Estimation of density ratio and its application to design a measure of dependence

Sohan Seth and José C. Príncipe

Computational NeuroEngineering Lab, University of Florida, Gainesville

Introduction

A measure of dependence is an important concept that has recently received considerable attention in the area of machine learning due to its potential application in many practical problems such as regression, clustering, feature selection, independent component analysis, etc. Given random variable pair (X, Y), a possible measure of dependence $\delta(X, Y)$ is the divergence between the joint density $h(x, y)$ and the product of the marginal densities $f(x)$ and $g(y)$ i.e.

$$\delta(X,Y) = \int \phi\left(\frac{h(x,y)}{f(x)g(y)}\right) f(x)g(y)\mathrm{d}x\mathrm{d}y = \mathbb{E}_X\mathbb{E}_Y\phi\left(\frac{h(X,Y)}{f(X)g(Y)}\right)$$

where ϕ is a convex function. Therefore, dependence between two random variables can be estimated by estimating the ratio

$$l(x,y) = \frac{h(x,y)}{f(x)g(y)}.$$

Estimation of density ratio

- Let U and V be two random variables with densities $p(u)$ and $q(v)$, respectively. The objective is to estimate the density ratio

$$l(u) = \frac{q(u)}{p(u)}$$

from realizations $\{u_i\}_{i=1}^n$ and $\{v_i\}_{i=1}^n$.

- Assume that $l(u) \in \mathcal{H}$ where $\mathcal{H}$ is the reproducing kernel Hilbert space defined on $\mathcal{U}$ and U takes value in $\mathcal{U}$ i.e., to be specific, assume the model

$$\hat{l}(u) = \sum_{i=1}^n \alpha_i \kappa(u, u_i)$$

where κ is the reproducing kernel of $\mathcal{H}$.

- Minimize the regularized cost,

$$\mathcal{J} = \int \left(l(u) - \hat{l}(u)\right)^2 p(u)\mathrm{d}u + \lambda||\hat{l}||^2_{\mathcal{H}}$$

- Expanding the cost function, we get,

$$\begin{aligned}
\mathcal{J} =& \int \left(l(u) - \hat{l}(u)\right)^2 p(u)\mathrm{d}u + \lambda||\hat{l}||^2_{\mathcal{H}} \\
=& C - 2\int \hat{l}(u)q(u)\mathrm{d}u + \int \hat{l}^2(u)p(u) + \lambda||\hat{l}||^2_{\mathcal{H}} \\
=& C - 2\,\mathbb{E}\sum_{i=1}^n \alpha_i\kappa(V,u_i) + \mathbb{E}\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j\kappa(U,u_i)\kappa(U,u_j) + \lambda\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j\kappa(u_i,u_j) \\
\approx& C - \frac{2}{n}\sum_{j=1}^n\sum_{i=1}^n \alpha_i\kappa(v_j,u_i) + \frac{1}{n}\sum_{k=1}^n\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j\kappa(u_k,u_i)\kappa(u_k,u_j) \\
&\qquad + \lambda\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j\kappa(u_i,u_j) \\
=& C - \frac{2}{n}\alpha^\top\mathbf{K}_{UV}\mathbf{1} + \frac{1}{n}\alpha^\top(\mathbf{K}_{UU} + n\lambda\mathbf{I})\mathbf{K}_{UU}\alpha
\end{aligned}$$

where $C = \int l^2(u)p(u)\mathrm{d}u < \infty$ does not depend on α, $\mathbf{1}$ is a vector of ones and

$$[\mathbf{K}_{AB}]_{ij} = \kappa(a_i, b_j).$$

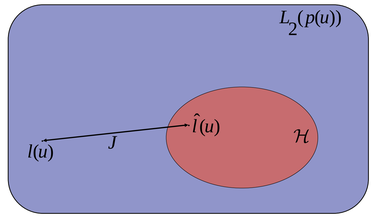

Figure 1: Graphical representation of the cost for density ratio estimation

- Differentiating with respect to α and equating to zero, we get,

$$(\mathbf{K}_{UU} + n\lambda\mathbf{I})\mathbf{K}_{UU}\alpha^* = \mathbf{K}_{UV}\mathbf{1}.$$

- Therefore, the values of $\hat{l}(u)$ at points $\{u_i\}_{i=1}^n$ are given by

$$\hat{\mathbf{l}} = \mathbf{K}_{UU}\alpha^* = (\mathbf{K}_{UU} + n\lambda\mathbf{I})^{-1}\mathbf{K}_{UV}$$

where $\hat{\mathbf{l}} = [\hat{l}(u_1), \ldots, \hat{l}(u_n)]^\top$.

Estimation of Dependence

- In order to estimate dependence, we need to estimate the ratio $l(x, y)$. However, the extension of the proposed method to this problem is not straightforward since, the model $\hat{l}(x, y)$ ($\hat{l}(u)$ in previous section) requires samples from the denominator density $f(x)g(y)$ (i.e. $p(u)$ in the previous section) whereas, in practice, we only have samples from the numerator density $h(x, y)$ (i.e. $q(u)$ in previous section). To resolve this issue we propose to estimate the ratio

$$\tilde{l}(x,y) = \frac{f(x)g(y)}{h(x,y)}$$

instead. Note that this ratio can be undefined but we are only interested in the region where it is defined. For example, we use this fact in defining the cost function.

- In a similar fashion, given samples $\{(x_i, y_i)\}_{i=1}^n$, assume the model $\hat{l}(x,y) = \sum_{i=1}^n \alpha_i\kappa_1(x,x_i)\kappa_2(y,y_i)$ and minimize the regularized cost

$$\mathcal{J} = \int \left(\hat{l}(x,y) - \hat{\tilde{l}}(x,y)\right)^2 h(x,y)\mathrm{d}x\mathrm{d}y + \lambda||\hat{\tilde{l}}(x,y)||^2_{\mathcal{H}}$$

- Using similar derivation, we get,

$$n(\mathbf{K}_{XX}\circ\mathbf{K}_{YY} + n\lambda\mathbf{I})(\mathbf{K}_{XX}\circ\mathbf{K}_{YY})\alpha^* = \mathbf{K}_{XX}\mathbf{1}\circ\mathbf{K}_{YY}\mathbf{1}$$

where $\circ$ denotes the entrywise product.

- Using the estimated density ratio we estimate the dependence, in particular, the mutual information ($\phi(t) = t\log t$), as follows,

$$\begin{aligned}
\mathcal{MI}(X,Y) =& -\mathbb{E}\log(\tilde{l}(X,Y)) \approx -\frac{1}{n}\sum_{j=1}^n \log\hat{\tilde{l}}(x_i,y_i) = -\frac{1}{n}\mathbf{1}^\top\log\hat{\tilde{\mathbf{l}}} \\
=& -\frac{1}{n}\mathbf{1}^\top\log((\mathbf{K}_{XX}\circ\mathbf{K}_{YY} + n\lambda\mathbf{I})^{-1}(\mathbf{K}_{XX}\mathbf{1}\circ\mathbf{K}_{YY}\mathbf{1}))
\end{aligned}$$

- Although the proposed method requires inverting an $n \times n$ matrix, the computational load can be reduced by exploiting the fact that this matrix often has a fast decaying eigen structure. Methods such as incomplete Cholesky decomposition can be used for this purpose [2].

Simulation

- **Density estimation:**

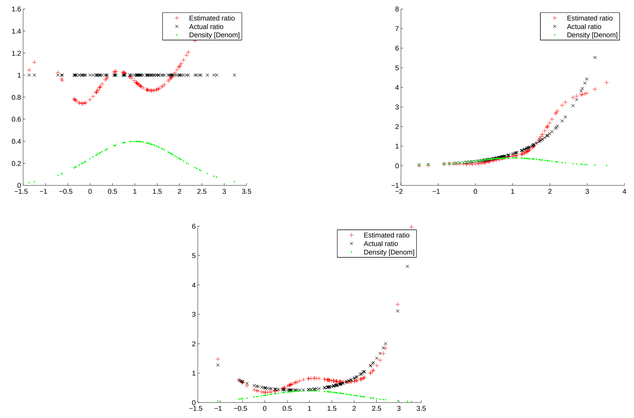

Figure 2: Estimation of density ratio with 100 samples where the actual densities are both Gaussian with following parameters (a) $\mu_1 = 1, \sigma_1 = 1, \mu_2 = 1, \sigma_2 = 1$, (b) $\mu_1 = 1, \sigma_1 = 2, \mu_2 = 1, \sigma_2 = 1$ and (c) $\mu_1 = 1, \sigma_2 = 1, \mu_2 = 1, \sigma_2 = 2$

- **Test of dependence:** Let $X', Y' \sim \mathcal{N}(0,1)$ and $U \sim \mathcal{U}(0,2)$. Define $X = UX'$ and $Y = UY'$. Then, X and Y are not independent. However, it is difficult to see from the scatterplot.

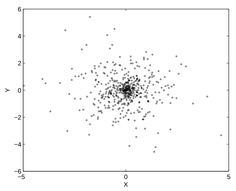

Figure 3: The variables shown in this figure are dependent.

Method || Sample size	100	200	300	400	500
Mutual information estimated by proposed method	72	87	94	98	99
Characteristic function based measure of dependence [1]	19	45	79	87	97

Table 1: The table shows the number of times dependence has been accepted out of 100 times. The threshold of the test has been determined by a permutation test. The proposed method shows better small sample performance.

Summary

We present a novel way of estimating the ratio of two PDFs and use it to estimate the dependence between two random variables. The initial results are promising. Further study of the effects of kernel size and regularization parameter is being undertaken.

References

[1] Andrey Feuerverger. A consistent test for bivariate dependence. *International Statistical Review*, 61-3:419–433, 1993.

[2] Shai Fine, Katya Scheinberg, Nello Cristianini, John Shawe-taylor, and Bob Williamson. Efficient svm training using low-rank kernel representations. *Journal of Machine Learning Research*, 2:243–264, 2001.