Research Overview

Salimeh Yasaei Sekeh

University of Michigan-Ann Arbor

March, 2017

Background and Motivation

- Consider the classification problem of feature vector $\mathbf{X}$, into one of two classes, $\{0, 1\}$. The Bayes classifier assigns a vector $\mathbf{X}$ to the class with the highest posterior probability and Bayes error rate (BER):

$$\epsilon^{\text{Bayes}} = \int\limits_{pf_0(\mathbf{x}) \leq qf_1(\mathbf{x})} p\ f_0(\mathbf{x})\ \mathrm{d}\mathbf{x} + \int\limits_{pf_0(\mathbf{x}) \geq qf_1(\mathbf{x})} q\ f_1(\mathbf{x})\ \mathrm{d}\mathbf{x}. \tag{1}$$

  where f_0, f_1 are the conditional distributions and p, q are the prior probabilities.
- **Problem:** Computing BER requires evaluating a complicated multi-dimensional integral.
- **Solution:** One can evaluate simpler expressions that specify bounds for BER in terms of measures of distance or divergence between probability functions, such as Bhattacharyya distance, see Kailath (1967).

Background and Motivation

- **One more problem:** When the distributions f_0, f_1 are unknown, these bounds cannot be evaluated. **So** it may be interesting to estimate f_0, f_1 and subsequently these bounds from empirical data.
- **Better solution?**

* Nonparametric Divergence Measure (Henze and Penrose divergence), Berisha and Hero (2015):

$$D_p(f_0, f_1) = \frac{1}{4pq}\left[\int \frac{(pf_0(\mathbf{x}) - qf_1(\mathbf{x}))^2}{pf_0(\mathbf{x}) + qf_1(\mathbf{x})}\mathrm{d}\mathbf{x} - (p-q)^2\right].$$

D_p belongs to the class of f-divergences and

1. $0 \le D_p \le 1$
2. $D_p = 0 \Leftrightarrow f_0(\mathbf{x}) = f_1(\mathbf{x})$
3. $D_p(f_0, f_1) = D_q(f_1, f_0)$.

Background and Motivation

Remarkable Properties:

- D_p can be estimated directly without estimation or plug-in of the densities f_0 and f_1 based on an extension of the Friedman-Rafsky (FR) multi-variate two sample test statistic: Consider sample realizations $\mathbf{X}_0 \in \mathbb{R}^{m\times d}$ from f_0 and $\mathbf{X}_1 \in \mathbb{R}^{n\times d}$ from f_1. As $m \to \infty$ and $n \to \infty$ such that $\frac{m}{m+n} \to p$,

$$1 - \mathcal{C}(\mathbf{X}_0, \mathbf{X}_1)\frac{m+n}{2mn} \to D_p(f_0, f_1), \quad a.s.$$

  Here $\mathcal{C}(\mathbf{X}_0, \mathbf{X}_1)$: # edges connecting a data point from f_0 to a data from f_1 in first generating a Euclidean minimal spanning tree (MST) on data set $\mathbf{X}_0 \cup \mathbf{X}_1$.
- There exists a local relationship between D_p and Chernoff α-divergence.
- D_p gives tighter bounds on the BER than those based on the Battacharya distance.
- Given a hypothesis, h, the target error can be bounded by the error on the source data, the difference between labels and D_p between source and target distributions in case of classification problem that they come from different distributions.

Convergence rate

- Bias: Let $d \geq 2$, and $\mathcal{C}(\mathbf{X}_0, \mathbf{X}_1)$ be FR-statistics. Then for Hölder continuous density functions f_0 and f_1

$$\left| \frac{\mathbb{E}\left[\mathcal{C}(\mathbf{X}_0, \mathbf{X}_1)\right]}{m+n} - 2pq \int \frac{f_0(\mathbf{x}) f_1(\mathbf{x})}{p f_0(\mathbf{x}) + q f_1(\mathbf{x})} \, \mathrm{d}\mathbf{x} \right|$$

$$\leq \begin{cases} O\left((m+n)^{-\eta^2 \big/ (d(\eta+1))} \right), & d \geq 4, \\ \text{mín}\left\{ O\left((m+n)^{-\eta d((1/d)+s-1) \big/ (1-d(\eta+1))} \right), \right. \\ \left. O\left((m+n)^{-\eta^2 \big/ (d(\eta+1))} \right) \right\}, & d = 2, 3. \end{cases} \tag{2}$$

Here $s = \dfrac{(1-1/d)\eta}{d\left((1-1/d)\eta + 1\right)}$ and the smoothness Hölder parameter $0 > \eta \leq 1$.

Convergence rate

- Variance: The variance of the estimator HP-integral, $\mathfrak{R}_{m,n}/(m+n)$ is bounded by

$$Var\left(\frac{\mathfrak{R}(\mathfrak{X}_m,\mathfrak{Y}_n)}{m+n}\right) \leq \frac{32\ c_d^2\ p}{(m+n)}, \tag{3}$$

where constant c_d depends only on d.

Some More Theoretical Properties

- (Convexity of the D_p): For given $\lambda_1, \lambda_2 \in [0,1]$ with $\lambda_1 + \lambda_2 = 1$,

$$D_p(\lambda_1 f_1 + \lambda_2 f_2, \lambda_1 g_1 + \lambda_2 g_2) \leq \lambda_1 D_p(f_1, g_1) + \lambda_2 D_p(f_2, g_2).$$

  Equality occurs iff $\lambda_1 \lambda_2 = 0$ or $f_1 = f_2$ and $g_1 = g_2$.

- (Bounds on D_p): For appropriately smooth families of distributions $\{f_{\boldsymbol{\theta}}\}$, under a specific set $\mathbb{S}_p(\mathbb{S}_p^c)$, one can bound the D_p by Fisher information matrix $\mathbf{J}_{\boldsymbol{\theta}}$:

$$D_p(f_{\boldsymbol{\theta}_1}, f_{\boldsymbol{\theta}_2}) \leq (\geq) \\ 1 - \left(p \exp\left\{ \frac{1}{2} (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^{\mathrm{t}} \mathbf{J}_{\boldsymbol{\theta}_1} (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) - o(\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2) \right\} + q \right)^{-1}.$$

♣ Remark: We have obtained some of these and other properties, by using properties of MST such as subadditivity, superadditivity for bounded MST, smoothness and so on.

P-Mutual Information

For parameters $p \in (0,1)$, $p+q=1$, P-mutual information, I_p is defined by

$$I_p(\mathbf{X}:\mathbf{Y}) = \frac{1}{4pq}\left[\int \frac{pf(\mathbf{x},\mathbf{y}) - qf(\mathbf{x})g(\mathbf{y})}{pf(\mathbf{x},\mathbf{y}) + qf(\mathbf{x})g(\mathbf{y})}\,\mathrm{d}\mathbf{xy} - (p-q)^2\right],$$

where $f(\mathbf{x},\mathbf{y})$ denotes joint and $f(\mathbf{x})$, $g(\mathbf{y})$ stand marginal PDFs for RVs $\mathbf{X}$,$\mathbf{Y}$.

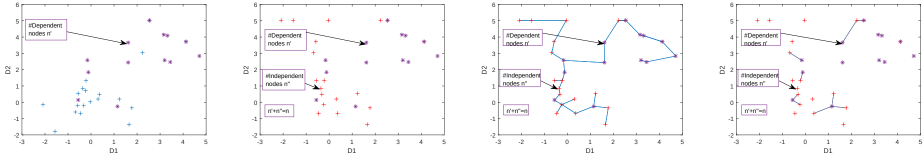
Properties of I_p:

- I_p has concavity in $f(\mathbf{x})$ and convexity in $f(\mathbf{y}|\mathbf{x})$.
- The chain rule for I_p can be established.
- We can also represent an analogue form of the data processing inequality.

Friedman-Rafsky Estimator for HP-MI

- Why HP-mutual information? It can be estimated directly without estimation or plug-in of the densities.

  Consider n samples from joint pdf $f_{\mathbf{XY}}$:



Friedman-Rafsky Estimatorfor HP-MI

- Consider the sample $\mathbf{Z}_n = \mathbf{X}_n \times \mathbf{Y}_n = \left\{ \left(\mathbf{x}_i, \mathbf{y}_i \right)_{i=1}^{n} \right\}$ from joint density function $f_{\mathbf{X},\mathbf{Y}}$ with marginal density functions $f_{\mathbf{X}}$, $f_{\mathbf{Y}}$.
- Divide the sample set $\mathbf{Z}_n$ in two subsets $\mathbf{Z}'_{n'}$ and $\mathbf{Z}''_{n''}$ with the sample proportion $\alpha = n'/n$.
- Denote $\widetilde{\mathbf{Z}}_{n''} = \left\{ (\mathbf{x}_{i_k}, \mathbf{y}_{j_k}),\ k = 1, \dots, n'',\ \text{selected at random from } \mathbf{Z}''_{n''} \right\}$.
- Construct the minimal spanning tree (MST) on the concatend data set, $\mathbf{Z}'_{n'} \cup \widetilde{\mathbf{Z}}_{n''}$.
- Count the edges connecting a node in $\mathbf{Z}'_{n'}$ to a node of $\widetilde{\mathbf{Z}}_{n''}$.
- Output: The FR estimator $\mathfrak{R}_{n',n''}$.

♣ Convergence rate for this estimator is under progress.

I_p In Terms of Copula Density

A multivariate generalization I_p for a d RV $\mathbf{X} = (X_1, \ldots, X_d)$ with marginal PDFs $f_i(x_i)$ and copula density $c(\mathbf{u})$ is given by

$$I_p(\mathbf{X}) = \frac{1}{4pq}\left[\int \frac{pf(\mathbf{x}) - q\prod_i f_i(x_i)}{pf(\mathbf{x}) + q\prod_i f_i(x_i)}\,\mathrm{d}\mathbf{x} - (p-q)^2\right]$$

$$1 - \int_{[0,1]^d} \frac{c(\mathbf{u})}{p\,c(\mathbf{u}) + q}\,\mathrm{d}\mathbf{u} = 1 - \mathbb{E}_C\left[(p\,c(\mathbf{U}) + q)^{-1}\right] := I_p(c)(\text{say}).$$

Interesting relations:

- Pearson's ϕ^2-statistic:

$$I_p(c) = \sum_{n=2}^{\infty}(-1)^n p^{n-1} q \int_{[0,1]^d} (c(\mathbf{u}) - 1)^n \mathrm{d}\mathbf{u}.$$

- Renyi copula entropy, $h_\alpha(c) = -I_\alpha(\mathbf{X})$:

$$1 - e^{p\,h_q(c)} \leq I_p(c) \leq \frac{p\left(e^{-h_2(c)} - 1\right)}{pe^{-h_2(c)} + q}.$$

I_p In Terms of Copula Density

♣ **How can copula be used to estimate I_p:**

Theorem

Consider sample realizations from $c(\mathbf{u})$ and uniform copula density $c_0(\mathbf{u})$ denoted by $\mathbf{U} \in [0,1]^{n\times d}$ and $\mathbf{U}_0 \in [0,1]^{m\times d}$. The FR statistics, $\mathfrak{R}^c_{m,n}$, is constructed on the data set $\mathbf{U} \cup \mathbf{U}_0$ estimates $I_p(\mathbf{X})$ as $n \to \infty$ and $m \to \infty$ in the regime $\dfrac{n}{n+m} \to p$ and $\dfrac{n}{n+m} \to q$, that is

$$1 - \mathfrak{R}^c_{m,n}(\mathbf{U}, \mathbf{U}_0)\frac{n+m}{2nm} \to I_p(c) = I_p(\mathbf{X}).$$

References

- N. Henze and M. D. Penrose, **On the multivarite runs test**, *Ann. Statist.*, vol. 27, no. 1, pp. 290-298.
- V. Berisha and A. O. Hero, **Empirical non-parametric estimation of the Fisher information**, *IEEE Signal Process. Lett.*, vol. 22, no. 7, pp. 988-992, July 2015.
- V. Berisha, A. Wisler, A. Hero and A. Spanias, **Empirically estimable classification bounds based on a nonparametric divergence measure**, *IEEE Trans. on Signal Process.*, vol. 64, no. 3, pp. 580–591, Febuarary 2016.
- H. Liu, J. Lafferty and L. Wasserman, **The nonparanormal: Semiparametric estimation of high dimensional undirected graphs**, *Journal of Machine Learning Research*, 10, pp. 2295–2328, 2009.