Errors in Entropy Estimation

Roy Schestowitz

22nd April 2006

Abstract

We consider the problem of estimating the overlap between two data clouds in a high-dimensional hyperspace. In order to do so, we measure the distance (shuffle or Euclidean) between each possible pairing of points, where each point corresponds to one data instance encoded as a vector. A large collection of points shapes the cloud from which they are sometimes derived in a process that involves generative point distribution models. For any point in a given cloud A, we compute its distance to each example in another cloud B and vice versa. All these distances can be arranged in the form of a matrix, which can then be analysed to estimate the level of overlap between these two clouds of data. The principles of entropic graphs are used to infer complexity from the relationships embedded in the matrix. In our attempts to compute entropy, there is a level of *uncertainty* involved. We are most concerned about the error in the calculation, but we also consider an error which is the side-effect of repeating the experiments with multiple and distinct instantiations. Each instantiation, which is derived using the same stochastic process, tends to entail slightly different values. The document explains the derivation of these errors, as well as the calculation of entropy itself.

1 Calculating the Entropy

Entropy is a measure of uncertainty, which can often be used to assess the complexity among data patterns. In the context of our experiments, entropy is used to estimate the complexity of clouds of data by treating and interpreting them as a connected graphs.

Distance between point in the data clouds form a matrix. That matrix can in turn be evaluated for its complexity using an extended idea that is related to Shannon's entropy. In the context of data clouds, we seek to identify the level of point dispersion, as well as the correlation of that dispersion when two Gaussian distributions are involved.

As a simplistic example, we consider a spherical normal distribution in hyperspace. We consider yet another such distribution and let it gradually drift away from the first. We can then estimate the entropy of the two distributions, the joining of these, and come up with a certain measure of similarity. We observe a well-behaved decrease as the clouds gradually differ, as expected.

The formulation we use to calculate entropy involves the notion of a graph G and and a symbol for entropy, H. It also involves the two data clouds, which in the name of simplicity, we shall refer to as A and B. For the two clouds, we may assume for the sake of the argument, that we have obtained the distances between all points which they comprise of.

Our estimation of overall entropy is as follow:

$$H_{total} = (H(G[A \rightarrow B] - H(G[A \rightarrow A_{sample}]) \tag{1}$$

More latterly, we replaced our antiquated and confusing notation. In practice, we ought to replace A with S_o, which corresponds to the word "synthetic". In our experiments, we tend to deal with synthetic images that are generated from a model of appearance (combining shape and intensity). A_{sample} is also known as S_i, whose size is arbitrary and can be extended at will. S_i used to be merely a subset of the full set S_0, but it must not be contained in S_0. It is only derived from the same model as S_0 so it is *not* the case that $S_i \subseteq S_0$. Likewise, and even more strictly, no instance in S_i should be contained in S_0.

We can extend the number of S_i's to consider in order to improve our estimations, whenever/if time permits. Ultimately, we are left with a graph which shows the formulation to be rather helpful. The calculation of entropy itself is as follows:

$$H(Z_n) = 1/(1-\alpha)[log L_\gamma(Z_n)/n^\alpha - const \tag{2}$$

where Z_n is the distribution (of varying density) , L_γ is the length of the graph, α is a value that lies between 0 and 1 and *const* is an unimportant constant, at least at this stage.

2 Calculating the Error

There remains an uncertainty which is due to the varying value of the graph length. That error propagates to the overall, larger-scale calculation as listed in Equation 1. This leads to imbalance in the value of entropy. The error can be estimated in the following way: for each of the entropy 'sub-components' above, entropy is estimated which is dependent on the graph distance. Thus, considering the standard error

$$\sigma = \frac{L_\gamma}{\sqrt{N-1}} \tag{3}$$

where N is the number of instantiation used for error estimation. In line with rules for error propagation in logarithms

$$\sigma_{propagated} = \frac{log(\sigma)}{L_\gamma} \tag{4}$$

This gets applied to both cloud comparisons. Then, in order to combine the contribution of both entropies

$$\sigma_{total} = \sqrt{\sigma^2_{S_0 \Rightarrow T} + \sigma^2_{S_0 \Rightarrow S_i} - 2\sigma_{S_0 \Rightarrow S_i}\sigma_{S_0 \Rightarrow T}} \tag{5}$$

$H(Z_n)$ together with σ_{total} provide the final estimation of entropy and its level of (un)certainty.