Research Overview

Salimeh Yasaei Sekeh

University of Michigan-Ann Arbor

June, 2016

- 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):

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

  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).

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

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.

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 $\mathbf{J}_{\boldsymbol{\theta}}$:

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

  Question: Can we obtain some of these and/or other properties, by using properties of MST such as subadditivity, superadditivity for bounded MST, smoothness and so on?

  ♣ In fact this is one of our goals and we're working on it!

P-Mutual Information

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

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

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.

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

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

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

  Let $\mathbf{X}_1, \ldots, \mathbf{X}_m$ be i.i.d samples having distributions $F_{\mathbf{X}}$ and $\mathbf{U} = F(\mathbf{X})$ be a RV drawn from the copula density. Further $(\widehat{\mathbf{U}}_1, \ldots \widehat{\mathbf{U}}_m) \in [0,1]^{m \times d}$ denotes the empirical copula sample, where the j-th coordinator of $\widehat{\mathbf{U}}_i$, $\widehat{U}_i^j$, is the ratio of the number of elements in $\{X_1^j, \ldots, X_m^j\}$ less than or equal to X_i^j over m:

  $$\widehat{U}_i^j = \frac{1}{m} \text{rank}\Big(X_i^j, \{X_1^j, X_2^j, \ldots, X_m^j\}\Big),$$

  here $\text{rank}(x, A)$ is the number of elements of A less than or equal to x.

- Generate sample $\widehat{\mathbf{U}}^0 = (\widehat{U}_1^0, \ldots, \widehat{U}_n^0) \in [0,1]^{n \times d}$ from uniform copula, $c^0(\mathbf{u}) = 1$. Let $\mathcal{R}_c(\widehat{\mathbf{U}}, \widehat{\mathbf{U}}^0)$ denote the Friedman-Rafsky (FR) statistic: The number of edges connecting a data point from $c(\mathbf{u})$ to a data to $c^0(\mathbf{u})$ in MST on data set $\widehat{\mathbf{U}}, \widehat{\mathbf{U}}^0$.

Conjecture: For asymptotic, take $m \to \infty$ and $n \to \infty$ in a linked manner so that $\dfrac{m}{m+n} \to p \in (0,1)$ and $p + q = 1$, then

$$\frac{\mathcal{R}_c(\widehat{\mathbf{U}}, \widehat{\mathbf{U}}^0)}{m+n} \to 2pq \int_{[0,1]^d} \frac{c(\mathbf{u})}{p\, c(\mathbf{u}) + q}\, \mathrm{d}\mathbf{u} \quad a.s.$$

- This implies

$$1 - \frac{m+n}{2mn} \mathcal{R}_c(\widehat{\mathbf{U}}, \widehat{\mathbf{U}}^0) \to I_p(c).$$

Main references:

- 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.

- V. Berisha and A. Hero. Empirical non-parametric estimation of the Fisher information, *IEEE Signal Process. Lett.*, vol. 22, no. 7, pp. 988–992, July 2015.

- N. Henze and M. D. Penrose. On the multivariate runs test, *Ann. Statist.*, vol. 27, no. 1, pp. 290-298.