Neural Network Interest Group

Título/*Title*:
Information Theoretical Learning

Autor(es)/*Author(s)*:
Jorge M. Santos

Relatório Técnico/*Technical Report* No. 4 /2003

FEUP/INEB, Rua Dr. Roberto Frias, 4200-465, Porto, Portugal

Título/*Title*:

Information Theoretical Learning

Autor(es)/*Author(s)*:

Jorge M. Santos

Relatório Técnico/*Technical Report* No. 4 /2003

Publicado por/*Published by*: NNIG. http://paginas.fe.up.pt/~nnig/

Outubro de 2003

Information Theoretic Learning

Jorge M. Santos

October 10, 2003

Abstract

We make were a brief introduction to some entropy principles and information theoretic learning and also an overview of the Unsupervised Learning with Renyi´s Quadratic Entropy.

Contents

1. *Entropy*

Consider the Hartley's measure that defines the amount of information associated with the measurement of an equally likely event x which occurs with probability p as:

$$I(x) = \log \frac{1}{p}.$$

Shannon defined the entropy as the expectation of $I(p_k)$,

$$H(x) = \sum_{k=1}^{N} p_k I(p_k), \qquad H(x) = \sum_{k=1}^{N} p_k \frac{1}{\log p_k}, \qquad H(x) = -\sum_{k=1}^{N} p_k \log p_k$$

i.e the entropy measures de average amount of information conveyed by the event x. The more uncertain the event x, the larger is its information content which can be measured by its entropy.

2. Joint Entropy and Conditional Entropy

Let's extend the definition of entropy to a pair of random variables. The pair (x, y) can be considered to be a single vector-valued random variable.

Definition: The *joint entropy* $H(x, y)$ of a pair of discrete random variables (x, y) with a joint distribution $p(x, y)$ is defined as

$$H(x, y) = -\sum_i \sum_j p(x_i, y_j) \log p(x_i, y_j)$$

We define the conditional entropy of a random variable given another as the expected value of the entropies of the conditional distributions, averaged over the conditioning random variable.

Definition: The *conditional entropy* $H(y \mid x)$ is

$$\begin{aligned} H(y \mid x) &= -\sum_i p(x_i) H(y \mid x = x_i) \\ &= -\sum_i p(x_i) \sum_j p(y_j \mid x_i) \log p(y_j \mid x_i) \\ &= -\sum_i \sum_j p(x_i, y_j) \log p(y_j \mid x_i) \end{aligned}$$

We can prove that the entropy of a pair of random variables is the entropy of one plus the conditional entropy of the other.

$$H(x, y) = H(x) + H(y \mid x)$$

$$
\begin{aligned}
H(x,y) &= -\sum_i \sum_j p(x_i, y_j) \log p(x_i, y_j) \\
&= -\sum_i \sum_j p(x_i, y_j) \log p(x_i) p(y_j \mid x_i) \\
&= -\sum_i p(x_i) \log p(x_i) - \sum_i \sum_j p(x_i, y_j) \log p(y_j \mid x_i) \\
&= H(x) + H(y \mid x)
\end{aligned}
$$

We can easily see that $H(y \mid x) \neq H(x \mid y)$. However, $H(x) - H(x \mid y) = H(y) - H(y \mid x)$.

3. Relative Entropy and Mutual Information

The relative entropy is a measure of the distance between two distributions.

Definition: The *relative entropy*, or the *Kullback-Leibler divergence measure,* $D(p(x);q(x))$ is

$$
D(p(x);q(x)) = \sum_i p(x_i) \log \frac{p(x_i)}{q(x_i)} \qquad \textbf{Eq. 1}
$$

The relative entropy measures the penalty of using a wrong statistical model (i.e. using $q(x)$ when the true model is $p(x)$). The relative entropy is always non-negative, is zero only if $p(x) = q(x)$, and since it is not symmetric it is not a true distance measure.

Entropy measures the amount of information required on the average to describe a random event or message. More generally we may be interested in quantifying the amount of information between joint events. For instance, we may be interested in quantifying the degree of uncertainty in the input *x* of a noisy system after observing its output *y*.

Definition: The *mutual information* $I(x,y)$ is the relative entropy between the joint distribution and the product of the marginal distributions.

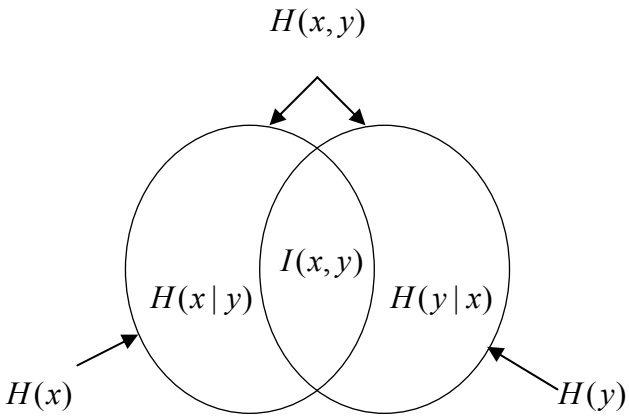
$$
\begin{aligned}
I(x,y) &= D(p(x,y); p(x)p(y)) \\
&= \sum_i \sum_j p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \\
&= \sum_i \sum_j p(x_i, y_j) \log \frac{p(x_i \mid y_j)}{p(y_j)} \\
&= H(x) - H(x \mid y) \\
I(x,y) &= H(y) - H(y \mid x)
\end{aligned}
$$

$$
I(x,x) = H(x) - H(x \mid x) = H(x) \qquad \text{and} \qquad I(y,y) = H(y)
$$

Since $H(y \mid x) = H(x,y) - H(x)$ than:

$$I(x,y) = H(x) + H(y) - H(x,y)$$

The relationship between $H(x)$, $H(y)$, $H(x,y)$, $H(x \mid y)$, $H(y \mid x)$ and $I(x,y)$ can be expressed in the following diagram. The mutual information corresponds to the intersection of the information in x with the information in y.



All the above definitions have been presented for the case of discrete random events. The definitions can be extended to the case of continuous random variables by substituting sums with integrals (as long as the integrals exist). For instance, differential entropy $h(x)$ is defined as

$$h(x) = \int_D P(x) \log P(x)\, dx$$

where $P(x)$ is the probability density function (PDF) for the continuous random variable $x \in D$.

4. Entropy Optimization Principles

The most common entropy optimization principles involve the K-L divergence with respect to the uniform target distribution (*q(x)* in Eq. 1 is set as the uniform distribution). Let us assume that *y* is a function of some parametric mapper ***y*** = *g(**x**,**w**),* where *g(.)* is the mapping, ***x*** the input vector, ***y*** the output vector, and ***w*** the adjustable parameters. So by analogy to optimization in Euclidean space, we can adapt the parameters ***w*** by manipulating *p(y)* and minimize the K-L distance to find de Kullback's Minimum cross-entropy (MinxEnt), or maximize the K-L divergence to find de Jayne's maximization of entropy (MaxEnt).
Examples of this optimization principle are the work of Bell & Sejnowski on blind source separation and the work of Barlow and Attick in neural networks. Linsker also used the information principle in neural networks. He used a linear network assuming that the output was Gaussian distributed as well as the noise. The problem seems to become more complex if we work with arbitrary distributions and nonlinear networks.

One of the difficulties of the application of these information-theoretic criteria is that analytic solutions are known only for very restricted cases, e.g. Gaussianity and linear mappings. Otherwise mathematical approximations and computationally complex algorithms result.
The two fundamental issues in the application of information-theoretic criteria to neurocomputing are: the choice of the criterion for the quantitative measure of information, and the estimation of the probability density function from data samples.

5. Information-Theoretic Learning

(This is an overview of Principe's article in Information-Theoretic Learning)

The concepts of entropy and mutual information are all that is needed to pose and solve optimization problems with information theoretic criteria. Consider a parametric mapping $g : \Re^K \to \Re^M$, $M < K$ of a random vector $\mathbf{x} \in \Re^K$, which is described by the following equation

$$\mathbf{y} = g(\mathbf{x}, \mathbf{w}) \qquad \textbf{Eq. 2}$$

where $\mathbf{y}$ is also a random vector $\mathbf{y} \in \Re^M$, and $\mathbf{w}$ is a set of parameters. The goal is to choose the parameters $\boldsymbol{w}$ of the mapping g(.) such that a figure of merit based on IT is optimized at the output space of the mapper (Figure 1).

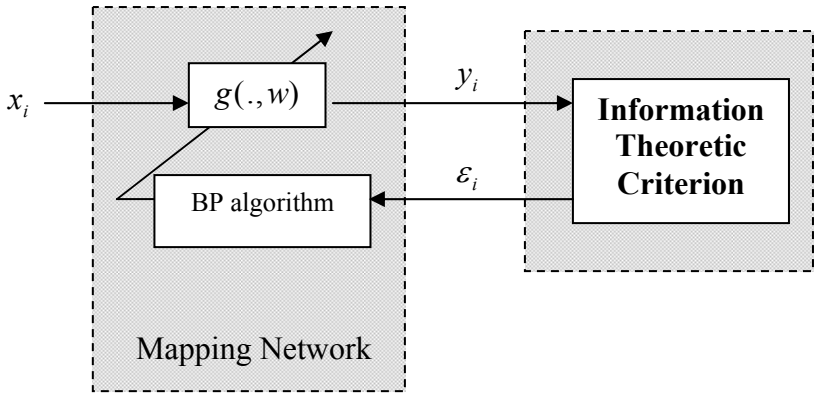


Figure 1: Training a mapper (linear or non linear) with ITL

This is what we call information-theoretic learning (ITL). Notice that we are only requiring the availability of observations x_i and y_i of random vectors without assuming any a priori knowledge about their probability density functions. Notice also that the mapper can either be linear or non-linear, and that the criterion may or may not exploit an external input normally called the desired response, i.e. information theoretic learning spans both the unsupervised and supervised frameworks. We also want the method to be general purpose and not developed for a single application.

5.1 Probability Density Function Estimation

One obstacle of using information theoretic criteria (entropy or mutual information) is that information measures are a weighted sum of the logarithm of the PDF for discrete random variables (or an integral function of the logarithm of the PDF of continuous random variables). Since we can not work directly with the PDF (unless assumptions are made about its form), we rely on nonparametric estimators. Density estimation is an ill-posed problem, and in particular nonparametric density estimation is very unreliable in high dimensional spaces. The approach described here, however, relies on such estimates in the output space of a nonlinear mapper, where the dimensionality is under control of the designer, and is generally manageable.
Principe uses the Parzen window method. The Parzen estimator is a kernel based estimator, which estimates the PDF, $f_Y(\mathbf{y})$, of a random vector $\mathbf{Y} \in \Re^M$ at a point $\mathbf{y}$ as

$$\hat{f}_Y(\mathbf{y}, \mathbf{a}) = (\frac{1}{N}) \sum_{i=1}^{N} \kappa(\mathbf{y} - a_i) .$$

The vectors $\mathbf{a}_i \in \Re^M$ are observations of the random vector and $\kappa(\ )$ is a kernel function which itself satisfies the properties of PDFs. The Parzen window can be viewed as a convolution of the estimator kernel with the observations. Principe says we can choose the symmetric Gaussian kernel

$$\kappa(\mathbf{y}) = G(\mathbf{y}, \sigma^2 \mathbf{I}) = \frac{1}{(2\pi)^{\frac{M}{2}} \sigma^M} \exp(-\frac{\mathbf{y}^T \mathbf{y}}{2\sigma^2})$$

with covariance matrix $\sigma^2 \mathbf{I}$ since we require that $\kappa(\ )$ be differentiable everywhere.

5.2 Unsupervised Learning with Renyi´s Quadratic Entropy

One of Principe's proposed methods to achieve information-theoretic learning is to use the Renyi´s Quadratic Entropy.

In chapter 1 we have defined the information associated with an event $I(x) = \log\frac{1}{p}$.

In the general theory of means the mean of the real numbers $x_1,...,x_n$ with weights $p_1,...,p_n$ has the form:

$$\overline{x} = \varphi^{-1}\left(\sum_{k=1}^{n} p_k \varphi(x_k)\right)$$

where $\varphi(x_k)$ is the Kolmogorov-Nagumo function, which is an arbitrary continuous and strictly monotonic function defined on the real numbers. So, in general, an entropy measure satisfies:

$$H = \varphi^{-1}\left(\sum_{k=1}^{n} p_k \varphi(I(pk))\right)$$

and $\varphi(I(pk))$ is a measure of information. By being a measure of information $\varphi(\ )$ can not be arbitrary since information is "additive". To meet the additivity condition $\varphi(\ )$ can be either $\varphi(x) = x$ or $\varphi(x) = 2^{(1-\alpha)x}$. If the first is used we get the Shannon's entropy. If the second is used we get the Renyi's entropy with order α, which we denote by $H_{R\alpha}$

$$H_{R\alpha} = \frac{1}{1-\alpha}\log\left(\sum_{K=1}^{n} P_K^{\alpha}\right).$$

When $\alpha = 2$,

$$H_{R2} = -\log\left(\sum_{K=1}^{n} P_K^2\right) \text{ is called the Quadratic entropy.}$$

For the continuous random variable $\mathbf{Y}$ with PDF $f_Y(\mathbf{y})$, we can obtain the differential version of Renyi´s quadratic entropy:

$$H_{R2}(Y) = -\log\left(\int_{-\infty}^{+\infty} f_Y(y)^2\, dy\right)$$

5.2.1 Integration of Products of Gaussian Kernels

Let $G(y,\Sigma) = \frac{1}{(2\pi)^{\frac{M}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}y^T\Sigma^{-1}y\right)$ be the Gaussian kernel in M dimensional space, where Σ is the covariance matrix, $y \in R^M$. Let $a_i \in R^M$ and $a_j \in R^M$ be two

data points in the space, Σ_1 and Σ_2 be two covariance matrices for two Gaussian kernels in the space, then it can be shown that the following relation holds:

$$\int_{-\infty}^{+\infty} G(y-a_i,\Sigma_1)G(y-a_j,\Sigma_2)dy = G((a_i-a_j),(\Sigma_1+\Sigma_2))$$

The previous equation can also be interpreted as a convolution between two Gaussian kernels centered at a_i and a_j and it is easy to see that the result should be Gaussian function with a covariance equal to the sum of the individual covariances and centered at $(a_i - a_j)$.

5.2.2 Quadratic Entropy Cost Function for Discrete Samples

Let $a_i \in R^M, i = 1,...,N$, be a set of samples from a random variable $Y \in R^M$ in M-dimensional space. An interesting question is what will be the entropy associated with this set of data samples, without imposing any assumptions about the PDF. Part of the answer lies in the methodology of estimating the data PDF by the Parzen window method using a Gaussian kernel:

$$\hat{f}_Y(\mathbf{y},\mathbf{a}) = \frac{1}{N}\sum_{t=1}^{N} G(\mathbf{y}-\mathbf{a}_i,\sigma^2\mathbf{I})$$

where $G(\ ,\ )$ is the Gaussian kernel as above and $\sigma^2\mathbf{I}$ is the covariance matrix.
If we use Shannon entropy along with Parzen estimation, an algorithm to estimate entropy becomes unrealistically complex. Renyi's quadratic entropy leads to a much simpler form. Using the last two equations we obtain an entropy estimator for a set of discrete data points $\{\mathbf{a}_i\}$ as

$$\begin{cases} H(\{\mathbf{a}_i\}) = H_{R2}(Y \mid \{\mathbf{a}_i\}) = -\log\left(\int_{-\infty}^{+\infty} f_Y(y)^2\,dy\right) = -\log V(\{\mathbf{a}_i\}) \\ V(\{\mathbf{a}_i\}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\int_{-\infty}^{+\infty} G(\mathbf{y}-\mathbf{a}_i,\sigma^2 I)G(\mathbf{y}-\mathbf{a}_j,\sigma^2 I)dy = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G(\mathbf{a}_i-\mathbf{a}_j,2\sigma^2 I) \end{cases} \quad \textbf{Eq. 3}$$

The combination of Renyi's quadratic entropy with the Parzen window leads to an estimation of entropy by computing interactions among pairs of samples which is a practical cost function for ITL. The only approximation in this evaluation is the PDF estimation.

5.2.3 Quadratic Entropy and Information Potential

We wrote Eq. 3 in this way because there is a very interesting physical interpretation for this estimator of entropy. Let us assume that we place physical particles in the locations prescribed by a_i and a_j. Since is always positive and is inversely proportional to the distance between the particles, we can consider that a potential field was created in the

space of interaction with a field strength dictated by the Gaussian kernel, i.e. an exponential decay with the distance square. Physical particles interact with an inverse of distance rule, but Renyi's quadratic entropy with the Gaussian kernel imposes a different interaction law (which by the way is controlled by the kernel utilized in the Parzen estimator).
Now $V(\{\mathbf{a}_i\})$, which is the sum of all pairs of interactions, can be regarded as an overall potential energy of the data set. *We will call this potential energy an information potential*. So maximizing entropy becomes equivalent to minimizing information potential. The quadratic entropy is the negative logarithm of the information potential, so it measures the density of samples throughout the space. This procedure resembles the world of interacting physical particles which originated the concept of entropy.
We can also see from

$$\begin{cases} H(\{\mathbf{a}_i\}) = H_{R2}(Y \mid \{\mathbf{a}_i\}) = -\log\left(\int_{-\infty}^{+\infty} f_Y(y)^2\, dy\right) = -\log V(\{\mathbf{a}_i\}) \\ V(\{\mathbf{a}_i\}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\int_{-\infty}^{+\infty} G(\mathbf{y}-\mathbf{a}_i,\sigma^2 I)G(\mathbf{y}-\mathbf{a}_j,\sigma^2 I)dy = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G(\mathbf{a}_i-\mathbf{a}_j,2\sigma^2 I) \end{cases} \quad \text{Eq. 3}$$

that the Parzen window method implemented with the Gaussian kernel and coupled with Renyi's entropy of higher order $(\alpha > 2)$ will compute interactions among α-tuples of samples, providing even more information about the complex structure of the data set.

5.2.4 Information Forces

Just like in mechanics, the derivative of the potential energy is a force, in this case an information driven force, that moves the data samples in the space of the interactions. Therefore,

$$\frac{\partial}{\partial a_i} G(\mathbf{a}_i-\mathbf{a}_j,2\sigma^2 I) = G(\mathbf{a}_i-\mathbf{a}_j,2\sigma^2 I)(\mathbf{a}_j-\mathbf{a}_i)/(2\sigma^2)$$

can be regarded as the force that a particle in the position of sample a_j impinges upon a_i, and *will be called an information force.* If all the data samples are free to move in a certain region of the space, then the information forces between each pair of samples will drive all the samples to a state with minimum information potential. If we add all the contributions of the information forces from the ensemble of samples on a_i we have the net effect of the information potential on sample a_i, i.e.

$$\frac{\partial}{\partial a_i} V(\{\mathbf{a}_i\}) = \frac{1}{N^2}\sum_{j=1}^{N} G(\mathbf{a}_i-\mathbf{a}_j,2\sigma^2 I)(\mathbf{a}_j-\mathbf{a}_i)/(\sigma^2) \quad \text{Eq. 4}$$

5.2.5 "Force" Back-Propagation

The concept of information potential creates a criterion for ITL, which is external to the mapper of Figure 1. The only missing step is to integrate the criterion with the adaptation of a parametric mapper as the MLP. Suppose the data samples $\{\mathbf{a}_i\}$ are the

outputs of our parametric mapper of Eq. 2, $\{\mathbf{y}_i \in R^M, i = 1,...,N\}$. If we want to adapt the MLP such that the mapping maximizes the entropy at the output $H(\{\mathbf{y}_i\})$, the problem is to find the MLP parameters $\{w_{ij}\}$ so that the information potential $V(\{\mathbf{y}_i\})$ is minimized. In this case, the data samples are not free but are a function of the MLP parameters. So, the information forces applied to each data sample by the information potential can be back-propagated to the parameters using the chain rule, i.e.

$$\frac{\partial}{\partial w} V(\{\mathbf{y}_i\}) = \sum_{1=1}^{N} \left[\frac{\partial}{\partial \mathbf{y}_i} V(\{\mathbf{y}_i\}) \right]^T \frac{\partial \mathbf{y}_i}{\partial w} = \sum_{1=1}^{N} \varepsilon_i^T \frac{\partial}{\partial w} g(\mathbf{w}, \mathbf{x}_i) \qquad \textbf{Eq. 5}$$

where $\mathbf{y}_i = (y_{i1},..., y_{iM})^T$ is the M dimensional MLP output. The quantity

$$\frac{\partial}{\partial y_i} V(\{\mathbf{y}_i\}) = \left(\frac{\partial}{\partial y_{i1}} V(\{y_i\}),..., \frac{\partial}{\partial y_{1M}} V(\{y_i\}) \right)^T \qquad \textbf{Eq. 6}$$

is the information force given by Eq. 4. Notice that from Eq. 5, the sensitivity of the output with respect to a MLP parameter $\frac{\partial \mathbf{y}_i}{\partial w}$ is the "*transmission mechanism" through which information forces are back-propagated to the parameter* (Figure 2). From the analogy of Eq. 5 with the backpropagation formalism we conclude that *information forces take the place of the injected error in the backpropagation algorithm*. So, we obtain a general, nonparametric, and samplebased methodology to adapt arbitrary nonlinear (smooth and differentiable) mappings for entropy maximization. Notice that we are adapting a MLP without a desired response. We have established an ITL criterion that adapts the MLP with a global property of its output sample distribution.

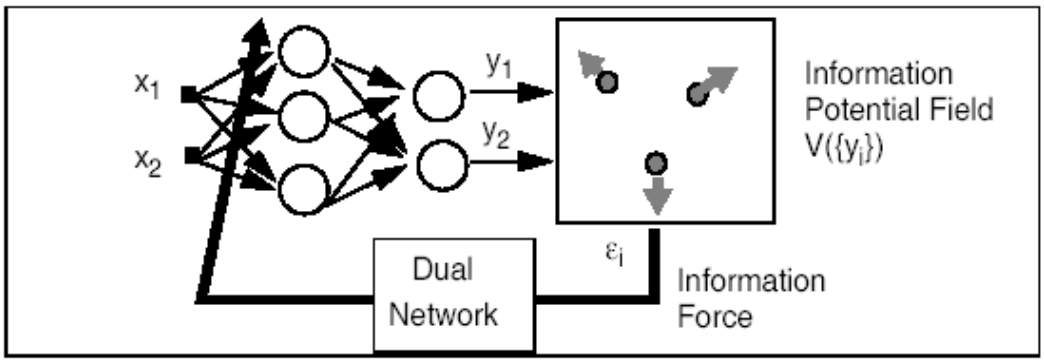


Figure 2: Training a MLP with the information potential

Principe states that the methodology presented here lays down the framework to construct an "*entropy machine*", i.e. a learning machine that is capable of estimating entropy directly from samples in its output space, and can through backpropagation modify its weights to minimize or maximize output entropy. The algorithm has complexity $O(N^2)$ since the criterion needs to examine the interactions among all pairs of output samples.

Principe extends Bell and Sejnowski approach to Independent Component Analyses (ICA). Bell's approach can not be easily extended to MLPs nor to data distributions which are multimodal in nature. In this approach he has decoupled the mapper from the criterion and the optimization problem, *Renyi'squadratic entropy becomes essentially a general purpose criterion as widely applicable as the MSE.*

6. References

Príncipe J., Fisher J., Xu D., "Information-Theoretic Learning", CNEL, University of Florida, 1998

Príncipe J., Xu D., "Information-Theoretic Learning Using Renyi´s Quadratic Entropy", CNEL, University of Florida, 1999

Cover T., Thomas G, "Elements of Information Theory", Wiley, 1991