{Sphere}}(\alpha) \subseteq \Theta(\alpha)$: Clearly $P_e(\Theta(\alpha))$ is no smaller than $P_e(\Theta_{\text{Sphere}}(\alpha))$.

VC Theory ?!

Theorem (Anthony and Biggs 1990)

*Let $\mathcal{H}$ be a hypothesis space for labeling function with VC dimension d. For any learning algorithm $\mathcal{A}$ working with $\mathcal{H}$ (***which is only aware of*** $\mathcal{T}_n$), there* ***exist*** *distributions such that with probability at least δ over n random samples, the error probability of $\widehat{y} = \mathcal{A}(\mathcal{T}_n)$ given $\mathcal{T}_n$ is at least*

$$\max\left(\frac{d-1}{32n}, \frac{1}{n}\log\left(\frac{1}{\delta}\right)\right)$$

VC Theory ?! (cont.)

Theorem (Devroye and Lugosi 1995)

*Assume that the optimal Bayes rule is contained in $\mathcal{H}$ with VC dimension of d. For any learning algorithm $\mathcal{A}$ (***which is only aware of $\mathcal{T}_n$***), we have*

$$\sup_{p(\mathbf{x},y):P_e^*=L} \mathbb{E}\left[\Pr(\widehat{y}(\mathbf{x}) \neq y|\mathcal{T}_n) - L\right] = \Omega\left(\sqrt{\frac{d}{n}}\right)$$

with $\widehat{y} = \mathcal{A}(\mathcal{T}_n)$

VC Theory ?! (cont.)

- Learning is impossible due to these results even for the class of linear classifiers in our scaling regime!
- Why it doesn't completely solve our impossibility problem?

Main Results

Theorem

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

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

Main Results (cont.)

Corollary

For any sequence of parameter sets Θ *with* $\Theta_{Sphere} \subseteq \Theta$*, and any sequence of classifiers* $\widehat{y}_{\mathcal{T}_n}$*, we have*

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

Discussion

- Consistent with the impossibility results for plug-in classifiers (PIC) :
  - First estimate the parameters of generative distributions. Then, plug the estimations in the optimal Bayes rule.
  - [Bickel and Levina 2004] and [Orten et al 2011] have shown that the classification error of PIC converges to $\frac{1}{2}$ in the general setting of Θ.
  - [Orten et al 2011] has shown that the error probability of PIC converges to $\frac{1}{2}$ in a simpler setting of $\Theta_{\text{Sensing Aware}}$:

$$\Theta_{\text{Sensing Aware}}(\alpha) := \left\{ \left(m_1\mathbf{h}, m_2\mathbf{h}, \gamma^2\mathbf{h}\mathbf{h}^\top + \beta^2\mathbf{I}\right) : \|\mathbf{h}\| = 1, \gamma \geq 0, \beta > 0, |m_1 - m_2| = \alpha\sqrt{\gamma^2 + \beta^2} \right\}.$$

Main Results (cont.)

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

- There are achievability results for $\Theta_{\text{Sensing Aware}}$ (and hence Θ_{Sphere}) based on the sparsity of $\mathbf{h}$:
  - Assume that sorted absolute values of components of $\mathbf{h}$ $(h_{(1)}, \ldots, h_{(d)})$ decay exponentially or polynomially fast :

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

  - Consistent estimation of $\mathbf{h}$ is possible through some soft thresholding of ML estimate of $\mathbf{h}$.

Discussion (cont.)

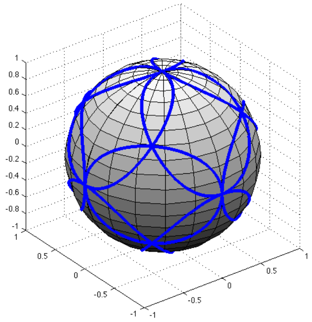

Figure: Exponential sparsity class $\mathcal{H}_{exp}$ (solid curves, top figure) and polynomial sparsity class $\mathcal{H}_{poly}$ (solid curves, bottom figure) for $d = 3$.

Proof Idea

- The key idea is to randomize the selection of θ.
- The worst case error probability is lower bounded by the average error probability of the classification scheme over different selections.
- It is in turn lower bounded by the average error probability of the so called marginalized MAP classifier :

$$\widehat{y}_{\mathsf{MAP}}(X_0) \triangleq \underset{y_0 \in \{-1,+1\}}{\arg\max}\; p_{Y_0|X_0,\mathcal{T}_n}(y_0|x_0, \mathcal{T}_n, \Theta_{\mathsf{Sphere}})$$
$$= \underset{y_0 \in \{-1,+1\}}{\arg\max} \int_{\theta \in \Theta_{\mathsf{Sphere}}} p_{Y_0,X_0,\mathcal{T}_n|\theta}(y_0, x_0, \mathcal{T}_n|\theta) p(\theta) d\theta \tag{3}$$

Proof Idea (cont.)

- Issues :
  - Choose a suitable distribution over θ.
  - Evaluate the integral to find $\widehat{y}_{\text{MAP}}$.
  - Find the average error probability of $\widehat{y}_{\text{MAP}}$.
- For $\mathbf{h}$ sampled from the uniform distribution over the unit sphere :

$$\widehat{y}_{\text{MAP}}(\mathbf{x}_0) = \text{sign}\left(\mathbf{x}_0^\top \left(\sum_{i=1}^{n} y_i \mathbf{x}_i\right)\right)$$

- The Bayes rule is $y^*(\mathbf{x}_0) = \text{sign}\left(\mathbf{x}_0^\top \mathbf{h}\right)$.
- $\widehat{y}_{\text{MAP}}$ looks like a plug-in classifier.

Conclusions

- Prior knowledge is essential in high dimensions setting. Otherwise there is no hope to get something meaningful even for a simple Gaussian distribution.
- Future work : Is it possible to extend this result to other distributions?