Infomax: information theory lecture 11

COMSM0075 Information Processing and Brain

`comsm0075.github.io`

October 2020

Source separation

$$\mathsf{s} \xrightarrow{\text{mixing}} \mathsf{r} = M\mathsf{s} \xrightarrow{\text{unmixing}} \mathsf{x} = W\mathsf{r}$$

Mutual information

$$\mathrm{s} \xrightarrow{\text{mixing}} \mathrm{r} = M\mathrm{s} \xrightarrow{\text{unmixing}} \mathrm{x} = W\mathrm{r}$$

Two-dimensional case: we are assuming the two sources S_1 and S_2 are independent, so we want to find independent X_1 and X_2.

Mutual information

$$\mathbf{X} = W\mathbf{R}$$

Two-dimensional case: we are assuming the two sources S_1 and S_2 are independent, so we want to find independent X_1 and X_2:

$$I(X_1, X_2) = 0$$

or at the very least we'll try to minimize $I(X_1, X_2)$.

Infomax

We want to minimize $I(X_1, X_2)$ but this is very hard to calculate!

$$I(X_1, X_2) = h(X_1) + h(X_2) - h(X_1, X_2)$$

Let's maximize $h(X_1, X_2)$ instead.

Infomax

We want to minimize $I(X_1, X_2)$ but this is very hard to calculate!

$$I(X_1, X_2) = h(X_1) + h(X_2) - h(X_1, X_2)$$

Let's maximize $h(X_1, X_2)$ instead: Infomax.

Infomax

We want to minimize $I(X_1, X_2)$ but this is very hard to calculate!

$$I(X_1, X_2) = h(X_1) + h(X_2) - h(X_1, X_2)$$

Let's maximize $h(X_1, X_2)$ instead.

- This means ignoring $h(X_1)$ and $h(X_2)$.
- It isn't obvious $h(X_1, X_2)$ is any easier to calculate than $I(X_1, X_2)$.

An obvious problem

The differential entropy isn't scale invariant

$$h(\lambda X_1, X_2) = h(X_1, X_2) + \log_2 |\lambda|$$

so it tells us nothing about mixing and unmixing.

An obvious problem

The differential entropy isn't scale invariant

$$h(X_1, \lambda X_2) = h(X_1, X_2) + \log_2 |\lambda|$$

so it tells us nothing about mixing and unmixing.

An obvious problem

The differential entropy isn't scale invariant

$$h(\lambda X_1, \lambda X_2) = h(X_1, X_2) + 2\log_2|\lambda|$$

so it tells us nothing about mixing and unmixing.

A very clever solution

Inspire by the behaviour of neurons Bell and Sejnowski added a saturating non-linearity:

$$\begin{aligned} y_1 &= g(x_1 + w_1) \\ y_2 &= g(x_2 + w_2) \end{aligned}$$

where w_1 and w_2 are parameters and, for example,

$$g(u) = \frac{1}{1 + e^{-u}}$$

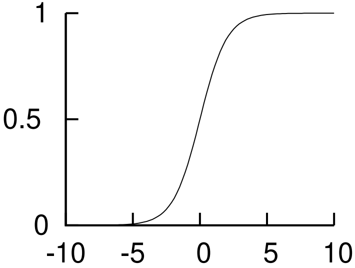
is a saturating non-linearity so $g : (-\infty, \infty) \rightarrow (0, 1)$.

Bell, A. J. and Sejnowski, T. J. (1995)

Saturating non-linearity

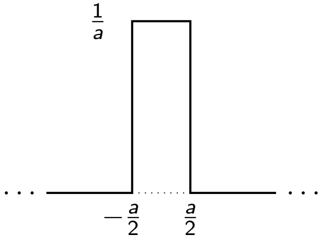
$$g(u) = \frac{1}{1 + e^{-u}}$$

is a saturating non-linearity so $g : (-\infty, \infty) \to (0, 1)$.



Saturating non-linearity

Say X is uniform



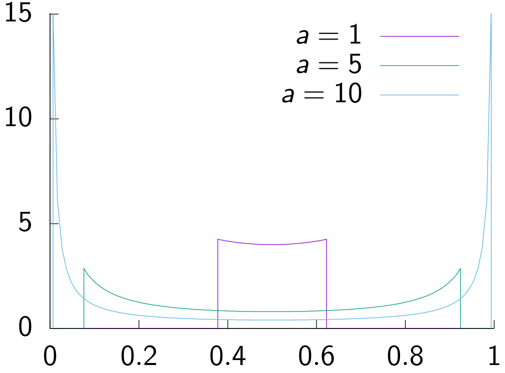
Saturating non-linearity

Now calculate

$$p_G(g) = \frac{p_X(x(g))}{dg/dx}$$

Saturating non-linearity

Now calculate $p_G(g)$



Saturating non-linearity

Now calculate $p_G(g)$

$$\begin{array}{ll} a = 1 & h(G) = -1.41 \\ a = 5 & h(G) = -0.26 \\ a = 10 & h(G) = -1.03 \\ a = 15 & h(G) = -11.6 \end{array}$$

Unmixing

$$\mathsf{s} \xrightarrow{\text{mixing}} \mathsf{r} = M\mathsf{s} \xrightarrow{\text{unmixing}} \mathsf{x} = W\mathsf{r} \xrightarrow{\text{non-linearity}} \mathsf{y} = g.(\mathsf{x} + \mathsf{w})$$

using the broadcast notation

$$g.(\mathsf{x} + \mathsf{w}) = (g(x_1 + w_1), g(x_2 + w_2))$$

Unmixing

$$\mathsf{s} \xrightarrow{\text{mixing}} \mathsf{r} = M\mathsf{s} \xrightarrow{\text{unmixing}} \mathsf{x} = W\mathsf{r} \xrightarrow{\text{non-linearity}} \mathsf{y} = g.(\mathsf{x} + \mathsf{w})$$

For later convenience:

$$\mathsf{y} = g.(\mathsf{x} + \mathsf{w}) = \mathsf{f}(\mathsf{r}; W, \mathsf{w})$$

One-dimensional problem

$$r \xrightarrow{\text{multiply}} x = Wr \xrightarrow{\text{non-linearity}} y = g(x + w) = f(r; w, W)$$

where W and w are both scalars. We want to maximize the entropy $h(Y)$, this should also maximize the information in Y about R:

$$I(R; Y) = h(Y) - h(Y|R)$$

but $h(Y|R)$ is constant since R determines Y.

Don't panic

$$I(R;Y) = h(Y) - h(Y|R)$$

and $h(Y|R)$ is constant since R determines Y. For differential entropy the constant is minus infinity not zero as it would be for discrete entropy, but since we are interested in derivative all that counts is that it's a constant!

Estimating $h(Y)$

As we know

$$h(Y) = -\int p(y) \log p(y) dy$$

and this is estimated by

$$\tilde{h}(y) = -\log p(y)$$

meaning if n values y^t are drawn from Y then

$$\frac{1}{n} \sum_t \tilde{h}(y^t) \to h(Y)$$

as n gets large.

We don't need $p_Y(y)$

The plan is to maximize $h(Y)$ by gradient ascent.

Picture of Pendle from Wikipedia

We don't need $p_Y(y)$

We can't estimate $h(Y)$ because we don't have $p_Y(y)$, but it turns out we can still calculate the derivative.

$$p_Y(y) = \frac{p_R[r = f^{-1}(y)]}{|f'(f^{-1}(y)|}$$

so

$$\tilde{h}(y) = \tilde{h}(r) + \log |f'|$$

and $p_R(r)$ is independent of the parameters.

We can do the calculation

We want dh/dW and dh/dw.

$$\begin{aligned} g(u) &= \frac{1}{1+\exp(-u)} \\ \frac{dg}{du} &= g(1-g) \end{aligned}$$

and hence

$$\log|f'| = \log W + \log f + \log(1-f)$$

We continue doing the calculation

$$f = g(Wr + w)$$

so

$$\frac{df}{dW} = rf(1-f)$$

and hence,

$$\frac{d\tilde{h}(y)}{dW} = \frac{1}{W} + \frac{1}{f}rf(1-f) - \frac{1}{1-f}rf(1-f) = \frac{1}{W} + r(1-2y)$$

Similarly

$$\frac{d\tilde{h}(y)}{dw} = 1 - 2y$$

This is all stuff we know

$$\frac{d\tilde{h}(y)}{dW} = \frac{1}{W} + r(1-2y)$$

and

$$\frac{d\tilde{h}(y)}{dw} = 1-2y$$

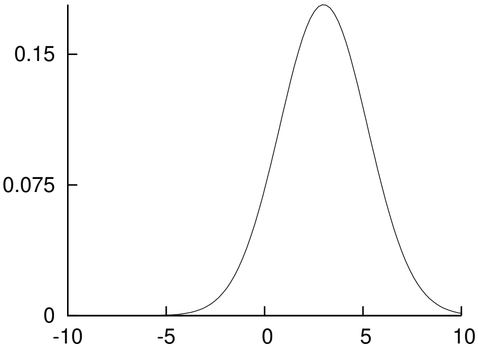
r is the recorded value which we can sample; W and w are the variables we want to work out.

We can do hill-climbing

If we have the derivatives we can use an hill-climbing algorithm like steepest ascent, conjugate gradient or metric gradient; the latter works particularly well here.

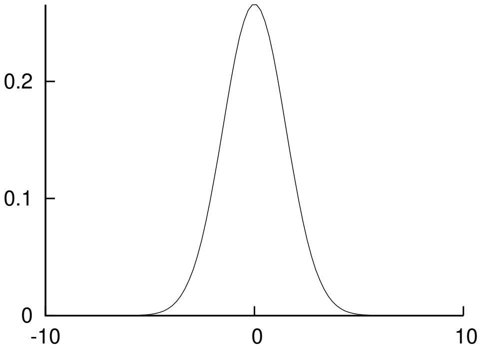
Example

Initial distribution $p_R(r)$:



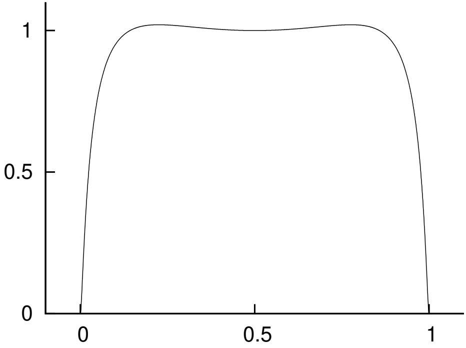
Example

After $u = Wr + w$ we have $p_U(u)$:



Example

The non-linearity give $y = g(Wr + w)$ we have $p_Y(y)$:



Back to the 2×2 case

$$\mathrm{s} \xrightarrow{\text{mixing}} \mathrm{r} = M\mathrm{s} \xrightarrow{\text{unmixing}} \mathrm{x} = W\mathrm{r} \xrightarrow{\text{non-linearity}} \mathrm{y} = g.(\mathrm{x} + \mathrm{w})$$

A similar calculation gives

$$\begin{aligned} \frac{d\tilde{h}(\mathrm{y})}{dW_{ab}} &= (W^T)^{-1}_{ab} + r_a(1 - 2y_b) \\ \frac{d\tilde{h}(\mathrm{y})}{dw_a} &= 1 - 2y_a \end{aligned}$$

allowing use to maximize $h(Y_1, Y_2)$. This is Infomax!