Information Theory with Applications, Math6397
Lecture Notes from September 02, 2014

taken by Kelley Walker

1.3 Concavity of Entropy (Continued)

1.3.12 Corollary. *If X and Y are discrete random variables then $H(X,Y) \leq H(X) + H(Y)$.*

Proof. In Proposition 1.1.7 we proved the inequality between conditional and unconditional entropy, $H(Y|X) \leq H(Y)$. Additionally, we have shown additivity of entropy, $H(X,Y) = H(X) + H(Y|X)$, in Proposition 1.2.9. Hence

$$H(X,Y) = H(X) + H(Y|X) \leq H(X) + H(Y).$$

□

1.3.13 Corollary. *It follows by induction that for finitely many random variables $X_1, X_2, \ldots, X_n$, we have*

$$H(X_1, X_2, \ldots, X_n) \leq \sum_{j=1}^{n} H(X_j).$$

1.3.14 Definition. Divergence for relative entropy between random variables X and Y with induced probability measures $\mathbb{P}_X$ and $\mathbb{Q}_Y$ on a common alphabet $\mathbb{A}$ is defined by

$$D(\mathbb{P}_X||\mathbb{Q}_Y) := \sum_{a\in\mathbb{A}} \mathbb{P}(a) \ln\left(\frac{\mathbb{P}_X(a)}{\mathbb{Q}_Y(a)}\right)$$

and is commonly denoted by $D(X||Y)$. We adopt the convection that $0\ln({}^0/_x) = 0$ if $x \geq 0$ and $x\ln({}^x/_0) = \infty$ if $x > 0$.

1.3.15 Remark. We observe that in general, $D(X||Y) \neq D(Y||X)$.

1.3.16 Theorem. *For random variables X and Y with induced measures $\mathbb{P}$ and $\mathbb{Q}$ on a common alphabet $\mathbb{A}$, $D(X||Y) \geq 0$ with equality if and only if $\mathbb{P} = \mathbb{Q}$.*

Proof. By the definition for $D(X||Y)$ and Jensen's Inequality we have

$$D(X||Y) = -\sum_{a\in\mathbb{A}} \mathbb{P}(a) \ln\left(\frac{\mathbb{Q}(a)}{\mathbb{P}(a)}\right) \geq \ln \sum_{a\in\mathbb{A}} \mathbb{P}(a) \left(\frac{\mathbb{Q}(a)}{\mathbb{P}(a)}\right) = -\ln \sum_{a\in\mathbb{A}} \mathbb{Q}(a) = 0.$$

If $D(X||Y) = 0$ and $\mathbb{P}(a) \neq 0$ then it follows that $\ln(\mathbb{P}(a)/\mathbb{Q}(a)) = 0$ for all $a \in \mathbb{A}$. Equivalently, $\mathbb{P}(a) = \mathbb{Q}(a)$ for all $a \in \mathbb{A}$, or $\mathbb{P} = \mathbb{Q}$. If, conversely, $\mathbb{P} = \mathbb{Q}$ then the sum defining $D(X||Y)$ is 0 for all terms $a \in \mathbb{A}$. Hence $D(X||Y) = 0$. □

1.3.17 Remark. With positivity, the relative divergence has one property of a metric. However, the lack of symmetry prohibits it from being a metric.

1.3.18 Corollary. *Suppose the order of set $\mathbb{A}$ is n, and X is a random variable with induced probability measure $\mathbb{P}$ on $\mathbb{A}$. Then $H(X) \leq \ln(n)$ with equality if and only if $\mathbb{P}$ is the uniform distribution.*

Proof. Let $\mathbb{Q}$ have the uniform distribution, so $\mathbb{Q}(a) = 1/n$ for all $a \in \mathbb{A}$. Then $0 \leq D(\mathbb{P}||\mathbb{Q})$ from above and

$$D(\mathbb{P}||\mathbb{Q}) = \sum_{a\in\mathbb{A}} \mathbb{P}(a) \ln\left(\frac{\mathbb{P}(a)}{\mathbb{Q}(a)}\right) = \sum_{a\in\mathbb{A}} \mathbb{P}(a)\ln(\mathbb{P}(a)) + \sum_{a\in\mathbb{A}} \mathbb{P}(a)\ln n.$$

Since $\sum_{a\in\mathbb{A}} \mathbb{P}(a) = 1$ we have

$$0 \leq D(\mathbb{P}||\mathbb{Q}) = \sum_{a\in\mathbb{A}} \mathbb{P}(a)\ln(\mathbb{P}(a)) + \ln(n).$$

Hence $-\ln(n) \leq \sum_{a\in\mathbb{A}} \mathbb{P}(a)\ln(\mathbb{P}(a))$. Equivalently,

$$\ln(n) \geq -\sum_{a\in\mathbb{A}} \mathbb{P}(a)\ln(\mathbb{P}(a)) = H(X).$$

□

1.3.19 Remark. If we have the infinite alphabet $\mathbb{A} = \mathbb{N}$, then we can assume some additional knowledge about X, for example the expected value of X, $\mathbb{E}(X) = \sum_{a=1}^{\infty} a\mathbb{P}(a)$, to get a non-trivial upper bound for the entropy.

1.3.20 Corollary. *Let $X : \Omega \to \mathbb{A}$ be a random variable with induced measure $\mathbb{P}_X$ on $\mathbb{A} = \mathbb{N}$. Let $\mu = \mathbb{E}(X)$. Then Then*

$$H(X) \leq \mu\ln\mu - (\mu-1)\ln(1-\mu) = \mu\left(\frac{-1}{\mu}\ln\left(\frac{1}{\mu}\right) - \left(1-\frac{1}{\mu}\right)\ln\left(1-\frac{1}{\mu}\right)\right)$$

with equality if and only if $\mathbb{P}_X(n) = (1-\alpha)\alpha^{n-1}$ where $\alpha = 1 - \frac{1}{\mu}$.

Proof. Let $\mathbb{Q}(n) = (1-\alpha)\alpha^{n-1}$ with $0 \leq \alpha < 1$, then

$$0 \leq D(\mathbb{P}_X||\mathbb{Q}) = \sum_{n\in\mathbb{N}} \mathbb{P}_X(n)\ln\left(\frac{\mathbb{P}_X(n)}{\mathbb{Q}(n)}\right) = -H(X) + \sum_{n\in\mathbb{N}} \mathbb{P}_X(n)\Big(\ln\left(\frac{\alpha}{1-\alpha}\right) - \ln\alpha^n\Big)$$

$$= -H(X) - \sum_{n\in\mathbb{N}} \mathbb{P}_X(n) n\ln\alpha = -H(X) + \ln\left(\frac{\alpha}{1-\alpha}\right) - \mu\ln\alpha.$$

Hence $H(X) \leq \ln\frac{\alpha}{1-\alpha} - \mu\ln\alpha$. Minimizing the right hand side with respect to α gives the best bound for the choice $\alpha = 1 - \frac{1}{\mu}$, so $H(X) \leq \mu\ln\mu - (\mu-1)\ln(1-\mu)$. Equality holds by the usual argument for relative entropy if and only if $D(\mathbb{P}_X||\mathbb{Q}) = 0$, that is, $\mathbb{P}_X(n) = \mathbb{Q}(n)$ for all $n \in \mathbb{N}$. □

An Aside for Relative Entropy

Given $X_1, X_2, \ldots, X_n$ independent and identically distributed random variables (i.i.d. r.v.'s) with a discrete alphabet $\mathbb{A}$ that are all distributed according to either of the probability measures $\mathbb{P}_X$ or $\mathbb{P}_{\widehat{X}}$, we wish to decide which measure is present via observing their values once. We follow the **Neyman-pearson** hypothesis test strategy.

Set-up

Let the null hypothesis $H_0 = \mathbb{P}_X$ and let $H_1 = \mathbb{P}_{\hat{X}}$. Let $\phi : \mathbb{A}^n \to \{0, 1\}$ be defined by

$$\phi(X_1, X_2, \ldots, X_n) = \begin{cases} 0 & \text{if } H_0 \text{ is accepted} \\ 1 & \text{if } H_0 \text{ is rejected} \end{cases}.$$

The map ϕ has an associated acceptance region for the null hypothesis: $\mathcal{A}_n = \{x \in \mathbb{A}^n : \phi(x) = 0\}$. Similarly, ϕ has an associated acceptance region for H_1: $\mathcal{A}_n^c = \{x \in \mathbb{A}^n : \phi(x) = 1\}$.

Minimizing Error

There exist two types of error: false positive and false negatives.

Type I: $\alpha_n := \mathbb{P}_{X_1, X_2, \ldots, X_n}(\phi(X_1, \ldots, X_n) = 1)$, so H_0 is rejected although it is true.

Type 2: $\beta_n := \mathbb{P}_{\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_n}(\phi(X_1, \ldots, X_n) = 0)$ so H_0 is accepted although H_1 is true.

Neyman-Pearson Testing

For Neyman-pearson testing, having a limit for false positives is the first priority. Thus, we set a threshold for an acceptable rate (α_n) of false positives and then minimize the probability of false negatives (β_n) among all possible choices:

> Given a constant $\epsilon > 0$ and the requirement $\alpha_n \leq \epsilon$, choose Φ such that β_n is minimal.

1.3.21 Theorem. *Given $X_1, \ldots, X_n$ with probability distributions $\mathbb{P}_X$ and $\mathbb{P}_{\hat{X}}$ as above, define the acceptance region for parameter $\tau > 0$ by*

$$\mathcal{A}_n(\tau) = \left\{ x \in \mathbb{A}^n : \frac{\mathbb{P}_{X_1, \ldots, X_n}}{\mathbb{P}_{\hat{X}_1, \ldots, \mathbb{P}_{\hat{X_n}}}} > \tau \right\}$$

and let $\alpha_n(\tau) = \mathbb{P}_{X_1, \ldots, X_n}(\mathcal{A}_n^c(\tau))$. Then if α_n and β_n are associated with another acceptance region $\mathcal{A}_n'$ we have for every $\alpha_n' \leq \alpha_n$ that $\beta_n' \geq \beta_n$.

Proof. Consider the acceptance region $\mathcal{A}_n'$ and let $\tau > 0$ then

$$\alpha_n' + \tau \beta_n' = \sum_{x \in \mathcal{A}_n'^c} \mathbb{P}_{X_1, \ldots, X_n}(x) + \tau \sum_{x \in \mathcal{A}_n'} \mathbb{P}_{\hat{X}_1, \ldots, \hat{X}_n}(x)$$

$$= \sum_{x \in \mathcal{A}_n'^c} \mathbb{P}_{X_1,\ldots,X_n}(x) + \tau \left(1 - \sum_{x \in \mathcal{A}_n'^c} \mathbb{P}_{\hat{X}_1,\ldots,\hat{X}_n}(x) \right)$$

$$= \tau + \sum_{x \in \mathcal{A}_n'^c} (\mathbb{P}_{X_1,\ldots,X_n}(x) - \tau \mathbb{P}_{\hat{X}_1,\ldots,\hat{X}_n}(x)).$$

We observe that $\mathcal{A}_n(\tau) = \{x \in \mathbb{A}^n : \mathbb{P}_{X_1,\ldots,X_n}(x) - \tau \mathbb{P}_{\hat{X}_1,\ldots,\hat{X}_n}(x) > 0\}$. Choosing $\mathcal{A}_n' = \mathcal{A}_n(\tau)$ minimizes the right-hand side, because then the sum only contains nonpostive terms. Consequently, $\alpha_n' + \tau\beta_n' \geq \alpha_n + \tau\beta_n$. That is, $\alpha_n - \alpha_n' \leq \tau(\beta_n' - \beta_n)$. Thus if α_n' is bounded above by α_n then $0 \leq \beta_n' - \beta_n$ and we have $\beta_n' \geq \beta_n$. $\square$

We conclude that the likelihood ratio test is optimal for the purposes of minimizing the probability of false negatives while keeping a fixed upper bound on the probability of false positives.