Distribution of Mutual Information

Marcus Hutter

Istituto Dalle Molle di Studi sull'Intelligenza Artificiale
IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland
marcus@idsia.ch, http://www.idsia.ch/~marcus

NIPS-2001, December 3–8

Consider (Dependent) Random Variables

- p_{ij} = joint probability of (i,j), $i \in \{1,...,r\}$ and $j \in \{1,...,s\}$.
- $p_{i+} = \sum_j p_{ij}$ = marginal probability of i,
- $p_{+j} = \sum_i p_{ij}$ = marginal probability of j.

(In)Dependence of Random Variables $\imath$ and $\jmath$

Widely used measure: Mutual Information (= CrossEntropy)

$$I(\mathbf{p}) \;=\; \sum_{i=1}^{r}\sum_{j=1}^{s} p_{ij}\log\frac{p_{ij}}{p_{i+}p_{+j}}$$

Example Application: Connecting Nodes in Bayesian Nets

Contingency Table

Data:

- n_{ij} = # of times (i,j) occurred.
- $n_{i+} = \sum_j n_{ij}$ = # of times i occurred.
- $n_{+j} = \sum_i n_{ij}$ = # of times j occurred.
- $n = \sum_{ij} n_{ij}$ = size of data set.

$j \setminus i$	1	2	$\cdots$	r
1	n_{11}	n_{12}	$\cdots$	n_{1r}
2	n_{21}	n_{22}	$\cdots$	n_{2r}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
s	n_{s1}	n_{s2}	$\cdots$	n_{rs}

Sample Frequency (Point) Estimate of p_{ij}

$$p_{ij} \approx \hat{p}_{ij} := \frac{n_{ij}}{n}$$

Problems of Point Estimate

- $I(\hat{\mathbf{p}})$ gives no information about its accuracy.
- $I(\hat{\theta}) \neq 0$ can have to origins:
  a true dependency of the random variables $\imath$ and $\jmath$ or
  just a fluctuation due to the finite sample size.

Questions of Interest

What is the probability that

- the true mutual information $I(\mathbf{p})$ is larger/smaller than a given threshold I^*,
- the estimate $I(\hat{\mathbf{p}})$ is (in)consistent with $I(\mathbf{p})=0$,

Baysian Solution: 2nd Order Prior

Change convention to avoid confusion: $p_{ij} \rightsquigarrow \theta_{ij}$.

Prior distribution $p(\theta_{ij})$ for the unknown θ_{ij} on the probability simplex.
(e.g. non-informative Dirichlet prior).

$\Rightarrow$ Posterior: $p(\theta|\mathbf{n}) \propto p(\theta) \cdot \prod_{ij} \theta_{ij}^{n_{ij}}$ (the n_{ij} are multinomially distributed).

$\Rightarrow$ Posterior probability density of the mutual information is:

$$p(I|\mathbf{n}) = \int \delta(I(\theta) - I) p(\theta|\mathbf{n}) d^{rs}\theta$$

Hard to Compute:

$\neg$ Monte Carlo (slow),

$\surd$ Exact (partially possible)

$\neg$ Wild approximation (unreliable)

$\surd$ Systematic expansion in $1/n$ (fast and sufficiently accurate)

Results for I under Dirichlet P(oste)rior

- Exact expression for mean:

$$E[I] = \frac{1}{n}\sum_{ij} n_{ij}[\psi(n_{ij}+1)-\psi(n_{i+}+1)-\psi(n_{+j}+1)+\psi(n+1)], \quad \psi(n) = \sum_{k=1}^{n-1}\frac{1}{k}$$

- Leading and next to leading order (n.l.o.) term for variance:

$$\text{Var}[I] = \frac{1}{n}\sum_{ij}\frac{n_{ij}}{n}\left(\log\frac{n_{ij}n}{n_{i+}n_{+j}}\right)^2 - \frac{1}{n}\left(\sum_{ij}\frac{n_{ij}}{n}\log\frac{n_{ij}n}{n_{i+}n_{+j}}\right)^2 + n.l.o. + O(n^{-3}).$$

- For $n.l.o.$ variance and leading order for skewness and kurtosis (3^{rd} and 4^{th} central moments) **come to my poster** or read the paper.
- Computation time: $O(r\cdot s)$, i.e. as fast as point estimate.
- Sytematic expansion of all moments to arbitrary order possible, but cumbersome.
- Leading order is as exact as one can specify prior knowledge.

Mutual Information Density Example Graph

$$p(I|\mathbf{n}) = \int \delta(I(\theta)-I) \prod_{ij} \theta_{ij}^{n_{ij}-1} \delta(\theta_{++}-1)\, d^{rs}\theta$$