Infotheory for Statistics and Learning

Lecture 3

- Conditional relative entropy, mutual information and f-divergence [PW:2.5,3.4,7],[CT:2]
- Data processing inequalities [PW:2.5,3.5,7.2],[CT:2.8]
- Sufficient statistics [PW:3.5],[CT:2.9]
- The information bottleneck [GP]

Conditional Relative Entropy

Given two random transformations $P_{Y|X=x}$ and $Q_{Y|X=x}$, define

$$D(P_{Y|X}\|Q_{Y|X}|P_X) = \int D(P_{Y|X=x}\|Q_{Y|X=x})dP_X$$
$$= \int \left\{ \int \log \frac{dP_{Y|X=x}}{dQ_{Y|X=x}} dP_{Y|X=x} \right\} dP_X$$

For discrete $P_X \to p(x), P_{Y|X} \to p(y|x), Q_{Y|X} \to q(y|x)$,

$$D(p(y|x)\|q(y|x)|p(x)) = \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)}$$

For abs. continuous $P_X \to f(x), P_{Y|X} \to f(y|x), Q_{Y|X} \to g(y|x)$,

$$D(f(y|x)\|g(y|x)|f(x)) = \int f(x) \left\{ \int f(y|x) \log \frac{f(y|x)}{g(y|x)} dy \right\} dx$$

Equivalent definition

$$D(P_{Y|X} \| Q_{Y|X} | P_X) = D(P_{Y|X} \times P_X \| Q_{Y|X} \times P_X)$$

Chain rule

$$D(P_{XY} \| Q_{XY}) = D(P_{Y|X} \| Q_{Y|X} | P_X) + D(P_X \| Q_X)$$

Consequently, for $P_Y = P_{Y|X} \circ P_X$ and $Q_Y = Q_{Y|X} \circ P_X$

$$\begin{aligned} D(P_{Y|X} \times P_X \| Q_{Y|X} \times P_X) &= D(P_{Y|X} \| Q_{Y|X} | P_X) + D(P_X \| P_X) \\ &= D(P_{X|Y} \| Q_{X|Y} | P_Y) + D(P_Y \| Q_Y) \end{aligned}$$

$\Rightarrow D(P_Y \| Q_Y) \leq D(P_{Y|X} \| Q_{Y|X} | P_X)$
with $=$ only if $D(P_{X|Y} \| Q_{X|Y} | P_Y) = 0$

For instead $P_Y = P_{Y|X} \circ P_X$ and $Q_Y = P_{Y|X} \circ Q_X$, we get

$$\begin{aligned} D(P_{Y|X} \times P_X \| Q_{Y|X} \times Q_X) &= D(P_{Y|X} \| P_{Y|X} | P_X) + D(P_X \| Q_X) \\ &= D(P_{X|Y} \| Q_{X|Y} | P_Y) + D(P_Y \| Q_Y) \end{aligned}$$

$\Rightarrow D(P_Y \| Q_Y) \leq D(P_X \| Q_X)$
with $=$ only if $D(P_{X|Y} \| Q_{X|Y} | P_Y) = 0$

Data processing inequality
Passing P_X and Q_X through the same transformation decreases the distance

Mutual Information

We get

$$I(X;Y) = D(P_{XY} \| P_X \otimes P_Y) = D(P_{Y|X} \| P_Y | P_X) = D(P_{X|Y} \| P_X | P_Y)$$

Also define

$$\begin{aligned} I(X;Y|Z) &= D(P_{XY|Z} \| P_{X|Z} \otimes P_{Y|Z} | P_Z) \\ &= \int \left\{ \int \log \frac{dP_{XY|Z=z}}{d(P_{X|Z=z} \otimes P_{Y|Z=z})} dP_{XY|Z=z} \right\} dP_Z \\ &= \int I(X;Y|Z=z) dP_Z \\ &= H(Y|Z) - H(Y|X,Z) \quad \text{[discrete]} \\ &= h(Y|Z) - h(Y|X,Z) \quad \text{[abs. cont.]} \end{aligned}$$

Note that $I(X;Y|Z) \geq 0$ with $=$ only if $X \to Z \to Y$

Chain rule

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

Consequently, if $X \to Y \to Z$

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

so $I(X;Z) \leq I(X;Y)$ with $=$ only if $I(X;Y|Z) = 0$

Follows also from $D(P_{Z|X} \| P_Z | P_X) \leq D(P_{Y|X} \| P_Y | P_X)$

Data processing inequality
Further processing/randomness decreases information

The Golden Formula

Given P_X, $P_{Y|X}$ and $P_Y = P_{Y|X} \circ P_X$, let Q_Y be any other output distribution, then

$$I(X;Y) = D(P_{Y|X} \| Q_Y | P_X) - D(P_Y \| Q_Y)$$

Thus $I(X;Y) \leq D(P_{Y|X} \| Q_Y | P_X)$ and

$$I(X;Y) = \min_{Q_Y} D(P_{Y|X} \| Q_Y | P_X)$$

achieved at $Q_Y = P_Y$

Conditional f-divergence

Let

$$D_f(P_{Y|X} \| Q_{Y|X} | P_X) = \int D_f(P_{Y|X=x} \| Q_{Y|X=x}) dP_X$$

For $P_Y = P_{Y|X} \circ P_X$ and $Q_Y = Q_{Y|X} \circ P_X$,

$$D_f(P_Y \| Q_Y) \leq D_f(P_{Y|X} \| Q_{Y|X} | P_X)$$

and for $P_Y = P_{Y|X} \circ P_X$ and $Q_Y = P_{Y|X} \circ Q_X$,

$$D_f(P_Y \| Q_Y) \leq D_f(P_X \| Q_X)$$

(data processing inequality)

Sufficient Statistics

Let $P_X^{(\theta)}$ be parameterized by $\theta \in \mathbb{R}$

Map X to T via $P_{T|X}$, i.e. $P_T^{(\theta)} = P_{T|X} \circ P_X^{(\theta)}$

Given $P_{T|X}$, T is a sufficient statistic of X for θ if there is a kernel $Q_{X|T}$ such that

$$P_{T|X} \times P_X^{(\theta)} = Q_{X|T} \times P_T^{(\theta)}$$

When interested in θ, if T is known one can forget X

For a random θ: $P_X^{(\theta)} \to P_{X|\theta}$, and with some arbitrary P_θ

Given $P_{X|\theta}$, $P_{T|X}$ and $\theta \to X \to T$, the following are equivalent

1) T is a sufficient statistic of X for θ
2) $\theta \to T \to X$, for any P_θ
3) $I(\theta; X|T) = 0$, for any P_θ
4) $I(\theta; X) = I(\theta; T)$, for any P_θ

A sufficient statistic T^* is minimal if $\theta \to X \to T \to T^*$ for all sufficient T, i.e. $I(X; T^*) \leq I(X; T)$ for all T while

$$I(\theta; X) = I(\theta; T) = I(\theta; T^*)$$

The Information Bottleneck

A minimal sufficient statistic may not exist

$\Rightarrow$ consider instead the problem

$$\inf_{P_{T|X}} \left(I(X;T) - \lambda I(\theta;T) \right)$$

for $\lambda \geq 0$

Varying λ traces out $(I(X;T), I(\theta;T))$, the information curve that separates achievable from non-achievable in the information plane

Equivalently, require $I(\theta;T) \geq \alpha$, $0 \leq \alpha \leq I(\theta;X)$, and define

$$d(\omega) = -\log \frac{dP_{T|\theta}}{dP_T}(\omega)$$

Then we have

$$R(\alpha) = \inf_{P_{T|X}:E[d(\omega)] \leq \alpha} I(X;T)$$

to get a rate–distortion characterization

$I(X;T)$ can be interpreted as the complexity of the description T, versus relevance $I(\theta;T)$ parameterized by α

When learning T from data: higher complexity $\Rightarrow$ harder to learn

$R(\alpha)$ = the complexity–relevance function