A CONDITIONAL INDEPENDENCE PERSPECTIVE OF VARIABLE SELECTION

Sohan Seth and José C. Príncipe
Computational NeuroEngineering Lab, University of Florida, Gainesville

Introduction

- Context : Variable selection seeks to find a smaller subset of informative variables ($\mathcal{S}_m$) from the set of all input variables ($\mathcal{S} \supset \mathcal{S}_m$). This problem can be easily phrased in terms of conditional independence as follows: given a set of d variables $\mathcal{S}$, select a set of m important variables $\mathcal{S}_m$, such that given these variables the target Y is least dependent on the rest of the variables $\mathcal{S} \backslash \mathcal{S}_m$ [1]. In practice, however, this approach is avoided due to the inherent difficulty in assessing conditional independence, and a correlation or mutual information based approach is preferred [3]. We introduce a novel measure of conditional independence, discuss its properties, and apply it in the context of variable selection.
- Background : (X, Y) are said to be conditionally independent (CI) given Z i.e. $X \perp Y | Z$ if Y does **not** contain any additional information about X other than that contained in Z i.e. $F_{X|YZ}(x|y,z) = F_{X|Z}(x|z) \, \forall \, (x,y,z)$ where $F_{U|V} = P(U \leq u | V \leq v)$. The simplest measure of conditional independence is [2]

$$\mathcal{M}_2^2 = \int \left(F_{XYZ}(x,y,z) F_Z(z) - F_{XZ}(x,z) F_{YZ}(y,z) \right)^2 \mathrm{d}F_{XYZ}(x,y,z).$$

This statistic is estimated by replacing the joint distributions with their respective empirical estimates i.e. $\hat{F}_U(u) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(u \leq u_i)$. This approach, however, suffers from estimation issues in higher dimensions, especially when the sample size is low. We explore an alternate method that alleviates these drawbacks.

Measure of conditional independence (CI)

X, Y, and Z are continuous

- **Conditional independence :=**

$$X \perp Y | Z \iff P(X \leq x | Y = y, Z = z) = P(X \leq x | Z = z) \, \forall \, (x,y,z).$$

- **Measure** of CI := a statistic that attains 0 if and only CI is satisfied e.g,

$$\mathcal{M}_1^2 = \int (P(X \leq u | Y = v, Z = w) - P(X \leq u | Z = w))^2 \mathrm{d}F_X(u) \mathrm{d}F_{YZ}(v,w).$$

- **Estimator**

$$\hat{\mathcal{M}}_1^2 = \int (\hat{P}(X \leq u | Y = v, Z = w) - \hat{P}(X \leq u | Z = w))^2 \mathrm{d}\hat{F}_X(u) \mathrm{d}\hat{F}_{YZ}(v,w)$$

where $\hat{\cdot}$ denotes estimates.

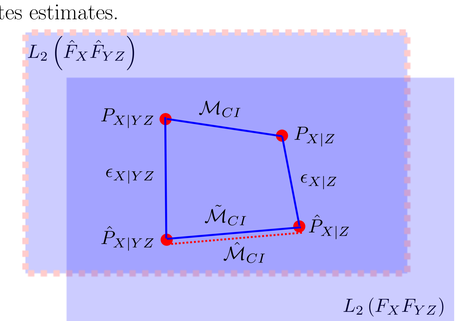

FIGURE 1: Illustration of $\mathcal{M}_1$ ($\mathcal{M}_{CI}$ in Figure) and its estimator

$$\Rightarrow |\mathcal{M}_1 - \hat{\mathcal{M}}_1| < \epsilon_{X|Z} + \epsilon_{X|YZ} + \underbrace{|\hat{\mathcal{M}}_1 - \tilde{\mathcal{M}}_1|}_{\to 0 \text{ as } n \to \infty}$$

- **Problem** Find estimator $p_u(v)$ of $P(U \leq u | V = v)$ s.t.

$$\epsilon_{U|V}^2 = \int (P(U \leq u | V = v) - p_u(v))^2 \mathrm{d}F_U(u) \mathrm{d}F_V(v)$$

is minimized.

- **Solution**

1 $F_U(u)$ is non-negative, therefore,

$$\text{minimize } \epsilon_{U|V}^2 \Rightarrow \text{ minimize } \epsilon_{U|V}^2(u) = \int (P(U \leq u | V = v) - p_u(v))^2 \mathrm{d}F_V(v) \, \forall \, u.$$

2 $P(U \leq u) = \mathbf{E}\mathbb{I}(U \leq u) \Rightarrow$

$$\begin{aligned} \epsilon_{U|V}^2(u) =& C - 2 \int P(U \leq u | V = v) p_u(v) \mathrm{d}F_V(v) + \int p_u^2(v) \mathrm{d}F_V(v) \\ =& C - 2 \int \mathbb{I}(u' \leq u) \mathrm{d}F_{U|V}(u'|v) p_u(v) \mathrm{d}F_V(v) + \int p_u^2(v) \mathrm{d}F_V(v) \\ =& C - 2 \int \mathbb{I}(u' \leq u) p_u(v) \mathrm{d}F_{UV}(u', v) + \int p_u^2(v) \mathrm{d}F_V(v). \end{aligned}$$

where $C = \int P^2(U \leq u | V = v) \mathrm{d}F_V(v)$.

3 Assume $p_u(v) = \sum_{i=1}^m \alpha_i^u \phi_i(v)$, then,

$$\begin{aligned} \epsilon_{U|V}^2(u) =& C - 2\mathbf{E} \sum_{j=1}^m \alpha_j^u \mathbb{I}(U \leq u) \phi_j(V) + \mathbf{E} \sum_{i=1}^m \sum_{j=1}^m \alpha_i^u \alpha_j^u \phi_i(V) \phi_j(V) \\ =& C - 2\mathbf{b}^\top \alpha_u + \alpha_u^\top \mathbf{A} \alpha_u \end{aligned}$$

where $[\mathbf{b}]_j = \mathbf{E}[\mathbb{I}(U \leq u) \phi_j(V)]$ and $[\alpha_u]_i = \alpha_i^u$ are column vectors, and $[A]_{ij} = \mathbf{E}[\phi_i(V) \phi_j(V)]$ is a matrix.

4 Given realizations $\{(u_i, v_i)\}_{i=1}^n$, $\hat{\mathbf{b}} = \frac{1}{n} \Phi \mathbf{i}$ and $\hat{\mathbf{A}} = \frac{1}{n} \Phi \Phi^\top$, where $[\Phi]_{ij} = \phi_j(v_i)$ is a matrix and $[\mathbf{i}_u]_i = \mathbb{I}(u_i \leq u)$ is a column vector.

5 Regularized solution; $\hat{P}(U \leq u | V = v) = \sum_{i=1}^n \alpha_i^{*(u)} \phi_i(v)$ where

$$\alpha_u^* = (\Phi^\top \Phi + \lambda \mathbf{I})^{-1} \Phi^\top \mathbf{i}_u.$$

- **Estimator**

$$\begin{aligned} \Rightarrow \hat{\mathcal{M}}_1^2 =& \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n (\hat{P}(X < x_i | Y = y_j, Z = z_j) - \hat{P}(X < x_i | z_j))^2 \\ =& \frac{1}{n^2} || (\Xi (\Xi^\top \Xi + \lambda \mathbf{I})^{-1} \Xi^\top - \Psi (\Psi^\top \Psi + \lambda \mathbf{I})^{-1} \Psi^\top) \mathbf{i} ||_F^2. \end{aligned}$$

where $[\mathbf{i}]_{ij} = \mathbb{I}(x_i \leq x_j)$ is a matrix of 0s and 1s, and Ξ and Ψ are matrix generated by basis functions ξ and ψ, chosen to be Gaussian kernel.

X **is categorical** (such as class labels)

- **Conditional independence :=**

$$X \perp Y | Z \iff P(X = x | Y = y, Z = z) = P(X = x | Z = z) \, \forall \, (x,y,z).$$

- **Measure**

$$\mathcal{M}_1^2 = \sum_{i=1}^c \int \left(P(X = u_i | Y = v, Z = w) - P(X = u_i | Z = w) \right)^2 \mathrm{d}F_{YZ}(v,w).$$

- **Estimator**

$$\hat{\mathcal{M}}_1^2 = \frac{1}{n^2} || (\Xi (\Xi^\top \Xi + \lambda \mathbf{I})^{-1} \Xi^\top - \Psi (\Psi^\top \Psi + \lambda \mathbf{I})^{-1} \Psi^\top) \mathbf{i} ||_F^2.$$

where $[\mathbf{i}]_{ij} = \mathbb{I}(x_i = u_j)$. This is a $(n \times c)$ matrix rather than $(n \times n)$ as in continuous X.

Simulation

- **Comparison of $\mathcal{M}_1$ and $\mathcal{M}_2$: Regression** $X \sim U[0,1]^{16}, \epsilon \sim \mathcal{N}(0, 0.01^2)$

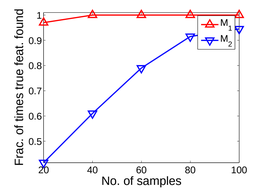

FIGURE 2: $y = x_1 \exp(-x_1^2 - x_2^2) + \epsilon$

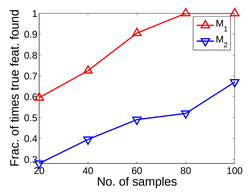

FIGURE 3: $y = 0.9 x_1 + \frac{0.2}{1 + x_2} + \epsilon$

- **Comparison of $\mathcal{M}_1$ and $\mathcal{M}_2$: Classification** $X_3, \ldots, X_{16} \sim U[0,1]$

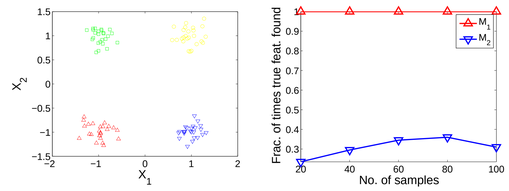

FIGURE 4: Four class problem

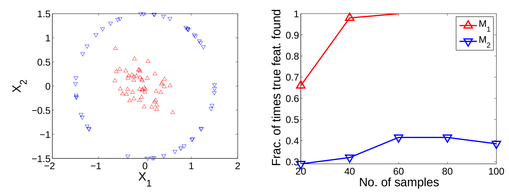

FIGURE 5: Two class problem

- **Real world data**
  - Forward selection Select $S_m = \underset{S \in \mathcal{S} \backslash \mathcal{S}_{m-1}}{\operatorname{argmin}} \, \mathcal{M}(Y, \mathcal{S} \backslash (\mathcal{S}_{m-1} \cup S_m), \mathcal{S}_{m-1} \cup S_m)$
  - Backward elimination Eliminate $S_m = \underset{S \in \mathcal{S} \backslash \mathcal{S}_{m-1}}{\operatorname{argmin}} \, \mathcal{M}(Y, S, \mathcal{S}_m \backslash S)$

Dataset	$F_{\mathcal{M}_2}$	$F_{\mathcal{M}_1}$	$B_{\mathcal{M}_2}$	$B_{\mathcal{M}_1}$	mRMR[3]
Glass	3.74	4.67	5.61	5.61	**2.80**
Wine	3.37	**1.12**	5.62	2.25	2.25
Parkinsons	4.08	4.08	**3.06**	4.08	5.10
Breast Cancer	10.00	**8.95**	11.58	9.47	11.58
Ionosphere	**3.31**	4.64	4.64	4.64	**3.31**
Connec. bench	12.50	13.46	14.42	**10.58**	11.54
Mean	6.17	6.15	7.49	6.10	6.10

TABLE 1: Percentage of misclassified samples using a kNN classifier

References

[1] Daphne Koller and Mehran Sahami. Toward optimal feature selection. Technical Report 1996-77, Stanford InfoLab, February 1996. Previous number = SIDL-WP-1996-0032.

[2] O. Linton and P. Gozalo. Conditional independence restrictions: Testing and estimation. Cowles Foundation Discussion Papers 1140, Cowles Foundation, Yale University, November 1996.

[3] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell*, 27(8):1226–1238, August 2005.