Lecture 23

AoI in supervised learning (2)

Reading:

Shisher et al AoI workshop 2021

Generalized Information theoretic measures for ERM:

- Entropy: $H_L(Y) = \min_{a \in \mathcal{A}} E_{Y \sim P_Y}[L(Y, a)]$.

  entropy associated with a loss function L.

  minimum loss without knowing X.

- Conditional entropy of Y given $X = x$:

$$H_L(Y|X=x) = \min_{a \in \mathcal{A}} E_{Y \sim P_{Y|X=x}}[L(Y, a)].$$

$$= \min_{\psi(x) \in \mathcal{A}} E_{Y \sim P_{Y|X=x}}[L(Y, \psi(x))].$$

- Conditional entropy of Y given X:

$H_L(Y|X)$

$$= \min_{\psi \in \Psi} E_{XY \sim P_{XY}}[L(Y, \psi(X))].$$

$$= \min_{\psi \in \Psi} \sum_{x,y} P_{XY}(x,y) \, L(y, \psi(x)).$$

$$= \min_{\psi \in \mathcal{Y}} \sum_{x \in X} P_X(x) \sum_{y \in Y} P_{Y|X=x}(y) L(y, \psi(x)).$$

$$= \min_{\psi(x) \in A, \forall x} \sum_{x \in X} P_X(x) \sum_{y \in Y} P_{Y|X=x}(y) L(y, \psi(x))$$

$$= \sum_{x \in X} P_X(x) \min_{\psi(x) \in A} \sum_{y \in Y} P_{Y|X=x}(y) L(y, \psi(x))$$

$$= \sum_{y \in X} P_X(x) \min_{\psi(x) \in A} E_{Y \sim P_{Y|X=x}}[L(Y, \psi(x))]$$

$$= \sum_{x \in X} P_X(x) H_L(Y|X=x).$$

$$H_L(Y|X) = E_{X \sim P_X}[H_L(Y|X=x)]$$

- Mutual information:

$$I_L(Y;X) = H_L(Y) - H_L(Y|X).$$

the reduction of the loss due to knowledge of X.

$$I_L(Y;XZ) = I_L(Y;Z) + I_L(Y;X|Z)$$

$$= [H_L(Y) - H_L(Y|Z)] + [H_L(Y|Z) - H_L(Y|XZ)]$$

Note:

$I_L(Y;X) \neq I_L(X;Y)$ in general.

Shannon's mutual function is a special case.

where. $I(Y;X) = I(X;Y)$.

- Divergence:

Def $a_{P_Y} = \arg\min_{a \in A} E_{Y \sim P_Y}[L(Y,a)]$.

be the optimal action, call "Bayes Action".

$$D_L(P_Y \| Q_Y]$$

$$= E_{Y \sim P_Y}(L(Y, a_{Q_Y})) - E_{Y \sim P_Y}[L(Y, a_{P_Y})]$$

$$= E_{Y \sim P_Y}(L(Y, a_{Q_Y})) - H_L(Y) \geq 0.$$

if $P_Y = Q_Y$, then $D_L(P_Y \| Q_Y) = 0$.

Then,

$$I_L(Y;X) = E_{X \sim P_X}[D_L(P_{Y|X} \| P_Y)].$$

$$= H_L(Y) - H_L(Y|X).$$

Example:

① log-loss:

$$L_{\log}(y, Q_Y) = -\log Q_Y(y).$$

$$H_{\log}(Y) = -\sum_{y \in \mathcal{Y}} P_Y(y) \log P_Y(y).$$

Shannon's entropy.

② 0-1 loss:

$$L_{0\text{-}1}(y, \hat{y}) = 1(y \neq \hat{y})$$

$$= \begin{cases} 1 & \text{if } y \neq \hat{y} \\ 0 & \text{if } y = \hat{y} \end{cases}.$$

$$H_{0\text{-}1}(Y) = 1 - \max_{y \in \mathcal{Y}} P_Y(y).$$

③ Quadratic loss:

action $a = \hat{y}$.

$$L_2(y, \hat{y}) = (y - \hat{y})^2.$$

$$H_2(Y) = E[Y^2] - E[Y]^2 = \mathrm{Var}[Y].$$

for general L,

$H_L(Y)$ may not be positive.

AoI in supervised learning:

$$H_L(Y_t \mid X_{t-\Delta(t)})$$

predict Y_t based on feature $X_{t-\Delta(t)}$.

Assume $(X_t, Y_t)_{t=1}^{\infty}$ is stationary.

$$H_L(Y_t \mid X_{t-\Delta(t)}) = H_L(Y_{\Delta(t)} \mid X_0).$$

Thm 1.

$g(\delta) = H_L(Y_\delta \mid X_0)$ is a function of the age δ,

where

$$H_L(Y_\delta \mid X_0) = g_1(\delta) - g_2(\delta).$$

$$g_1(\delta) = H_L(Y_0 \mid X_0) + \sum_{k=0}^{\delta-1} I_L(Y_t; X_{t-k} \mid X_{t-k-1}).$$

$$g_2(\delta) = \sum_{k=0}^{\delta-1} I_L(Y_t; X_{t-k-1} \mid X_{t-k}).$$

g_1, g_2 are non-decreasing functions.

For Markov chain $g_2(\delta) = 0$.

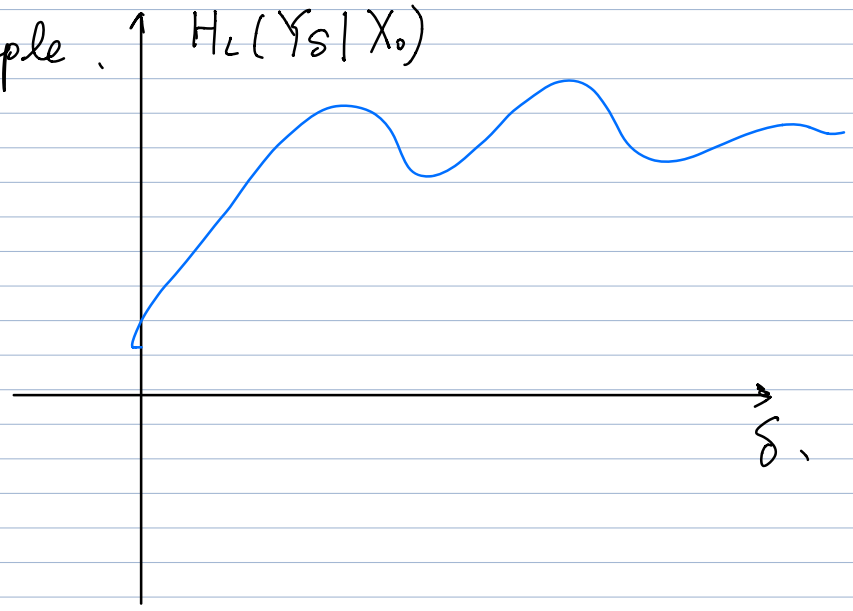
Hence,

$H_L(Y_\delta \mid X_0)$ can be increasing

— — .. decreasing.

Reading: Shisher et al. AoI workshop 2021.

Example.



Question: When is $H_L(Y_\delta | X_0)$ a non-decreasing function of δ?

Lemma 1: If $Y \leftrightarrow X \leftrightarrow Z$ is a Markov chain, then

$$H_L(Y|X) \leq H_L(Y|Z)$$

This is called Data Processing Inequality.

Lemma 2: If $Y_t \leftrightarrow X_{t-\delta} \leftrightarrow X_{t-\delta-\tau}$ is a Markov chain, then

$$H_L(Y_t | X_{t-\delta}) \leq H_L(Y_t | X_{t-\delta-\tau}).$$

Proof of Lemma 1:

If $Y \leftrightarrow X \leftrightarrow Z$, then

$$P_{Y,Z|X=x}(x,y,z) = P_{Y|X=x}(y)\, P_{Z|X=x}(z).$$

$$Y \perp Z \mid\mid X.$$

Given $X=x$, consider the term.

$$E_{Y,Z \sim P_{Y,Z|X=x}}\left[L(Y, \psi(x,Z))\right].$$

$$= E_{Y,Z \sim P_{Y|X=x} \otimes P_{Z|X=x}}\left[L(Y, \psi(x,Z))\right].$$

$$= \sum_{z \in \mathcal{Z}} P_{Z|X=x}(z) \underbrace{\sum_{y \in \mathcal{Y}} P_{Y|X=x}(y)\, L(y, \psi(x,z))}_{\alpha(x,z)}.$$

To minimize $\alpha(x,z)$, we let

$$\psi(x,z) = \arg\min_{a \in \mathcal{A}} \sum_{y \in \mathcal{Y}} P_{Y|X=x}(y)\, L(y,a),$$

$$= \arg\min_{a \in \mathcal{A}} E_{Y \sim P_{Y|X=x}}\left[L(Y,a)\right].$$

The last term is merely a function of X.

Hence, there exists an optimal action

$$\psi(x,z) = \phi(x).$$

such that

i $\phi(x)$ minimizes $\alpha(x, z)$.

ii $\phi(x)$ depends only on x.

$$\therefore \; E_{Y,Z \sim P_{Y,Z|X=x}}[L(Y, \psi(x, Z))]$$

$$\geq E_{Y,Z \sim P_{Y,Z|X=x}}[L(Y, \phi(x))].$$

$$H_L(Y|XZ)$$

$$= \min_{\psi} E_{X,Y,Z \sim P_{XYZ}}[L(Y, \psi(X, Z))]$$

$$\geq \min_{\phi} E_{X,Y,Z \sim P_{X,Y,Z}}[L(Y, \phi(X))].$$

$$= \min_{\phi} E_{X,Y \sim P_{X,Y}}[L(Y, \phi(X))]$$

$$= H_L(Y|X).$$

On the other hand,

$$H_L(Y|X)$$

$$= \min_{\phi} E_{X,Y,Z \sim P_{X,Y,Z}}[L(Y, \phi(X))]$$

$$\geq \min_{\psi} E_{X,Y,Z \sim P_{XYZ}}[L(Y, \psi(X, Z))]$$

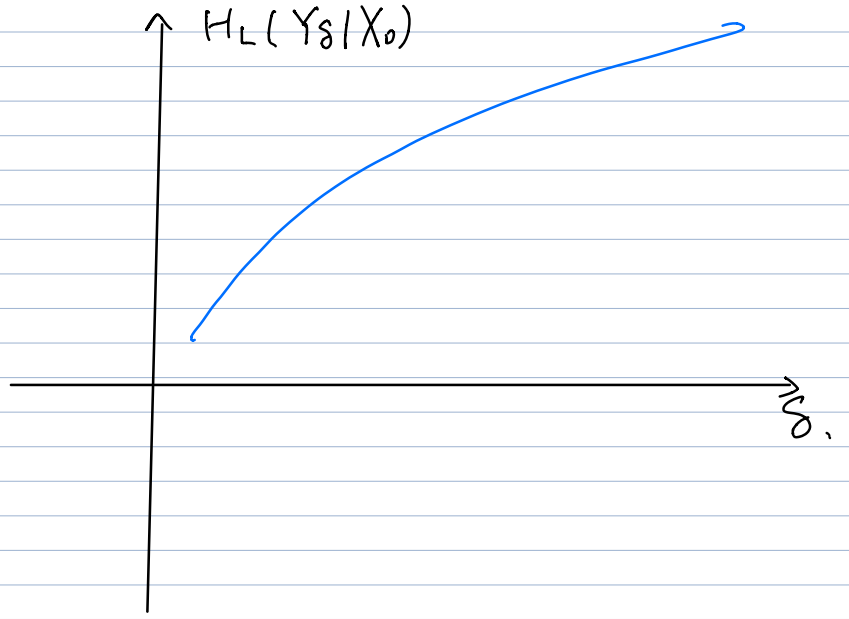
$= H_L(Y|XZ)$.

$$\begin{cases} H_L(Y|X) = H_L(Y|XZ) \\ H_L(Y|XZ) \le H_L(Y|Z), \end{cases}$$

$I_L(Y;Z|X) = H_L(Y|X) - H_L(Y|XZ) = 0.$

$$H_L(Y|X) \le H_L(Y|Z)$$ ∎

If $Y_t \leftrightarrow X_{t-s} \leftrightarrow X_{t-s-\tau}$ is a Markov chain for all $s, \tau \ge 0$.



Practical training data:

Never exactly a Markov chain.

sometimes it is close to a Markov chain.
sometimes it is quite different from
a Markov chain.

One can prove that.

If $(Y_t, X_{t-s}, X_{t-s-\tau})$ is close to
a Markov chain. then

$$H_L(Y_s | X_0)$$

is close to an non-decreasing function of s.

ε-Markov chain $\rightarrow$ ε-data processing inequality

Reading:

Shisher et al. AoI workshop 2021.