An Introduction to Alternating Conditional Expectation (ACE) Algorithm

Xiangxiang Xu

Department of Electronic Engineering
Tsinghua University

July 15, 2018

Outline

Correlation between random variables

Pearson correlation coefficient

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

- $|\rho(X,Y)| \leq 1$

HGR correlation

$$\rho_m(X,Y) \triangleq \max_{f,g} \rho(f(X), g(Y))$$

- $|\rho(X,Y)| \leq \rho_m(X,Y) \leq 1$

HGR correlation for discrete random variables

$$\rho_m(X,Y) \triangleq \max_{f,g} \rho(f(X), g(Y)) = \max_{\substack{f:\, \mathbb{E}[f]=0, \mathrm{Var}(f)=1 \\ g:\, \mathbb{E}[g]=0, \mathrm{Var}(g)=1}} \mathbb{E}[f(X)g(Y)]$$

Discrete (and finite) alphabet

X, Y take values from $\mathcal{X} = \{1, \cdots, |\mathcal{X}|\}$ and $\mathcal{Y} = \{1, \cdots, |\mathcal{Y}|\}$, respectively.

$$\begin{aligned}\mathbb{E}[f(X)g(Y)] &= \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} P_{X,Y}(x,y) f(x) g(y) \\ &= \sum_{x=1}^{|\mathcal{X}|} \sum_{y=1}^{|\mathcal{Y}|} f(x)\sqrt{P_X(x)} \cdot \frac{P_{X,Y}(x,y)}{\sqrt{P_X(x)}\sqrt{P_Y(y)}} \cdot g(y)\sqrt{P_Y(y)}\end{aligned}$$

HGR correlation for discrete random variables

$$\rho_m(X,Y) \triangleq \max_{f,g} \rho(f(X), g(Y)) = \max_{\substack{f:\, \mathbb{E}[f]=0, \mathrm{Var}(f)=1 \\ g:\, \mathbb{E}[g]=0, \mathrm{Var}(g)=1}} \mathbb{E}[f(X)g(Y)]$$

Discrete (and finite) alphabet

X, Y take values from $\mathcal{X} = \{1, \cdots, |\mathcal{X}|\}$ and $\mathcal{Y} = \{1, \cdots, |\mathcal{Y}|\}$, respectively.

$$\begin{aligned}\mathbb{E}[f(X)g(Y)] &= \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} P_{X,Y}(x,y) f(x) g(y) \\ &= \sum_{x=1}^{|\mathcal{X}|}\sum_{y=1}^{|\mathcal{Y}|} \underbrace{f(x)\sqrt{P_X(x)}}_{\phi(x)} \cdot \underbrace{\frac{P_{X,Y}(x,y)}{\sqrt{P_X(x)}\sqrt{P_Y(y)}}}_{B(y,x)} \cdot \underbrace{g(y)\sqrt{P_Y(y)}}_{\psi(y)}\end{aligned}$$

Outline

HGR correlation for discrete random variables

Notations

- $\boldsymbol{\phi} \leftrightarrow f : \phi(x) \triangleq \sqrt{P_X(x)} f(x)$
- $\boldsymbol{\psi} \leftrightarrow g : \psi(y) \triangleq \sqrt{P_Y(y)} g(y)$
- $\boldsymbol{B} \leftrightarrow P_{X,Y}$:

$$B(y,x) \triangleq \frac{P_{Y,X}(y,x)}{\sqrt{P_Y(y)}\sqrt{P_X(x)}}$$

HGR correlation for discrete random variables

Notations

- $\boldsymbol{\phi} \leftrightarrow f : \phi(x) \triangleq \sqrt{P_X(x)} f(x)$
- $\boldsymbol{\psi} \leftrightarrow g : \psi(y) \triangleq \sqrt{P_Y(y)} g(y)$
- $\boldsymbol{B} \leftrightarrow P_{X,Y}$:

$$B(y,x) \triangleq \frac{P_{Y,X}(y,x)}{\sqrt{P_Y(y)}\sqrt{P_X(x)}}$$

$$\mathbb{E}[f] = \langle \boldsymbol{\phi}, \boldsymbol{v}_0 \rangle, \ v_0(x) = \sqrt{P_X(x)},$$

$$\mathbb{E}[g] = \langle \boldsymbol{\psi}, \boldsymbol{u}_0 \rangle, \ u_0(y) = \sqrt{P_Y(y)},$$

$$\mathbb{E}[f^2] = \|\boldsymbol{\phi}\|^2, \ \mathbb{E}[g^2] = \|\boldsymbol{\psi}\|^2,$$

$$\mathbb{E}[f(X)g(Y)] = \boldsymbol{\psi}^{\mathrm{T}} \boldsymbol{B} \boldsymbol{\phi}.$$

$$\boldsymbol{B}\boldsymbol{\phi} \leftrightarrow \mathbb{E}[f(X)|Y], \boldsymbol{B}^{\mathrm{T}}\boldsymbol{\psi} \leftrightarrow \mathbb{E}[g(Y)|X]$$

HGR correlation for discrete random variables

Notations

- $\boldsymbol{\phi} \leftrightarrow f : \phi(x) \triangleq \sqrt{P_X(x)} f(x)$
- $\boldsymbol{\psi} \leftrightarrow g : \psi(y) \triangleq \sqrt{P_Y(y)} g(y)$
- $\boldsymbol{B} \leftrightarrow P_{X,Y}$:

$$B(y,x) \triangleq \frac{P_{Y,X}(y,x)}{\sqrt{P_Y(y)}\sqrt{P_X(x)}}$$

$$\mathbb{E}[f] = \langle \boldsymbol{\phi}, \boldsymbol{v}_0 \rangle, \ v_0(x) = \sqrt{P_X(x)},$$

$$\mathbb{E}[g] = \langle \boldsymbol{\psi}, \boldsymbol{u}_0 \rangle, \ u_0(y) = \sqrt{P_Y(y)},$$

$$\mathbb{E}[f^2] = \|\boldsymbol{\phi}\|^2, \ \mathbb{E}[g^2] = \|\boldsymbol{\psi}\|^2,$$

$$\mathbb{E}[f(X)g(Y)] = \boldsymbol{\psi}^{\mathrm{T}} \boldsymbol{B} \boldsymbol{\phi}.$$

$$\boldsymbol{B}\boldsymbol{\phi} \leftrightarrow \mathbb{E}[f(X)|Y], \boldsymbol{B}^{\mathrm{T}}\boldsymbol{\psi} \leftrightarrow \mathbb{E}[g(Y)|X]$$

$$\rho_m(X,Y) = \max_{\substack{f:\, \mathbb{E}[f]=0, \mathrm{Var}(f)=1 \\ g:\, \mathbb{E}[g]=0, \mathrm{Var}(g)=1}} \mathbb{E}[f(X)g(Y)] = \max_{\substack{\boldsymbol{\phi}:\, \langle \boldsymbol{\phi}, \boldsymbol{v}_0 \rangle = 0, \|\boldsymbol{\phi}\|=1 \\ \boldsymbol{\psi}:\, \langle \boldsymbol{\psi}, \boldsymbol{u}_0 \rangle = 0, \|\boldsymbol{\psi}\|=1}} \boldsymbol{\psi}^{\mathrm{T}} \boldsymbol{B} \boldsymbol{\phi}$$

Outline

Matrix $\boldsymbol{B}$

Definition

$$\boldsymbol{B} \leftrightarrow P_{X,Y}: \quad B(y,x) \triangleq \frac{P_{Y,X}(y,x)}{\sqrt{P_Y(y)}\sqrt{P_X(x)}}$$

Properties

- $\boldsymbol{u}_0 = \boldsymbol{B}\boldsymbol{v}_0, \boldsymbol{v}_0 = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{u}_0$.
- $\|\boldsymbol{B}\|_2 = 1$.
- The SVD of $\boldsymbol{B}$: $\boldsymbol{B} = \sum_{i=0}^{K} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}$, $\sigma_0 = 1 \geq \sigma_1 \geq \cdots \geq \sigma_K$.

Matrix $\boldsymbol{B}$

Definition

$$\boldsymbol{B} \leftrightarrow P_{X,Y}: \quad B(y,x) \triangleq \frac{P_{Y,X}(y,x)}{\sqrt{P_Y(y)}\sqrt{P_X(x)}}$$

Properties

- The SVD of $\boldsymbol{B}$: $\boldsymbol{B} = \sum_{i=0}^{K} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}$, $\sigma_0 = 1 \geq \sigma_1 \geq \cdots \geq \sigma_K$.
- $\sigma_1 = \max_{\substack{\boldsymbol{\phi}:\, \langle \boldsymbol{\phi}, \boldsymbol{v}_0 \rangle = 0, \|\boldsymbol{\phi}\| = 1 \\ \boldsymbol{\psi}:\, \langle \boldsymbol{\psi}, \boldsymbol{u}_0 \rangle = 0, \|\boldsymbol{\psi}\| = 1}} \boldsymbol{\psi}^{\mathrm{T}} \boldsymbol{B} \boldsymbol{\phi}$, with $\boldsymbol{\phi}^* = \boldsymbol{v}_1, \boldsymbol{\psi}^* = \boldsymbol{u}_1$.

Matrix $\tilde{\boldsymbol{B}}$

Definition

$\tilde{\boldsymbol{B}} \triangleq \boldsymbol{B} - \boldsymbol{u}_0 \boldsymbol{v}_0^{\mathrm{T}}$

$$\max_{\substack{\boldsymbol{\phi}:\, \langle \boldsymbol{\phi}, \boldsymbol{v}_0 \rangle = 0, \|\boldsymbol{\phi}\| = 1 \\ \boldsymbol{\psi}:\, \langle \boldsymbol{\psi}, \boldsymbol{u}_0 \rangle = 0, \|\boldsymbol{\psi}\| = 1}} \boldsymbol{\psi}^{\mathrm{T}} \boldsymbol{B} \boldsymbol{\phi} = \max_{\boldsymbol{\phi}, \boldsymbol{\psi}:\, \|\boldsymbol{\phi}\| = \|\boldsymbol{\psi}\| = 1} \boldsymbol{\psi}^{\mathrm{T}} \tilde{\boldsymbol{B}} \boldsymbol{\phi}$$

Outline

Power method

$\boldsymbol{\phi} \leftarrow \boldsymbol{\phi}_0$; /* Initialization */

$\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} - \langle \boldsymbol{\phi}, \boldsymbol{v}_0 \rangle \boldsymbol{v_0}$; /* $\boldsymbol{\phi} \perp \boldsymbol{v}_0$ */

repeat

- $\boldsymbol{\psi} \leftarrow \boldsymbol{B}\boldsymbol{\phi}$;
- $\boldsymbol{\phi} \leftarrow \boldsymbol{B}^{\mathrm{T}}\boldsymbol{\psi}$;
- $\boldsymbol{\phi} \leftarrow \boldsymbol{\phi}/\|\boldsymbol{\phi}\|$;

until $\boldsymbol{\psi}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{\phi}$ *stops to increase*;

Power method and ACE(Alternating Conditional Expectation)

Power method

$\phi \leftarrow \phi_0$; /* Initialization */

$\phi \leftarrow \phi - \langle \phi, v_0 \rangle \boldsymbol{v_0}$; /* $\phi \perp v_0$ */

repeat

$\quad \psi \leftarrow \boldsymbol{B}\phi;$

$\quad \phi \leftarrow \boldsymbol{B}^{\mathrm{T}}\psi;$

$\quad \phi \leftarrow \phi / \|\phi\|;$

until $\psi^{\mathrm{T}}\boldsymbol{B}\phi$ *stops to increase*;

ACE

$g \leftarrow g_0$; /* Initialization */

$g(y) \leftarrow g(y) - \mathbb{E}[g(Y)]$; /* Center */

repeat

$\quad f(X) \leftarrow \mathbb{E}[g(Y)|X];$

$\quad g(Y) \leftarrow \mathbb{E}[f(X)|Y];$

$\quad g(Y) \leftarrow g(Y)/\sqrt{\mathbb{E}[g^2(Y)]};$

until $\mathbb{E}[f(X)g(Y)]$ *stops to increase*;

ACE for multiple modes

Orthogonalization

Outline

Eckart-Young-Mirsky theorem

Suppose $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$, then $\boldsymbol{A}_r = \boldsymbol{U}_r\boldsymbol{\Sigma}_r\boldsymbol{V}_r^{\mathrm{T}}$ is the optimal solution to the following low rank approximation problem:

$$\begin{aligned} \underset{\hat{\boldsymbol{A}}}{\text{minimize}} \quad & \|\boldsymbol{A} - \hat{\boldsymbol{A}}\|_{\mathrm{F}}^2 \\ \text{subject to} \quad & \operatorname{rank}(\hat{\boldsymbol{A}}) \leq r. \end{aligned}$$

Corollary

The solutions of

$$\underset{(\boldsymbol{\phi},\boldsymbol{\psi})}{\text{minimize}} \|\tilde{\boldsymbol{B}} - \boldsymbol{\psi}\boldsymbol{\phi}^{\mathrm{T}}\|_{\mathrm{F}}^2$$

are $\{(c\boldsymbol{v}_1, \frac{\sigma_1}{c}\boldsymbol{u}_1) : c \neq 0\}$.

SVD & low rank approximation

Corollary

The solutions of

$$\underset{(\boldsymbol{\phi},\boldsymbol{\psi}):\boldsymbol{\phi}\perp\boldsymbol{v}_0,\boldsymbol{\psi}\perp\boldsymbol{u}_0}{\text{minimize}} \|\boldsymbol{B} - \boldsymbol{\psi}\boldsymbol{\phi}^{\mathrm{T}}\|_{\mathrm{F}}^2$$

are $\{(c\boldsymbol{v}_1, \frac{\sigma_1}{c}\boldsymbol{u}_1) : c \neq 0\}$.

SVD & low rank approximation

Corollary

The solutions of

$$\underset{(\boldsymbol{\phi},\boldsymbol{\psi}):\boldsymbol{\phi}\perp\boldsymbol{v}_0,\boldsymbol{\psi}\perp\boldsymbol{u}_0}{\text{minimize}} \|\boldsymbol{B}-\boldsymbol{\psi}\boldsymbol{\phi}^{\mathrm{T}}\|_{\mathrm{F}}^2$$

are $\{(c\boldsymbol{v}_1, \frac{\sigma_1}{c}\boldsymbol{u}_1) : c \neq 0\}$.

$$\frac{1}{2}\|\boldsymbol{B}-\boldsymbol{\psi}\boldsymbol{\phi}^{\mathrm{T}}\|_{\mathrm{F}}^2 = \frac{1}{2}\|\boldsymbol{B}\|_{\mathrm{F}}^2 - \left[\boldsymbol{\psi}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{\phi} - \frac{1}{2}\cdot(\boldsymbol{\phi}^{\mathrm{T}}\boldsymbol{\phi})\cdot(\boldsymbol{\psi}^{\mathrm{T}}\boldsymbol{\psi})\right]$$

SVD & low rank approximation

Corollary

The solutions of

$$\underset{(\boldsymbol{\phi},\boldsymbol{\psi}):\boldsymbol{\phi}\perp\boldsymbol{v}_0,\boldsymbol{\psi}\perp\boldsymbol{u}_0}{\text{minimize}} \quad \|\boldsymbol{B}-\boldsymbol{\psi}\boldsymbol{\phi}^{\mathrm{T}}\|_{\mathrm{F}}^2$$

are $\{(c\boldsymbol{v}_1, \frac{\sigma_1}{c}\boldsymbol{u}_1) : c \neq 0\}$.

$$\frac{1}{2}\|\boldsymbol{B}-\boldsymbol{\psi}\boldsymbol{\phi}^{\mathrm{T}}\|_{\mathrm{F}}^2 = \frac{1}{2}\|\boldsymbol{B}\|_{\mathrm{F}}^2 - \underbrace{\left[\boldsymbol{\psi}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{\phi} - \frac{1}{2}\cdot(\boldsymbol{\phi}^{\mathrm{T}}\boldsymbol{\phi})\cdot(\boldsymbol{\psi}^{\mathrm{T}}\boldsymbol{\psi})\right]}_{H(\boldsymbol{\phi},\boldsymbol{\psi})}$$

SVD & low rank approximation

Corollary

The solutions of

$$\underset{(\boldsymbol{\phi},\boldsymbol{\psi}):\boldsymbol{\phi}\perp\boldsymbol{v}_0,\boldsymbol{\psi}\perp\boldsymbol{u}_0}{\text{minimize}} \quad \|\boldsymbol{B} - \boldsymbol{\psi}\boldsymbol{\phi}^{\mathrm{T}}\|_{\mathrm{F}}^2$$

are $\{(c\boldsymbol{v}_1, \frac{\sigma_1}{c}\boldsymbol{u}_1) : c \neq 0\}$.

$$\frac{1}{2}\|\boldsymbol{B} - \boldsymbol{\psi}\boldsymbol{\phi}^{\mathrm{T}}\|_{\mathrm{F}}^2 = \frac{1}{2}\|\boldsymbol{B}\|_{\mathrm{F}}^2 - \underbrace{\left[\boldsymbol{\psi}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{\phi} - \frac{1}{2}\cdot(\boldsymbol{\phi}^{\mathrm{T}}\boldsymbol{\phi})\cdot(\boldsymbol{\psi}^{\mathrm{T}}\boldsymbol{\psi})\right]}_{H(\boldsymbol{\phi},\boldsymbol{\psi})}$$

$$H(f(X), g(Y)) \triangleq \mathbb{E}[f(X)g(Y)] - \frac{1}{2}\cdot\mathrm{Var}(f(X))\cdot\mathrm{Var}(g(Y)).$$

H-score

$$H(f(X), g(Y)) \triangleq \mathbb{E}[f(X)g(Y)] - \frac{1}{2} \cdot \mathrm{Var}(f(X)) \cdot \mathrm{Var}(g(Y)).$$

$$H(f(X)) \triangleq \max_{g} H(f(X), g(Y)) = \frac{1}{2} \cdot \frac{\mathrm{Var}([\mathbb{E}[f(X)|Y]])}{\mathrm{Var}(\mathbb{E}[f(X)])}.$$

Outline

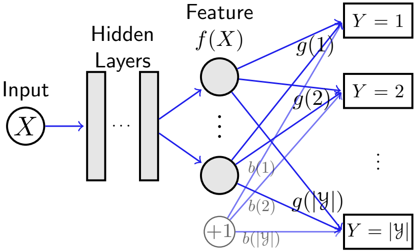
Principal Angles

Outline

Theoretical results

H-score v.s. Log-loss

- Local assumption: X and Y are nearly independent.
- $\ell_{\log} \approx -\frac{1}{2}H(\boldsymbol{f}(X), \boldsymbol{g}(Y)) + C$



$$\tilde{P}_{Y|X}(y|x) \triangleq \frac{e^{f^{\mathrm{T}}(x)g(y)+b(y)}}{\sum_{y' \in \mathcal{Y}} e^{f^{\mathrm{T}}(x)g(y')+b(y')}}$$

Softmax Output

NN for ACE

H-score & ACE-Net:

Warning

Outliers in ACE

Extensions

MACE: Modified(Multiple) ACE Algorithm

References I

[1] Leo Breiman and Jerome H Friedman.
Estimating optimal transformations for multiple regression and correlation.
Journal of the American statistical Association, 80(391):580–598, 1985.

[2] Hans Gebelein.
Das statistische problem der korrelation als variations-und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung.
ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik, 21(6):364–379, 1941.

[3] Hermann O Hirschfeld.
A connection between correlation and contingency.
In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 31, pages 520–524. Cambridge University Press, 1935.

[4] Shao-Lun Huang, Anuran Makur, Lizhong Zheng, and Gregory W Wornell.
An information-theoretic approach to universal feature selection in high-dimensional inference.
In *Information Theory (ISIT), 2017 IEEE International Symposium on*, pages 1336–1340. IEEE, 2017.

References II

[5] Shao-Lun Huang, Lin Zhang, and Lizhong Zheng.
An information-theoretic approach to unsupervised feature selection for high-dimensional data.
In *Information Theory Workshop (ITW), 2017 IEEE*, pages 434–438. IEEE, 2017.

[6] Anuran Makur, Fabián Kozynski, Shao-Lun Huang, and Lizhong Zheng.
An efficient algorithm for information decomposition and extraction.
In *Communication, Control, and Computing (Allerton), 2015 53rd Annual Allerton Conference on*, pages 972–979. IEEE, 2015.

[7] Alfréd Rényi.
On measures of dependence.
Acta mathematica hungarica, 10(3-4):441–451, 1959.

Thank you!