Information theoretic learning models

November 3, 2010

Motivation: Optimal adaptive filtering $\Rightarrow \mathbf{E}[Xe] = 0$

Uncorrelated is not independent! Consider $X \sim U[-1, 1]$ and $Y = X^2$.

Information: Which has more information?

1. NN project is due today. 2. NN project is not due today.

If two events has probability p and q of occurring, then

1. $I(p), I(q) \geq 0$,
2. $I(1) = 0$,
3. $I(p) < I(q)$ if $p > q$.

Take $I(p) = -\log p$

Entropy: Entropy is a measure of uncertainty.

Let X take values $\{x_1, \ldots, x_k\}$ with probability $\{p_1, \ldots, p_k\}$.

If $p_1 = 1$ and $p_2 = \ldots = p_k = 0$, only one event occur $\Rightarrow$ No uncertainty $\Rightarrow$ Zero entropy

If $p_1 = \ldots = p_k = \frac{1}{k}$, all events are equally probable $\Rightarrow$ Max uncertainty $\Rightarrow$ Max entropy

Desired properties

1. $H(P) = H(p_1, \ldots, p_k)$ is symmetric
2. $H(P)$ is continuous
3. $H(P * Q) = H(P) + H(Q)$ Additivity

Shannon's entropy $H(P) = \sum p_k I(p_k)$

Rényi's entropy

$$H(P) = \frac{1}{1-\alpha} \log \sum p_k^{\alpha}$$

Equivalent to $H(P) = g^{-1}\left(\sum p_k g(I(p_k))\right)$ with $g(x) = 2^{(1-\alpha)x}$.
Shannon's entropy is Rényi's entropy for $\alpha \rightarrow 1$

Note, $0 \log 0 = 0$

Conditional entropy:

1a Dr. Principe is in town next Monday.

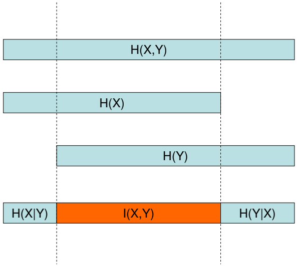
2 Dr. Principe is teaching next Monday.

1b You didn't finish your project.

$$
\begin{aligned}
& H(X|Y) \\
={}& \sum P(X = q_k) H(X|Y = q_k) \\
={}& -\sum P(Y = q_k) \sum P(X = p_j | Y = q_k) \log P(X = p_j | Y = q_k) \\
={}& -\sum_k \sum_j P(X = p_j, Y = q_k) \log \frac{P(X = p_j, Y = q_k)}{P(Y = q_k)} \\
={}& -\sum_k \sum_j P(X = p_j, Y = q_k) \log \frac{P(X = p_j, Y = q_k)}{P(X = p_j) P(Y = q_k)} + H(X) \\
={}& -MI(X, Y) + H(X)
\end{aligned}
$$

MI is mutual information!

Mutual information is zero $\Leftrightarrow$ Random variables are independent



Differential entropy: $H(X) = \int f_X(x) \log f_X(x) \mathrm{d}x$

$H(X + c) = H(X)$ and $H(aX) = H(X) + \log |a|$

If $X \sim \mathcal{U}[0, 1], H(X) = 0$ and if $X = c, H(X) = -\infty$

Conditional entropy: $H(X|Y)$

Mutual information:

$$MI(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(X) - H(X, Y$$

$$MI(X,Y) = \iint f_{XY}(x,y) \log \frac{f_{XY}(x,y)}{f_X(x) f_Y(y)} dxdy$$

MI is nonnegative! MI is invariant to invertible transformation.

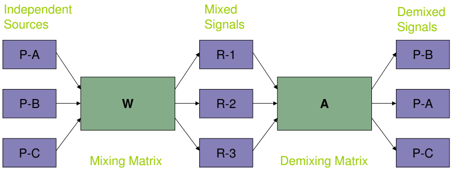
InfoMax: Train a network such that the mutual information $I(X, Y)$ between input X and Y is maximized.

$I(X, Y) = H(Y) - H(Y|X)$

Information theoretic learning works with other "forms" of MI.

$$QMI(X, Y) = \iint (f_{XY}(x, y) - f_X(x) f_Y(y))^2 dxdy$$

ICA



Assumptions: Mutual independence of sources, square mixing matrix, noise free model, zero mean, unit covariance.

InfoMax: Train a network such that the mutual information $I(X, Y)$ between input X and Y is maximized.

$I(X, Y) = H(Y) - H(Y|X)$.

If $Y = G(X) + N$ then $H(Y|X) = H(N)$ i.e. maximizing $I(X, Y)$ implies maximizing $H(Y)$

$$Y = \frac{1}{1 + \exp(-(aX + b))}$$

$$f_Y(y) = \frac{f_X(x)}{\left|\frac{\partial y}{\partial x}\right|}$$

$$\Rightarrow H(Y) = \mathbf{E}\left[\log\left|\frac{\partial y}{\partial x}\right|\right] + H(X)$$

Stochastic gradient rule

$$\Delta a \propto \frac{1}{a} + x(1 - 2y)$$

$$\Delta b \propto 1 - 2y$$

Multivariate case

$$\Delta A \propto \left[A^\top\right]^{-1} + (1 - 2\mathbf{y})\mathbf{x}^\top$$
$$\Delta B \propto 1 - 2\mathbf{y}$$

$\left[A^\top\right]^{-1}$ avoids redundancy

In the context of ICA

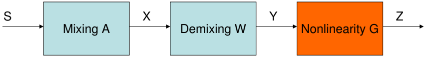
$$I(y_1, y_2) = H(y_1) + H(y_2) - H(y_1, y_2)$$

i.e. maximizing $H(y_1, y_2)$ implies minimizing $I(y_1, y_2)$

What happen to the individual entropies?

Choose nonlinearity such that it matches the source pdf.

InfoMax Bell & Sejnowski



$$f_Z(z) = \frac{f_S(s)}{\left|\frac{\mathrm{d}z}{\mathrm{d}y}\right|}$$

where

$$\left|\frac{\mathrm{d}z}{\mathrm{d}y}\right| = DWA$$

where $D = \mathrm{diag}\left(\frac{\partial z_1}{\partial y_1}, \frac{\partial z_2}{\partial y_2}\right)$

$$H(Z) = H(S) - \left[\log|A| + \log|W| + \sum_{i=1}^{2} \log\left(\frac{\partial z_i}{\partial y_i}\right)\right]$$

If $g_i(y_i) = 1/(1 + e^{-y_i})$

$$\frac{\partial H(Z)}{\partial W} = W^{-\top} + (1 - 2z)x^{\top}$$