Infotheory for Statistics and Learning

Lecture 1

- Entropy [PW:1],[CT:2,8]
- Relative entropy [PW:2], [CT:2]
- Mutual information [PW:3], [CT:2]
- f-divergence [PW:7]

Entropy

Over $(\mathbb{R}, \mathcal{B})$, consider a discrete RV X with all probability in a countable set $\mathcal{X} \in \mathcal{B}$, the alphabet of X

Let $p_X(x)$ be the pmf of X for $x \in \mathcal{X}$

The (Shannon) entropy of X

$$H(X) = -\sum_{x \in \mathcal{X}} p_X(x) \log p_X(x)$$

- the logarithm is base-2 if not declared otherwise
- sometimes denoted $H(p_X)$ to emphasize the pmf p_X
- $H(X) \geq 0$ with $=$ only if $p_X(x) = 1$ for some $x \in \mathcal{X}$
- $H(X) \leq \log |\mathcal{X}|$ (for $|\mathcal{X}| < \infty$) with $=$ only if $p_X(x) = 1/|\mathcal{X}|$
- $H(p_X)$ is concave in p_X

For two discrete RVs X and Y, with alphabets $\mathcal{X}$ and $\mathcal{Y}$ and a joint pmf $p_{XY}(x,y)$, we have the joint entropy

$$H(X,Y) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p_{XY}(x,y)\log p_{XY}(x,y)$$

Conditional entropy

$$\begin{aligned} H(Y|X) &= -\sum_x p_X(x)\sum_y p_{Y|X}(y|x)\log p_{Y|X}(y|x) \\ &= \sum_x p_X(x)H(Y|X=x) \\ &= H(X,Y) - H(X) \end{aligned}$$

Extension to > 2 variables straightforward

Relative Entropy

Assume P and Q are two prob. measures over $(\Omega,\mathcal{A})$

Emphasize expectation w.r.t. P (or Q) as $E_P[\cdot]$ (or $E_Q[\cdot]$)

The relative entropy between P and Q

$$D(P\|Q) = E_P\left[\log\frac{dP}{dQ}\right]$$

if $P \ll Q$ and $D(P\|Q) = \infty$ otherwise

- $D(P\|Q) \geq 0$ with $=$ only if $P = Q$ on $\mathcal{A}$
- $D(P\|Q)$ is convex in (P,Q), i.e.

$$D(\lambda P_1+(1-\lambda)P_2\|\lambda Q_1+(1-\lambda)Q_2) \leq \lambda D(P_1\|Q_1)+(1-\lambda)D(P_2\|Q_2)$$

Also known as divergence, or Kullback–Leibler (KL) divergence

$D(P\|Q)$ is not a metric (why?), but is still generally considered a measure of "distance" between P and Q

For discrete RVs: $P \to p_X$ and $Q \to p_Y$,

$$D(p_X \| p_Y) = \sum_x p_X(x) \log \frac{p_X(x)}{p_Y(x)}$$

For abs. continuous RVs : $P \to P_X \to f_X$ and $Q \to P_Y \to f_Y$,

$$D(P_X \| P_Y) = D(f_X \| f_Y) = \int f_X(x) \log \frac{f_X(x)}{f_Y(x)} dx$$

For a discrete RV X (with $|\mathcal{X}| < \infty$), note that

$$H(X) = \log |\mathcal{X}| - \sum_x p_X(x) \log \frac{p_X(x)}{1/|\mathcal{X}|}$$

$\Rightarrow H(p_X)$ is concave in p_X, entropy is negative distance to uniform

Mutual Information

Two variables X and Y with joint distribution P_{XY} on $(\mathbb{R}^2, \mathcal{B}^2)$ and marginals P_X and P_Y on $(\mathbb{R}, \mathcal{B})$

Mutual information

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

where $P_X \otimes P_Y$ is the product distribution on $(\mathbb{R}^2, \mathcal{B}^2)$

Discrete:

$$I(X;Y) = \sum_{x,y} p_{XY}(x,y) \log \frac{p_{XY}(x,y)}{p_X(x) p_Y(y)}$$

Abs. continuous:

$$I(X;Y) = \int f_{XY}(x,y) \log \frac{f_{XY}(x,y)}{f_X(x) f_Y(y)} dx dy$$

For discrete RVs, we see that

$$
\begin{aligned}
I(X;Y) &= H(X) + H(Y) - H(X,Y) \\
&= H(X) - H(X|Y) = H(Y) - H(Y|X)
\end{aligned}
$$

For abs. continuous P_X define differential entropy as

$$
h(X) = -D(P_X \| \lambda) = -\int f_X(x) \log f_X(x) d\lambda
$$

where λ is Lebesgue measure on $(\mathbb{R}, \mathcal{B})$, then we get

$$
\begin{aligned}
I(X;Y) &= h(X) + h(Y) - h(X,Y) \\
&= h(X) - h(X|Y) = h(Y) - h(Y|X)
\end{aligned}
$$

Saying $h(X) = -D(P_X \| \lambda)$ is a slight abuse, since λ is not a probability measure. Still, $h(X)$ can be interpreted as negative distance to "uniform"

Since

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

$I(X;Y) \geq 0$ with $=$ only if $P_{XY} = P_X \otimes P_Y$, i.e. X and Y indep.

Furthermore, since

$$
I(X;Y) = H(Y) - H(Y|X) \quad \text{or} \quad I(X;Y) = h(Y) - h(Y|X)
$$

we get $H(Y|X) \leq H(Y)$ and $h(Y|X) \leq h(Y)$,

conditioning reduces entropy

f-divergence

$f : (0, \infty) \to \mathbb{R}$ convex, strictly convex at $x = 1$ and $f(1) = 0$

Two probability measures P and Q on $(\Omega, \mathcal{A})$

μ any measure on $(\Omega, \mathcal{A})$ such that both $P \ll \mu$ and $Q \ll \mu$

Let

$$p(\omega) = \frac{dP}{d\mu}(\omega), \quad q(\omega) = \frac{dQ}{d\mu}(\omega)$$

The f-divergence between P and Q

$$D_f(P\|Q) = \int f\left(\frac{p(\omega)}{q(\omega)}\right) dQ = E_Q\left[f\left(\frac{p(\omega)}{q(\omega)}\right)\right]$$

When $P \ll Q$ we have

$$\frac{p(\omega)}{q(\omega)} = \frac{dP}{dQ}(\omega) \text{ and thus } D_f(P\|Q) = E_Q\left[f\left(\frac{dP}{dQ}(\omega)\right)\right]$$

When both P and Q are discrete, i.e. there is a countable set $K \in \mathcal{A}$ such that $P(K) = Q(K) = 1$, let $\mu =$ counting measure on K, i.e. $\mu(F) = |F|$ for $F \subset K$. Then p and q are pmf's and

$$D_f(P\|Q) = \sum_{\omega \in K} q(\omega) f\left(\frac{p(\omega)}{q(\omega)}\right)$$

When $(\Omega, \mathcal{A}) = (\mathbb{R}, \mathcal{B})$ and both P and Q have R–N derivatives w.r.t. Lebesgue measure $\mu = \lambda$ on $\mathcal{B}$, then p and q are pdfs and

$$D_f(P\|Q) = \int q(x) f\left(\frac{p(x)}{q(x)}\right) dx$$

In general, $D_f(P\|Q) \geq 0$ with $=$ only for $P = Q$ on $\mathcal{A}$

Also, $D_f(P\|Q)$ is convex in (P, Q)

Examples (assuming $P \ll Q$):

Relative entropy, $f(x) = x \log x$

$$D_f(P\|Q) = D(P\|Q) = E_Q\left[\frac{dP}{dQ}\log\frac{dP}{dQ}\right] = E_P\left[\log\frac{dP}{dQ}\right]$$

Total variation, $f(x) = \frac{1}{2}|x-1|$

$$D_f(P\|Q) = \mathsf{TV}(P,Q) = \frac{1}{2}E_Q\left|\frac{dP}{dQ} - 1\right| = \sup_{A\in\mathcal{A}}(P(A) - Q(A))$$

- discrete
$$\mathsf{TV}(P,Q) = \frac{1}{2}\sum_x |p(x) - q(x)|$$
- abs. continuous
$$\mathsf{TV}(P,Q) = \frac{1}{2}\int |p(x) - q(x)|dx$$

χ^2-divergence, $\chi^2(P,Q)$, $f(x) = (x-1)^2$

Squared Hellinger distance, $H^2(P,Q)$, $f(x) = (1-\sqrt{x})^2$

Hellinger distance, $H(P,Q) = \sqrt{H^2(P,Q)}$

Le Cam distance, $\mathsf{LC}(P\|Q)$, $f(x) = (1-x)/(2x+2)$

Jensen–Shannon symmetrized divergence,

$$f(x) = x\log\frac{2x}{x+1} + \log\frac{2}{x+1}$$
$$\mathsf{JS}(P\|Q) = D\left(P\middle\|\frac{P+Q}{2}\right) + D\left(Q\middle\|\frac{P+Q}{2}\right)$$

Inequalities for f-divergences

Consider $D_f(P\|Q)$ and $D_g(P\|Q)$ for P and Q on $(\Omega, \mathcal{A})$

Let

$$\mathcal{R}(f,g) = \{(D_f, D_g) : \text{over } P \text{ and } Q\}$$

and $\mathcal{R}_2(f,g) = \mathcal{R}(f,g)$ for the special case $\Omega = \{0,1\}$ and $\mathcal{A} = \sigma(\{0,1\}) = \{\emptyset, \{0\}, \{1\}, \{0,1\}\}$

Theorem: For any $(\Omega, \mathcal{A})$, $\mathcal{R} =$ the convex hull of $\mathcal{R}_2$

Let

$$F(x) = \inf\{y : (x,y) \in \mathcal{R}(f,g)\}$$

then

$$D_g(P\|Q) \geq F(D_f(P\|Q))$$

Example: For $g(x) = x \ln x$ and $f(x) = |x-1|$, it can be proved[1] that $(x, F(x))$ is obtained from

$$x = t\left(1 - (\coth(t) - \frac{1}{t})^2\right)$$
$$F = \log\left(\frac{t}{\sinh(t)}\right) + t\coth(t) - \frac{t^2}{\sinh^2(t)}$$

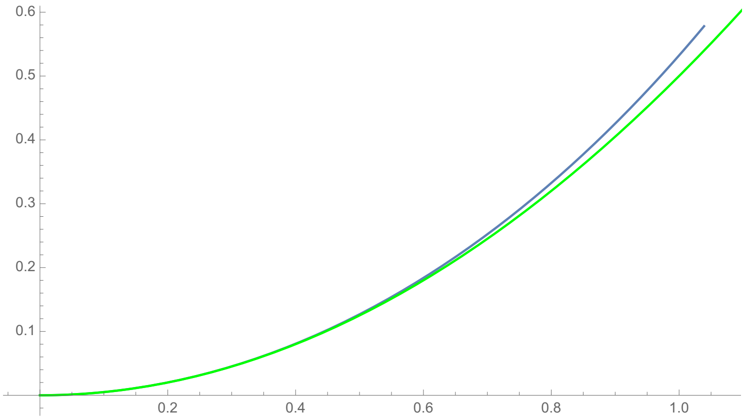
by varying $t \in (0, \infty)$

That is, given a t, resulting in (x, F), we have

$$D_g(P\|Q) = D(P\|Q) \geq F \quad \text{for} \quad D_f(P\|Q) = 2\text{TV}(P,Q) = x$$

(with $D(P\|Q)$ in nats, i.e. based on $\ln x$)

[1] See A. A. Fedotov, P. Harremoës and F. Topsøe, "Refinements of Pinsker's inequality," *IEEE Trans. IT*, 2003. The paper uses $V(P\|Q) = 2\text{TV}(P\|Q)$



Blue: The curve $(x(t), F(t))$ for $t > 0$

Green: The function $x^2/2$

Thus we have Pinsker's inequality

$$D(P\|Q) \geq \frac{1}{2}(D_f(P\|Q))^2 = 2\,(\mathsf{TV}(P,Q))^2$$

Or, for $D(P\|Q)$ in bits: $D(P\|Q) \geq 2\log e\,(\mathsf{TV}(P,Q))^2$

Other inequalities between f-divergences:

$$\frac{1}{2}H^2(P,Q) \leq \mathsf{TV}(P,Q) \leq H(P,Q)\sqrt{1 - H^2(P,Q)/4}$$

$$D(P\|Q) \geq 2\log\frac{2}{2 - H^2(P,Q)}$$

$$D(P\|Q) \leq \log(1 + \chi^2(P\|Q))$$

$$\frac{1}{2}H^2(P,Q) \leq \mathsf{LC}(P,Q) \leq H^2(P,Q)$$

$$\chi^2(P\|Q) \geq 4\,(\mathsf{TV}(P,Q))^2$$

For discrete p and q, "reverse Pinsker"

$$D(p\|q) \leq \log\left(1 + \frac{2}{\min_x q(x)}(\mathsf{TV}(p,q))^2\right) \leq \frac{2\log e}{\min_x q(x)}(\mathsf{TV}(p,q))^2$$