How to Explain Log-Linear Relation Between Amount of Computations and Effectiveness of the Result – a Relation that Motivates the Need for Big Data

Francisco Zapata, Olga Kosheleva, and Vladik Kreinovich
University of Texas at El Paso, El Paso, TX 79968
fazg74@gmail.com, olgak@utep.edu, vladik@utep.edu

1. Log-Linear Relation: A Brief Description

- It is known that:
  - the more computations we perform,
  - the more efficient the decisions and designs resulting from these computations.
- Empirical data shows that there is a log-linear dependence between:
  - the effectiveness e of an application and
  - the amount d of computations that led to this application.
- Specifically, we have $e = a + b \cdot \ln(d)$ for some constants a and b.

2. This Empirical Relation Explains Why We Need Big Data

- *Reminder:* the formula $e = a + b \cdot \ln(d)$ describes the relation between:
  - the effectiveness e of an application and
  - the amount d of computations that led to this application.
- This empirical relation can be reformulated as

$$d \sim \exp(\text{const} \cdot e).$$

- This reformulation *explains* why we need *big data:*
  - every time we want to increase efficiency by one unit,
  - we need to double the amount of processed data.
- *What we do:* we provide an explanation for the empirical log-linear dependence.

3. How to Estimate Effectiveness

- The effectiveness e of an application is proportional to the number m of useful features that this design has.
- For example, let us look at a headache medicine.
- Its first – and most important – feature is that it should cure headaches.
- If it also avoids negative effects on the stomach, this is better.
- If it also clears your sinuses, even better, etc.
- Let us denote the average probability that a randomly selected substance (or design) has a feature by p.
- The features are usually independent.
- So, the probability that a randomly selected design has m features is p^m.

4. Towards an Explanation

- The probability that a randomly selected design has m features is p^m.
- According to statistics:
  - if a rare event has probability q,
  - then we need, on average, a sample of size $\approx \dfrac{1}{q}$ to observe at least one such event.
- So, to find a design with m features, we need to test $\dfrac{1}{p^m}$ different designs.
- The resulting amount of computations d is proportional to the number of tested designs, i.e., to $\dfrac{1}{p^m}$:

$$d = c' \cdot \left(\frac{1}{p}\right)^m.$$

5. Resulting Explanation

- The amount of computations is

$$d = c' \cdot \left(\frac{1}{p}\right)^m.$$

- By taking logarithms of both sides, we get

$$\ln(d) = \ln(c') + m \cdot \ln\left(\frac{1}{p}\right).$$

- So $m = A + B \cdot \ln(d)$, where

$$B = \frac{1}{\ln\left(\frac{1}{p}\right)} \text{ and } A = -\frac{\ln(c')}{\ln(d)}.$$

- On the other hand, the effectiveness e of a design is proportional to m: $e = c' \cdot m$.
- Hence $e = a + b \cdot \ln(d)$, where $a = c' \cdot A$ and $b = c' \cdot B$.

6. Conclusions

- Thus, we indeed get a log-linear dependence between:
  - the effectiveness e of an application and
  - the amount d of computations that led to this application.

7. References

- M. Banko and E. Brill, "Scaling to very very large corpora for natural language disambiguation", *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, 2001, pp. 26–33.
- T. Brants et al., "Large language models in machine translation", *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Processing*, 2007, pp. 858–867.
- J. Lin, "Is big data a transient problem?", *IEEE Internet Computing*, 2015, September/October 2015, pp. 86–90.