Minimisation of Cross Entropy Loss

predict. For $x \in \mathcal{X}$, let $p(x)$ be the true probability of the observation falling in class x, and let $\hat{p}(x)$ be the predicted probability of the observation falling in class x. A common loss function that ML models try to minimize is the ***cross-entropy loss.*** In this setting, it is defined as the cross entropy of the predicted distribution relative to the true distribution, i.e.

$$H(P, \hat{P}) = -\sum_{x \in \mathcal{X}} p(x) \log \hat{p}(x).$$

True distribution versus predicted distribution. The loss needs to minimised to have an accurate value.

This is linked to Wittgenstein's ruler:

> Wittgenstein's ruler: Either "talkers"are smart & voters are indeed stupid, or Voters are smart & tawkers are stupid

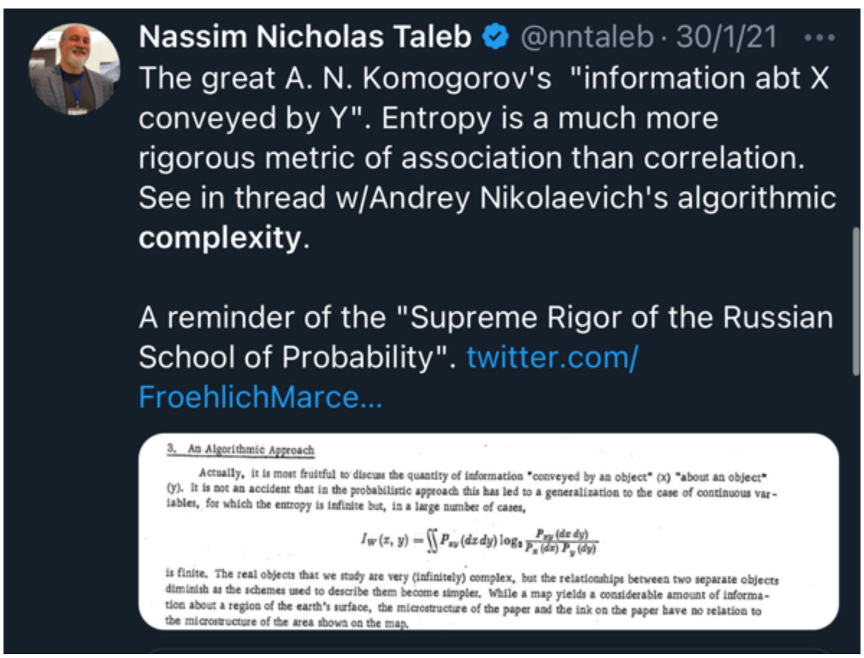
But Wittgenstein's ruler is objective oriented in this sense.

In this context, perplexity is simply the exponentiation of cross-entropy loss, i.e.

$$Perplexity(P, \hat{P}) = 2^{H(P,\hat{P})},$$

(assuming that 2 is the base of the logarithm used to compute the cross-entropy.)

Why use perplexity instead of cross-entropy loss? One reason I can see is that the value of cross-entropy loss depends on the base of the logarithm used, while perplexity is invariant to that choice. There could be other reasons, I'd love to hear from you if you know of them!



Nassim Nicholas Taleb @nntaleb · 30/1/21

The great A. N. Komogorov's "information abt X conveyed by Y". Entropy is a much more rigorous metric of association than correlation. See in thread w/Andrey Nikolaevich's algorithmic **complexity**.

A reminder of the "Supreme Rigor of the Russian School of Probability". twitter.com/FroehlichMarce...

3. An Algorithmic Approach

Actually, it is most fruitful to discuss the quantity of information "conveyed by an object" (x) "about an object" (y). It is not an accident that in the probabilistic approach this has led to a generalization to the case of continuous variables, for which the entropy is infinite but, in a large number of cases,

$$I_W(x, y) = \iint P_{xy}(dx\,dy) \log_2 \frac{P_{xy}(dx\,dy)}{P_x(dx)\,P_y(dy)}$$

is finite. The real objects that we study are very (infinitely) complex, but the relationships between two separate objects diminish as the schemes used to describe them become simpler. While a map yields a considerable amount of information about a region of the earth's surface, the microstructure of the paper and the ink on the paper have no relation to the microstructure of the area shown on the map.