Lecture 10

Mutual information

$$I(X;Y) = D_{KL}\left(p(x,y) \,\|\, p(x)p(y)\right) =$$

$$= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} .$$

$$I(X;Y) \geq 0 \; ; \; = 0 \text{ iff } p(x,y) = p(x)p(y).$$

Note that

$$I(X;Y) = \sum_{x,y} p(x|y)\, p(y) \log p(x|y) -$$

$$- \sum_{x,y} p(x,y) \log p(x) =$$

$$= \underbrace{\sum_{y} p(y) \sum_{x} p(x|y) \log p(x|y)}_{-H(X|Y)} + H(X) .$$

Likewise, $I(X;Y) = H(Y) - H(Y|X)$.

Thus, $I(X;Y)$ quantifies the reduction in uncertainty about X after observing Y, and vice versa.

Recall that $H(X|Y) \leq H(X)$, consistent with $0 \leq I(X;Y) = H(X) - H(X|Y)$.

One can show that

$$I(X;Y) = H(X) + H(Y) - \underbrace{H(X,Y)}_{H(X)+H(Y|X)} =$$

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

Likewise, $I(X;Y) = \underbrace{H(X,Y)}_{H(X)+H(Y|X)} - H(X|Y) - H(Y|X) =$

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

Conditional mutual information

$$I(X;Y|Z) \equiv \sum_z p(z) \left[\sum_{x,y} p(x,y|z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} \right] \quad (\equiv)$$

$$(\equiv) -\sum_z p(z) \left[\sum_x p(x|z) \log p(x|z) \right] - \sum_z p(z) \left[\sum_y p(y|z) \times \right.$$

$$\left. \times \log p(y|z) \right] + \sum_z p(z) \left[\sum_{x,y} p(x,y|z) \log p(x,y|z) \right] =$$

$$= H(X|Z) + H(Y|Z) - \underbrace{H(X,Y|Z)}_{H(Y|Z)+H(X|Y,Z)} = \underbrace{H(X|Z)}_{H(X,Z)-H(Z)} - \underbrace{H(X|Y,Z)}_{H(X,Y,Z)-H(Z,Y)} \quad (\Diamond)$$

see below

$$(\Diamond) \equiv H(X,Z) + H(Y,Z) - H(Z) - H(X,Y,Z) \quad (\equiv)$$

$$H(X|Y,Z) = H(X,Y,Z) - H(Y) - H(Z|Y) =$$

$$= H(X,Y,Z) - \cancel{H(Y)} - H(Z,Y) + \cancel{H(Y)}$$

$$\overbrace{H(Y,Z) - H(Y) - H(Z)}^{-I(Y;Z)} + H(Y) + H(X,Z) - \overbrace{H(X,Y,Z)}^{I(Y;X,Z)}$$

$$= I(Y;X,Z) - I(Y;Z).$$

Thus, conditional MI is the extra info that XZ tells us about Y, excluding what Z alone has told us about Y.

Finally, we can write

$$I(Y;X,Z) = I(X;Y|Z) + I(Y;Z)$$

$X \leftrightarrow Y$ yields

$$I(Y,Z;X) = I(\underset{\downarrow}{Z}\,;\underset{\downarrow}{X}) + I(\underset{\downarrow}{Y};X|\underset{\downarrow}{Z}).$$

$Z \to Z_1$, $Y \to Z_2$, $Z \to Z_1$

$$I(Z_1,Z_2;X) = I(Z_1;X) + I(Z_2;X|Z_1).$$

This generalizes to

$$I(Z_1,\dots,Z_N;X) = \sum_{n=1}^{N} I(Z_n;X|Z_1,\dots,Z_{n-1}).$$

[chain rule for mutual information]

Generalized correlation coefficient

Consider $\begin{pmatrix} x \\ y \end{pmatrix} \sim \mathcal{N}\left(\vec{0}, \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$ jointly gaussian ($d=2$)

Recall that if $X \sim \mathcal{N}(\mu, \sigma^2)$,

$$h(X) = \frac{1}{2} + \frac{1}{2}\log(2\pi\sigma^2).$$

Likewise, if $\vec{X} \sim \mathcal{N}(\vec{\mu}, \Sigma)$ d-dimensional gaussian

$$h(\vec{X}) = \frac{d}{2} + \frac{d}{2}\log(2\pi) + \frac{1}{2}\log \underbrace{|\Sigma|}_{\det(\Sigma)}.$$

In our case,

$x \sim \mathcal{N}(0, \sigma^2)$ and $y \sim \mathcal{N}(0, \sigma^2)$, s.t.

$$h(x) = h(y) = \frac{1}{2} + \frac{1}{2}\log(2\pi\sigma^2).$$

Moreover,

$$h(x,y) = 1 + \log(2\pi) + \frac{1}{2}\log[\sigma^4(1-\rho^2)].$$

Then $I(x;y) = h(x) + h(y) - h(x,y) =$

$$= \log(2\pi\sigma^2) - \log(2\pi) - \underbrace{\frac{1}{2}\log\sigma^4}_{\log\sigma^2} - \frac{1}{2}\log(1-\rho^2) \quad (=)$$

$$(=) \frac{1}{2}\log\frac{1}{1-\rho^2}.$$

Special cases:

(a) $\rho = 1 \Rightarrow X = Y \Rightarrow I(x;y) = \infty$

Observing Y tells us X exactly

(b) $\rho = 0 \Rightarrow X \perp Y \Rightarrow I(x;y) = 0$

Observing Y tells us nothing about X

(c) $\rho = -1 \Rightarrow X = -Y \Rightarrow I(x;y) = \infty$

Observing Y tells us X exactly, as in (a)

Normalized MI

Recall that

$$I(X;Y) = \begin{cases} H(X) - H(X|Y) \leq H(X) \\ H(Y) - H(Y|X) \leq H(Y) \end{cases}$$

So, $0 \leq I(X;Y) \leq \min\{H(X), H(Y)\}$.

Thus, we can define

$$NMI(X,Y) = \frac{I(X;Y)}{\min(H(X), H(Y))}$$

clearly, $0 \leq NMI(X,Y) \leq 1$.

If $NMI(X,Y) = 0 \Rightarrow I(X;Y) = 0 \Rightarrow X \perp Y$.

If $NMI(X,Y) = 1$ and $H(X) < H(Y)$, [$H(Y) < H(X)$ argued in the same way]

we have $I(X;Y) = H(X) - H(X|Y) = H(X) \Rightarrow$

$\Rightarrow H(X|Y) = 0$, Y determines X completely

NMI may be tricky to compute for continuous variables $\Rightarrow$ dependence on bin sizes.

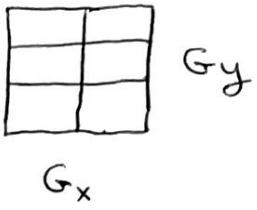
Maximal information coefficient (MIC)

Define $MIC(X,Y) = \max_{G} \frac{I(X;Y)|_G}{\log G_{min}}$,

where G = set of 2D grids, and

$G_{min} = \min\{G_x, G_y\}$

$G_x \times G_y$ grid

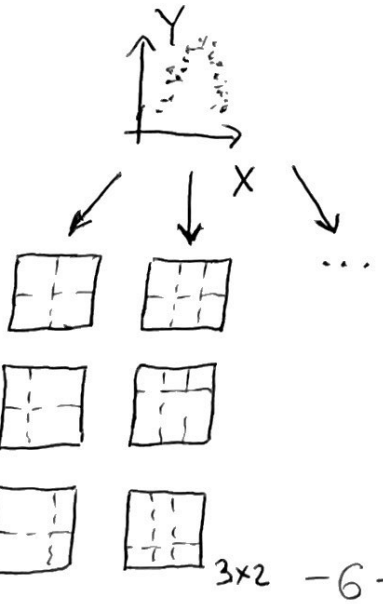


$\log G_{min}$ is the max entropy value for $H(X)$ or $H(Y)$ [the min of the two, actually] => this ensures $0 \le MIC(X,Y) \le 1$.

Finally, $I(X;Y)|_G$ is $I(X;Y)$ computed on a given grid G.

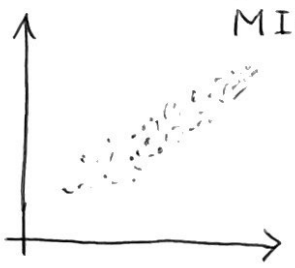
Empirically, $G_x G_y \le N^{0.6}$.

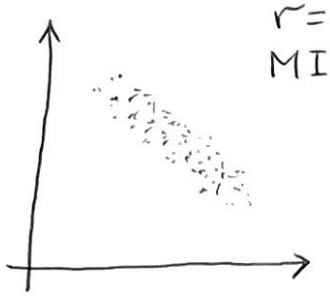
↑ sample size

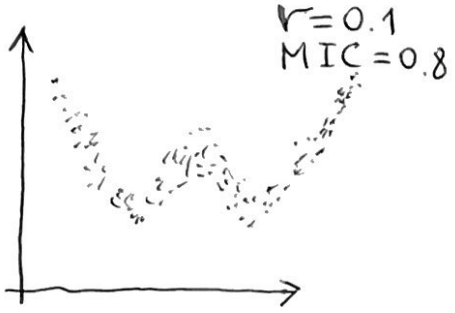


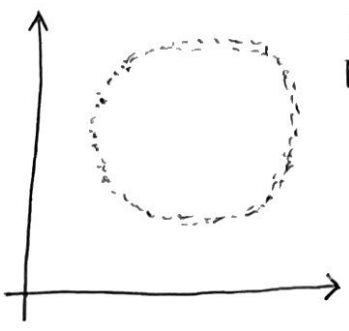
MIC is the max value on all these grids

Ex.









Note that MIC is not restricted to linear relations $\Rightarrow$ MIC ≈ 1.0 for any non-linear relationship.

Data processing inequality (DPI)

Suppose $X \to Y \to Z$ form a Markov chain: (Y: unobserved; Z: [potentially noisy] observ'n of a f'n of Y)

$X \perp Z \mid Y$.

(Z: noisy observation of a function of X)

Chain rule: $I(X; Y, Z) = I(X; Z) + \underbrace{I(X; Y \mid Z)}_{\geq 0} =$

$= I(X; Y) + \underbrace{I(X; Z \mid Y)}_{=0 \text{ since } X \perp Z \mid Y}$, or

$$I(X; Z) \leq I(X; Y).$$

Similarly, $I(X; Z) \leq I(Y; Z)$

We cannot increase the amount of information we have about unknown X by further processing.

Sufficient statistics

Consider $\vec{\theta} \to \mathcal{D} \to S(\mathcal{D})$ ← some function of the data

- $\vec{\theta}$: unknown model prms (hidden vars)
- $\mathcal{D}$: observed data

DPI gives: $I(\vec{\theta}; S(\mathcal{D})) \leq I(\vec{\theta}; \mathcal{D})$.

If $I(\vec{\theta}; S(\mathcal{D})) = I(\vec{\theta}; \mathcal{D})$,

$S(\mathcal{D})$ is a sufficient statistic of $\mathcal{D}$, for the purpose of inferring $\vec{\theta}$.

Define minimal sufficient statistic as the one that maximally compresses $\mathcal{D}$ without hurting inference of $\vec{\theta}$.

Ex. (a) N Bernoulli trials:

$S(\mathcal{D}) = \{N, N_1\}$ → to find Bernoulli $\vec{\theta}$

N_1: # pos. outcomes

no need to keep the entire sequence of events

(b) N samples from a 1D gaussian: $\mathcal{N}(\mu, \sigma^2)$

To find $\vec{\theta} = \{\mu, \sigma^2\}$, just need

$S(\mathcal{D}) = \{N, \text{mean of data}, \text{var. of data}\}$.

Fano's inequality

Consider $Y \to X \to \hat{Y}$ ← $f(X)$, estimator of Y (classifier)

Y: true class labels; X: features in the classifier

Define $E : \hat{Y} \neq Y$ misclassification event

$P_e = P(E=1)$ prob. of error

Chain rule for entropy:

$$H(E, Y|\hat{Y}) = H(Y|\hat{Y}) + \underbrace{H(E|Y,\hat{Y})}_{=0} = H(E|\hat{Y}) + H(Y|E,\hat{Y})$$

E completely defined by Y and $\hat{Y}$

Now, $H(E|\hat{Y}) \leq H(E)$ conditioning reduces entropy

$$H(Y|E,\hat{Y}) = \overbrace{P(E=0)}^{1-P_e} \overbrace{H(Y|\hat{Y}, E=0)}^{=0,\text{ no error}} + \underbrace{P(E=1)}_{P_e} \underbrace{H(Y|\hat{Y}, E=1)}_{\leq \log K} \leq P_e \log K.$$

K: # classes

Finally, DPI gives $I(Y;\hat{Y}) \leq I(Y;X)$, or

$H(Y) - H(Y|\hat{Y})$ $\leq$ $H(Y) - H(Y|X)$

$$H(Y|X) \leq H(Y|\hat{Y}) \leq \underbrace{H(E)}_{\leq 1} + P_e \log K .$$

≤ 1, entropy of a Bernoulli r.v. if $\log \to \log_2$

Thus, $H(Y|X) \leq 1 + P_e \log K$, or

$$P_e \geq \frac{H(Y|X) - 1}{\log K}$$ Fano's inequality

We want to minimize $H(Y|X)$ in order to minimize the RHS $\Rightarrow$

$\Rightarrow$ maximize $I(Y; X)$.

In other words, pick features (input) that have high MI with class labels Y.