About information theoretic optimum distribution comparision

Tomas Ukkonen

First we calculate entropy by calculating log likelihood of seen data. Assume we have a source of data and each symbol has probability of $p(x_i) = p_i$. Then randomly chosen geometric mean likelihood of data (N symbol long string) is:

$$P = (\prod\ p_j p_k ... p_l)^{1/N}$$

Now assume each symbol, on average, exists on a string $p_j N$ times when N is large.

$$P = \left(\prod\ p_1^{p_1 N} p_2^{p_2 N} ... p_M^{p_M N}\right)^{1/N}$$

$$P = \prod\ p_1^{p_1} p_2^{p_2} ... p_M^{p_M}$$

And by taking logarithm we have

$\log(P) = \sum_i p_i \log(p_i)$

This measure has non-positive values so we instead fix it to be positive by using minus sign which also makes it same as entropy where we measure average bits required by each symbol when the signal is encoded.

$$H(x) = -\log(P) = -\sum_i p_i \log(p_i)$$

Assume now instead that we want to **compare** two distributions. In this case, we may want to calculate mismatch between two different distributions by calculating ratios like:

$$R = \prod_i \max\left(p_i / q_i, q_i / p_i\right)^{(p_i + q_i)/2}$$

Here we calculate maximum ratio between each probability and its average mean value when distributions are blended together $(p_i + q_i)/2$.

$$\log(R) = \sum_i \frac{p_i + q_i}{2} \left|\log\left(\frac{q_i}{p_i}\right)\right|.$$

This entropy-like measure then is always better way to compare distributions unlike KL-divergence measure where the ratios between each parameter may cancel others (like encoding bit lenghts may compensate for shorter ones elsewhere) because we do **not** take maximum value of the ratios:

$$R' = \prod_i \left(p_i / q_i\right)^{(p_i + q_i)/2}$$

$$\log(R') = \sum_i \frac{p_i + q_i}{2} \log\left(\frac{q_i}{p_i}\right)$$

My thanks go to Tommi Jauhiainen who introduced me into this idea although I did the necessarily math to understand reason why absolute valued logarithms in KL-divergence are always better when comparing distributions.