Lecture 13.

HGR (Rényi) Maximal Correlation.

Given 2 discrete random variables X, Y, want to measure the correlation between X and Y, how can we do?

Example: Pearson correlation coefficient:

$$\gamma(X,Y) \triangleq \frac{\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]}{\sqrt{\mathrm{Var}(X)\cdot\mathrm{Var}(Y)}}$$

- If $\gamma(X,Y)=0 \not\Rightarrow X, Y$ are indep.

A good correlation measurement $\rho(X,Y)$ should satisfy:

(1) commutable $\rho(X,Y)=\rho(Y,X)$

(2) $0 \le \rho(X,Y) \le 1$

(3) $\rho(X,Y)=0$ if and only if X, Y are indep.

(4) $\rho(X,Y)=1$, if $Y=\xi(X)$ or $X=\eta(Y)$, for some deterministic functions ξ, η.

(5) For one-to-one functions ξ, η, $\rho(\xi(X),\eta(Y))=\rho(X,Y)$

Example: Pearson correlation coefficient $\gamma(X,Y)$

$$\gamma(X,Y) \triangleq \frac{\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]}{\sqrt{\mathrm{Var}(X)\cdot\mathrm{Var}(Y)}} \qquad -1 \le \gamma(X,Y) \le 1$$

$|\gamma(X,Y)|$ satisfies (1), (2), not (3), (4), (5) (HW)

$|\gamma(X,Y)|=1$, iff $X=aY+b$ or $Y=cX+d$

Example: mutual information $I(X;Y) \triangleq \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}$

(P_{XY}: joint dist. of X, Y)

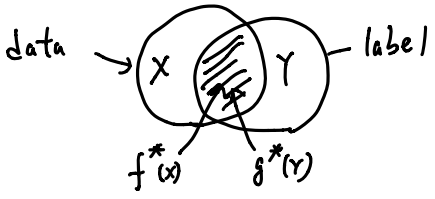
$I(X;Y)$ satisfies (1), (3), (5) not (2), (4)

- Can we find some $\rho(X,Y)$ satisfying (1) - (5) ?

Definition: The HGR maximal correlation $\rho(X,Y)$ is defined as

$$\rho(X,Y) \triangleq \max_{\substack{f: X\mapsto\mathbb{R}, g: Y\mapsto\mathbb{R} \\ \mathbb{E}[f(X)]=\mathbb{E}[g(Y)]=0 \\ \mathbb{E}[f^2(X)]=\mathbb{E}[g^2(Y)]=1}} \mathbb{E}[f(X)\cdot g(Y)]$$

- $\rho(X,Y) = \max_{f,g} \gamma(f(X), g(Y))$
- Data X, Y, want find features of X and Y such that these features are the most related ⇒ HGR maximal correlation functions



$\rho(X,Y)$ satisfies (1), (2), (5), (4), (3)

Check (5): for one-to-one functions ξ, η, $\rho(\xi(X), \eta(Y)) = \max_{f,g} \gamma(\underbrace{f(\xi(X))}_{f'}, \underbrace{g(\eta(Y))}_{g'})$

$= \max_{f',g'} \gamma(f'(X), g'(Y)) = \rho(X,Y)$

Check (4): If $Y = \xi(X)$, let $K(Y): \mathcal{Y} \mapsto \mathbb{R}$ be a one-to-one function such that $\mathbb{E}[K(Y)]=0$, $\mathbb{E}[K^2(Y)]=1$, then take $f(X) = K(\xi(X))$, $g(Y) = K(Y)$

$\Rightarrow \underbrace{\mathbb{E}[f(X)\cdot g(Y)]}_{\le 1} = \mathbb{E}[K(\underbrace{\xi(X)}_{Y})\cdot K(Y)] = \mathbb{E}[K^2(Y)] = 1 \Rightarrow \rho(X,Y) = 1$

Definition: The canonical dependence matrix $\widetilde{B}$ ($|\mathcal{Y}| \times |\mathcal{X}|$) is defined as

$$\widetilde{B}(y,x) \triangleq \frac{P_{XY}(x,y) - P_X(x)P_Y(y)}{\sqrt{P_X(x)\cdot P_Y(y)}}, \quad \forall x \in \mathcal{X},\ y \in \mathcal{Y}$$

Example: X, Y are binary, $P_{XY}(x,y) = \begin{cases} \frac{1}{3} & \text{if } (x,y)=(0,0) \\ \frac{1}{6} & \text{“ } (0,1) \\ \frac{1}{6} & (1,0) \\ \frac{1}{3} & (1,1) \end{cases}$

$Y \backslash X$	0	1	
0	$\frac{1}{3}$	$\frac{1}{6}$	$P_Y(0)=\frac{1}{2}$
1	$\frac{1}{6}$	$\frac{1}{3}$	$P_Y(1)=\frac{1}{2}$
	$P_X(0)=\frac{1}{2}$	$P_X(1)=\frac{1}{2}$	

$\widetilde{B}$ is a 2×2 matrix

$\widetilde{B}$:

	0	1
0	$\widetilde{B}(0,0) = \frac{\frac{1}{3} - \frac{1}{2}\cdot\frac{1}{2}}{\sqrt{\frac{1}{2}\cdot\frac{1}{2}}} = \frac{\frac{1}{12}}{\frac{1}{2}} = \frac{1}{6}$	$\widetilde{B}(0,1) = \frac{\frac{1}{6} - \frac{1}{2}\cdot\frac{1}{2}}{\sqrt{\frac{1}{2}\cdot\frac{1}{2}}} = \frac{-\frac{1}{12}}{\frac{1}{2}} = -\frac{1}{6}$
1	$\widetilde{B}(1,0) = -\frac{1}{6}$	$\widetilde{B}(1,1) = \frac{1}{6}$

$$\widetilde{B} = \begin{bmatrix} \frac{1}{6} & -\frac{1}{6} \\ -\frac{1}{6} & \frac{1}{6} \end{bmatrix}$$

Definition: The information vectors ϕ, ψ associated to functions f, g are defined:

$$\phi(x) = \sqrt{P_X(x)}\cdot f(x), \quad \psi(y) = \sqrt{P_Y(y)}\cdot g(y)$$

(ϕ: $|\mathcal{X}|$-dimensional vector; ψ: $|\mathcal{Y}|$-dim vector)

$$\phi = \begin{bmatrix} \phi(1) \\ \phi(2) \\ \vdots \\ \phi(|\mathcal{X}|) \end{bmatrix} \qquad \psi = \begin{bmatrix} \psi(1) \\ \psi(2) \\ \vdots \\ \psi(|\mathcal{Y}|) \end{bmatrix}$$

Whenever functions f, g are given, we have the corresponding ϕ, ψ.

- property: norm-square: $\|\phi\|^2 = \sum_x \phi^2(x) = \sum_x P_X(x) f^2(x) = \mathbb{E}[f^2(X)]$, $\|\psi\|^2 = \mathbb{E}[g^2(Y)]$

inner product: $\underline{\sqrt{P_X}} = \begin{bmatrix} \sqrt{P_X(1)} \\ \sqrt{P_X(2)} \\ \vdots \\ \sqrt{P_X(|\mathcal{X}|)} \end{bmatrix}$, $\langle \phi, \underline{\sqrt{P_X}} \rangle = \sum_x \phi(x)\cdot\sqrt{P_X(x)} = \sum_x P_X(x)\cdot f(x) = \mathbb{E}[f(X)]$

$$\underline{\sqrt{P_Y}} = \begin{bmatrix} \sqrt{P_Y(1)} \\ \vdots \\ \sqrt{P_Y(|\mathcal{Y}|)} \end{bmatrix}, \quad \langle \psi, \underline{\sqrt{P_Y}} \rangle = \mathbb{E}[g(Y)]$$

Thm: The HGR maximal correlation $\rho(X,Y)$ is the largest singular value of $\widetilde{B}$

Proof: Note that $\mathbb{E}[f(X)\cdot g(Y)] = \sum_{x,y} P_{XY}(x,y)\cdot f(x)\cdot g(y) = \sum_{x,y}\left(\frac{P_{XY}(x,y)}{\sqrt{P_X(x)P_Y(y)}}\cdot(\sqrt{P_X(x)}\cdot f(x))\cdot(\sqrt{P_Y(y)}\cdot g(y)) - \frac{P_X(x)P_Y(y)}{\sqrt{P_X(x)P_Y(y)}}\cdot(\sqrt{P_X(x)}f(x))(\sqrt{P_Y(y)}\cdot g(y))\right)$

$$\sum_{x,y}\frac{P_X(x)\cdot P_Y(y)}{\sqrt{P_X(x)P_Y(y)}}\cdot(\sqrt{P_X(x)}\,f(x))\cdot(\sqrt{P_Y(y)}\,g(y)) = \sum_{x,y} P_X(x)f(x)\cdot P_Y(y)\cdot g(y)$$

$= \mathbb{E}[f(x)] \cdot \mathbb{E}[g(Y)]$

$\Rightarrow \mathbb{E}[f(x)\cdot g(Y)] = \sum_{x,y} \underbrace{\frac{P_{XY}(x,y) - P_X(x)\cdot P_Y(y)}{\sqrt{P_X(x)\,P_Y(y)}}}_{\tilde{B}(y,x)} \cdot \underbrace{\left(\sqrt{P_X(x)}\, f(x)\right)}_{\phi(x)} \cdot \underbrace{\left(\sqrt{P_Y(y)}\, g(y)\right)}_{\psi(y)} = \sum_{x,y} \tilde{B}(y,x)\cdot\phi(x)\cdot\psi(y) = \psi^T \tilde{B}\, \phi$

$\Rightarrow \rho(X,Y) = \max_{\phi,\psi} \psi^T \tilde{B}\, \phi$ subject to: $\|\phi\|^2 = \|\psi\|^2 = 1$ $(\Leftrightarrow \mathbb{E}[f^2(x)] = \mathbb{E}[g^2(Y)] = 1)$

$\langle \phi, \sqrt{P_X} \rangle = \langle \psi, \sqrt{P_Y} \rangle = 0$

$(\Leftrightarrow \mathbb{E}[f(x)] = \mathbb{E}[g(Y)] = 0)$

Fact 1: $\underset{\|\phi\|^2=\|\psi\|^2=1}{\arg\max}\ \psi^T \tilde{B}\, \phi$ = largest right / left singular vectors of $\tilde{B}$

Fact 2: singular vector are orthogonal with each other.

Check: $\left[\tilde{B}\cdot\sqrt{P_X}\right](y) = \sum_x \tilde{B}(y,x)\cdot\sqrt{P_X(x)} = \sum_x \frac{P_{XY}(x,y) - P_X(x)\cdot P_Y(y)}{\sqrt{P_Y(y)}} = \frac{P_Y(y) - P_Y(y)}{\sqrt{P_Y(y)}} = 0$

$|\mathcal{Y}| \times |\mathcal{X}| \times 1$ $\quad$ $|\mathcal{X}| \times 1$

$\Rightarrow \sqrt{P_X}$ is a right singular vector of $\tilde{B}$ with singular value 0.

since $\tilde{B}\cdot\sqrt{P_X} = \underline{0}$

$\sqrt{P_Y}$ is also a left singular vector of $\tilde{B}$

$\Rightarrow \rho(X,Y) = \sigma_1$, with "=" iff ϕ and ψ are the largest right/left singular vectors of $\tilde{B}$

σ_1: the largest singular value of $\tilde{B}$

$\Rightarrow$ Let ϕ_1, ψ_1 be the largest right / left singular vectors of $\tilde{B}$, then

$f^*(x) = \frac{1}{\sqrt{P_X(x)}}\cdot\phi_1(x)$, $\quad g^*(y) = \frac{1}{\sqrt{P_Y(y)}}\cdot\psi_1(y)$

- $\text{svd}(\tilde{B}) = [\sigma_1, \phi_1, \psi_1]$

  σ_1: $\rho(X,Y)$; ϕ_1: $f(x)$; ψ_1: $g(y)$

Check (3): $\rho(X,Y) = 0$ iff $\sigma_1 = 0$ iff $\tilde{B} = 0$ iff $P_{XY}(x,y) = P_X(x)\cdot P_Y(y)$

iff X, Y are indep.

Power iteration / method

For a postive definite matrix M, eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_m$

eigenvectors $v_1, v_2, \ldots, v_m$

$$\Rightarrow M = \sum_{i=1}^{m} \lambda_i \cdot v_i \cdot v_i^T \quad : \text{eigen-decomposition.}$$

$$\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_m \end{bmatrix}, \quad V = [v_1 \cdots v_m] \Rightarrow M = V \Lambda \cdot V^T$$

The algorithm to compute v_1

(i) Choose $u^{(0)} \not\perp v_1$

(ii) compute $u^{(1)} = M \cdot u^{(0)}$ (power step), $u^{(1)} \leftarrow \frac{u^{(1)}}{\|u^{(1)}\|}$ (normalization)

$$u^{(2)} = M \cdot u^{(1)}$$

$$\vdots$$

$$u^{(t)} = M \cdot u^{(t-1)}, \quad \text{normalization}$$

$$\Rightarrow u^{(t)} = \frac{M^t \cdot u^{(0)}}{\|M^t \cdot u^{(0)}\|}$$

Thm: $u^{(t)} \xrightarrow{t \to \infty} v_1$

Proof: Let $u^{(0)} = \sum_{i=1}^{m} \alpha_i v_i$, then $M^t \cdot u^{(0)} = \sum_{i=1}^{m} \alpha_i \cdot \underbrace{M^t \cdot v_i}_{\lambda_i^t \cdot v_i} = \sum_{i=1}^{m} \alpha_i \cdot \lambda_i^t \cdot v_i$

$$= \alpha_1 \cdot \lambda_1^t \cdot v_1 + \alpha_2 \cdot \lambda_2^t \cdot v_2 + \cdots + \alpha_m \cdot \lambda_m^t \cdot v_m$$

since λ_1 is the largest among all λ_i's

$$\Rightarrow \frac{\lambda_i^t}{\lambda_1^t} \to 0, \; t \to \infty \qquad \Rightarrow \frac{M^t \cdot u^{(0)}}{\|M^t \cdot u^{(0)}\|} \to v_1, \; t \to \infty$$

largest right singular vector of $\tilde{B}$

Fact 3: ϕ_1 is also the largest eigenvector of $\tilde{B}^T \cdot \tilde{B}$

ψ_1 is " of $\tilde{B} \cdot \tilde{B}^T$

The algorithm to compute ϕ_1, ψ_1

(i) Choose $\phi^{(0)} \not\perp \phi_1$, and $\phi^{(0)} \perp \boxed{P_X}$

(ii) Iterating: $\begin{cases} \psi^{(i)} = \tilde{B} \cdot \phi^{(i)} \\ \phi^{(i+1)} = \tilde{B}^T \psi^{(i)} \end{cases}$, for $i = 0, 1, \ldots, t$

(iii) Normalization

$$\phi^{(0)} \to \psi^{(0)} = \tilde{B} \cdot \phi^{(0)} \to \phi^{(1)} = \tilde{B}^T \cdot \psi^{(0)} \to \psi^{(1)} = \tilde{B} \cdot \phi^{(1)} \to \cdots$$

For ϕ : $\phi^{(i+1)} = \tilde{B}^T \cdot \tilde{B} \cdot \phi^{(i)} \Rightarrow \phi^{(i+1)}$ converge to largest eigenvector of $\tilde{B}^T \cdot \tilde{B}$

We can write $f^{(i)}(x) = \frac{1}{\sqrt{P_X(x)}} \phi^{(i)}(x)$, $g^{(i)}(y) = \frac{1}{\sqrt{P_Y(y)}} \cdot \psi^{(i)}(y)$

Then, the updating step $\psi^{(i)} = \tilde{B} \cdot \phi^{(i)}$ can be expressed in terms of functions as:

$$g^{(i)}(y) = \frac{1}{\sqrt{P_Y(y)}} \cdot \psi^{(i)}(y) = \frac{1}{\sqrt{P_Y(y)}} \cdot \sum_x \tilde{B}(y,x) \cdot \phi^{(i)}(x) = \frac{1}{\sqrt{P_Y(y)}} \cdot \sum_x \frac{P_{XY}(x,y) - P_X(x)P_Y(y)}{\sqrt{P_X(x) \cdot P_Y(y)}} \cdot \left(\sqrt{P_X(x)}\, f^{(i)}(x)\right)$$

$$= \frac{1}{P_Y(y)} \cdot \left(\sum_x P_X(x,y) \cdot f^{(i)}(x) - \sum_x P_X(x) P_Y(y) f^{(i)}(x) \right)$$

$$= \sum_x P_{X|Y}(x|y) \cdot f^{(i)}(x) = E\left[f^{(i)}(x) \mid Y = y \right]$$

Alternative conditional Expectation (ACE)

(i) Choose zero-mean $f^{(0)}(x)$, $E[f^{(0)}(x)] = 0$

(ii) Iterating : $\begin{cases} g^{(i)}(y) = E\left[f^{(i)}(x) \mid Y = y \right] \\ f^{(i+1)}(x) = E\left[g^{(i)}(Y) \mid X = x \right] \end{cases}$, for $i = 0, \cdots, t$

(iii) Normalization.

Compute from samples : (x_1, y_1) , $\cdots\cdots$ (x_n, y_n)

for $y_1, \cdots, y_n$, $y_1 = y_2 = \cdots = y_k = y$

$$\Rightarrow E\left[f(x) \mid Y = y \right] = \frac{1}{k} \sum_{i=1}^{k} f(x_i)$$