Research Overview

Salimeh Yasaei Sekeh

University of Michigan-Ann Arbor
March, 2017
Consider the classification problem of feature vector $X$, into one of two classes, $\{0, 1\}$. The Bayes classifier assigns a vector $X$ to the class with the highest posterior probability and Bayes error rate (BER):

$$
\epsilon_{\text{Bayes}} = \int_{p f_0(x) \leq q f_1(x)} p f_0(x) \, dx + \int_{p f_0(x) \geq q f_1(x)} q f_1(x) \, dx.
$$

(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):

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

- \( D_p \) belongs to the class of \( f \)-divergences and
  1. \( 0 \leq D_p \leq 1 \)
  2. \( D_p = 0 \iff f_0(x) = f_1(x) \)
  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 $X_0 \in \mathbb{R}^{m \times d}$ from $f_0$ and $X_1 \in \mathbb{R}^{n \times d}$ from $f_1$. As $m \to \infty$ and $n \to \infty$ such that $\frac{m}{m+n} \to p$, 

$$1 - C(X_0, X_1) \frac{m+n}{2mn} \to D_p(f_0, f_1), \quad a.s.$$ 

Here $C(X_0, X_1)$: \# 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 $X_0 \cup X_1$.

- There exists a local relationship between $D_p$ and Chernoff $\alpha$-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.
**Bias:** Let $d \geq 2$, and $C(X_0, X_1)$ be FR-statistics. Then for Hölder continuous density functions $f_0$ and $f_1$

$$
\left| \frac{\mathbb{E}[C(X_0, X_1)]}{m + n} - 2pq \int \frac{f_0(x)f_1(x)}{pf_0(x) + qf_1(x)} \, dx \right|
$$

$$
\leq \left\{ \begin{array}{ll}
O \left( (m + n)^{-\eta^2/(d(\eta+1))} \right), & d \geq 4, \\
\min \left\{ O \left( (m + n)^{-\eta d((1/d)+s-1)/(1-d(\eta+1))} \right), O \left( (m + n)^{-\eta^2/(d(\eta+1))} \right) \right\}, & d = 2, 3.
\end{array} \right.
$$

Here $s = \frac{(1 - 1/d)\eta}{d ((1 - 1/d)\eta + 1)}$ and the smoothness Hölder parameter $0 > \eta \leq 1$. 
Variance: The variance of the estimator HP-integral, $\mathcal{R}_{m,n}/(m + n)$ is bounded by

$$ Var\left( \frac{\mathcal{R}(X_m, Y_n)}{m + n} \right) \leq \frac{32 \ c_d^2 \ p}{(m + n)} , $$

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 $\{f_\theta\}$, under a specific set $S_p(S_p^c)$, one can bound the $D_p$ by Fisher information matrix $J_\theta$:

\[
D_p(f_{\theta_1}, f_{\theta_2}) \leq \left(1 - \left(p \exp \left\{\frac{1}{2} (\theta_1 - \theta_2)^t J_\theta (\theta_1 - \theta_2) - o(\|\theta_1 - \theta_2\|^2)\right\} + q\right)\right)^{-1}.
\]

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

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

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

**Properties of $I_p$:**

- $I_p$ has concavity in $f(x)$ and convexity in $f(y|x)$.
- The chain rule for $I_p$ can be established.
- We can also represent an analogue form of the data processing inequality.
Why HP-mutual information? It can be estimated directly without estimation or plug-in of the densities.

Consider $n$ samples from joint pdf $f_{XY}$:
Consider the sample $Z_n = X_n \times Y_n = \{(x_i, y_i)_{i=1}^n\}$ from joint density function $f_{X,Y}$ with marginal density functions $f_X, f_Y$.

- Divide the sample set $Z_n$ in two subsets $Z_{n'}'$ and $Z_{n''}'$ with the sample proportion $\alpha = n'/n$.
- Denote $\tilde{Z}_{n''}' = \{(x_{i_k}, y_{j_k}), k = 1, \ldots, n''\}$, selected at random from $Z_{n''}'$.
- Construct the minimal spanning tree (MST) on the concatend data set, $Z_{n'}' \cup \tilde{Z}_{n''}'$.
- Count the edges connecting a node in $Z_{n'}'$ to a node of $\tilde{Z}_{n''}'$.
- Output: The FR estimator $\mathcal{R}_{n', n''}$.

- Convergence rate for this estimator is under progress.
A multivariate generalization $l_p$ for a $d$ RV $X = (X_1, \ldots, X_d)$ with marginal PDFs $f_i(x_i)$ and copula density $c(u)$ is given by

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

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

**Interesting relations:**

- **Pearson’s $\phi^2$-statistic:**

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

- **Renyi copula entropy, $h_\alpha(c) = -l_\alpha(X)$:**

$$1 - e^{p \, h_q(c)} \leq l_p(c) \leq \frac{p(e^{-h_2(c)} - 1)}{pe^{-h_2(c)} + q}.$$
How can copula be used to estimate $I_p$:

**Theorem**

Consider sample realizations from $c(u)$ and uniform copula density $c_0(u)$ denoted by $U \in [0, 1]^{n \times d}$ and $U_0 \in [0, 1]^{m \times d}$. The FR statistics, $R_{m,n}^c$, is constructed on the data set $U \cup U_0$ estimates $I_p(X)$ as $n \to \infty$ and $m \to \infty$ in the regime $\frac{n}{n+m} \to p$ and $\frac{n}{n+m} \to q$, that is

$$1 - R_{m,n}^c(U, U_0) \frac{n + m}{2nm} \to I_p(c) = I_p(X).$$

