INFORMATION AND PROBABILITY: Part 2

Principle of Maximum Entropy

Suppose discrete variable *x* has the set of possible values $X = \{x_1,...,x_N\} = \bigcup_{n=1}^{N} \{x_n\}$ and we want to choose a probability model for *x*. Let π_f specify that the probability model is *f(x)*, i.e

$$P(x \mid \pi_f) = f(x), \quad \forall x \in X$$

where $f : X \to [0,1]$ is to be established.

Suppose we wish to impose *R* constraints on the probability model *f*(*x*) that are the expected values of *R* functions $g_r(x)$. Let c_R be a proposition specifying the constraints:

$c_R =$ "$\sum_{n=1}^{N} g_r(x_n) f(x_n) = \mu_r, \quad r = 1,...,R,$ are the constraints on probability model f."

e.g. choose *R*=2 and $g_1(x) = x$, $g_2(x) = x^2$ to impose first and second moment constraints.

Principle of Maximum Entropy [Jaynes 1957]: Choose that probability model $\hat{f}$ among all possible probability models $f : X \to [0,1]$ that maximizes the entropy of *x*:

$$H(x \mid \pi_f) = H\left(\{x = x_n\}_1^N \mid \pi_f\right) = -K\sum_{n=1}^{N} f(x_n) \log f(x_n)$$

Note that any other f would give a smaller entropy and so a lesser amount of missing information, implying an information gain compared with $\hat{f}$, without any additional information being utilized.

Applying this principle, we introduce Lagrange multipliers $\lambda_r K, \; r = 0,1,...,R$ and impose stationarity of:

$$J\left[f(x_1),..., f(x_N)\right] \triangleq -K\sum_{n=1}^{N} f(x_n) \log f(x_n) - \lambda_o K\sum_{n=1}^{N} f(x_n) - \sum_{r=1}^{R} \lambda_r K \sum_{n=1}^{N} g_r(x_n) f(x_n)$$

The maximum entropy probability model is:

$$\hat{f}(x) = \exp\left[-\lambda_o - \sum_{r=1}^{R} \lambda_r g_r(x)\right], \quad \forall x \in X$$

which belongs to a family of generalized exponential probability distributions, where

$$\lambda_o = \ln Z(\lambda_1,...,\lambda_R) \quad (\text{from } \sum_{n=1}^{N} f(x_n) = 1)$$

$$Z = \sum_{n=1}^{N} \exp\left[-\sum_{r=1}^{R} \lambda_r g_r(x_n)\right]$$

$$\mu_r = -\frac{\partial}{\partial \lambda_r} \ln Z(\lambda_1,...,\lambda_R), \quad \forall r = 1,...,R$$

It can be shown using Jensen's inequality that $\hat{f}$ does indeed maximize $H\left(x \mid \pi_f\right)$ (e.g. see Cover and Thomas).

The Principle of Maximum Entropy can be stated in an equivalent form as:

Principle of Minimum Relative Entropy: Choose that probability model $\hat{f}$ among all possible probability models $f : X \to [0,1]$ that minimizes the relative entropy of *x*:

$$I\left(x \mid \pi_f / \pi_o\right) = I\left(\left\{x = x_n\right\}_1^N \mid \pi_f / \pi_o\right)$$

subject to all of the constraints, i.e. $\hat{f}$ minimizes the expected information gain coming from imposing the additional *R* constraints given by c_R. Here, π_o denotes the proposition stating that the probability model for x is the uniform probability distribution, $P\left(x \mid \pi_o\right) = \frac{1}{N}, \ \forall x \in X$, with entropy $K \log N = \log_2 N$, which is the maximum entropy value for a variable with *N* possible values.

Proof:

The equivalence of the principles follows from:

$$I\left(x \mid \pi_f / \pi_o\right) = \sum_{n=1}^{N} f\left(x_n\right) \log_2 \frac{f\left(x_n\right)}{1/N} = \log_2 N - H\left(x \mid \pi_f\right)$$

Reference: Shore and Johnson (1980). Axiomatic derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross Entropy, IEEE Trans. on Information Theory, **26**, 26-37.

Mutual Information

Defn: Let proposition a imply that $\left\{b_n\right\}_1^N$, $\left\{d_m\right\}_1^M$ are exhaustive and mutually exclusive, then the mutual information of these sets of propositions is defined by:

$$I\left(\left\{b_n\right\}_1^N, \left\{d_n\right\}_1^N \mid a\right) = \sum_{n=1}^{N} \sum_{m=1}^{M} P\left(b_n, d_m \mid a\right) \log_2 \frac{P\left(b_n, d_m \mid a\right)}{P(b_n \mid a) P\left(d_m \mid a\right)}$$

where $P(b_n \mid a) = \sum_{m=1}^{M} P\left(b_n, d_m \mid a\right), \quad \forall n = 1, ..., N$

$$P(d_m \mid a) = \sum_{n=1}^{N} P\left(b_n, d_m \mid a\right), \quad \forall m = 1, ..., N$$

by the Marginalization Theorem P7(a) (with $c = b_n$ and $c = d_m$, respectively).

By Jensen's inequality:

$$I\left(\left\{b_n\right\}_1^N, \left\{d_n\right\}_1^N \mid a\right) \geq 0$$

with equality occurring only if $P(b_n, d_m \mid a) = P(b_n \mid a) P(d_m \mid a), \ \forall n, m$, i.e the b_n and the d_m are mutually independent.

Mutual information is a fundamental measure of the dependence between two sets of exhaustive and mutually exclusive propositions, or between two variables $x \in \{x_1,...,x_N\}$ and $y \in \{y_1,...,y_M\}$ (using $b_n = "x = x_n"$ and $d_m = "y = y_m"$). It gives the amount of information that sets of propositions, or pair of variables, provide about each other. It is more fundamental than correlation coefficients or covariance matrices.

Further insight is provided by the following alternative form using axiom P4:

$$I\left(\{b_n\}_1^N, \{d_m\}_1^M \mid a\right) = \sum_{m=1}^{M} P\left(d_m \mid a\right) \sum_{n=1}^{N} P\left(b_n \mid d_m, a\right) \log_2 \frac{P\left(b_n \mid d_m, a\right)}{P(b_n \mid a)}$$

$$= \sum_{m=1}^{M} P\left(d_m \mid a\right) I\left(\{b_n\}_1^N \mid d_m / a\right)$$

This shows that the mutual information gives the expected relative entropy of $\{b_n\}_1^N$ for d_m relative to a, so it is a measure of the expected information gain about $\{b_n\}$ from $\{d_m\}$. Similarly,

$$I\left(\{b_n\}_1^N, \{d_m\}_1^M \mid a\right) = \sum_{n=1}^{N} P\left(b_n \mid a\right) I\left(\{d_m\}_1^M \mid b_n / a\right)$$

showing that it is also a measure of the expected gain about $\{d_m\}$ from $\{b_n\}$.

Note that to evaluate the mutual information, proposition a must specify a joint probability model, i.e. specify $P\left(b_n, d_m \mid a\right),\ \forall n, m$. Also, it is easy to show that

$$I\left(\{b_n\}_1^N, \{d_m\}_1^M \mid a\right) = H\left(\{b_n\}_1^N \mid a\right) + H\left(\{d_m\}_1^M \mid a\right) - H\left(\{b_n\}_1^N, \{d_m\}_1^M \mid a\right)$$

where the last term is the joint entropy defined in terms of $P(b_n, d_m \mid a),\ \forall n, m$.
Since the mutual information is never negative:

$$H\left(\{b_n\}_1^N, \{d_m\}_1^M \mid a\right) \le H\left(\{b_n\}_1^N \mid a\right) + H\left(\{d_m\}_1^M \mid a\right)$$

showing that dependence between the sets of propositions reduces the entropy (the uncertainty in which proposition in each set is the one that is true) compared with the case where they are considered separately. The exception occurs when the b_n and the d_m are independent, corresponding to I=0.

References: Some papers applying information-theoretic ideas to dynamical systems:

Vastano, J.A. and Swinney, H.L. (1988). Information transfer in Spatiotemporal Systems. Physical Review Letters, **60**, 1773-1776

Schreiber, T. (2000). Measuring information transfer. Physical Review Letters, **85**, 461-464

Kaiser, A. and Schreiber, T. (2002). Information transfer in continuous processes. Physica D, **166**, 43-62