Attention in machine learning: how to explain the empirical formula

Sobita Alam, Arman Hossain, Samin Islam,
Arin Rahman, and Vladik Kreinovich
Department of Computer Science, University of Texas at El Paso
500 W. University, El Paso, TX 79968, USA
salam@miners.utep.edu, arahman6@miners.utep.edu,
sislam3@miners.utep.edu, ahossain4@miners.utep.edu, vladik@utep.edu

Attention as a way to better classification. In many practical situations, we have several objects, each of which is characterized by vector $x_i = (x_{i,1}, \ldots, x_{i,N})$ consisting of this object's numerical characteristics. For example, we have many picture of pets, and you want to classify them into cats and dogs. One of the difficulties is that objects within each class are different: e.g., dogs can be large and small, of different breeds, etc. To make classification task easier, it is desirable to replace each specific vector x_i with a weighted average $y_i = \sum_j w_{ij} \cdot x_j$ of all the objects x_j which are similar to x_i. This way, the role of individual characteristics will diminish, and the classification task will become easier.

A natural way to describe the closeness between the objects x_i and x_j is to use the usual metric $d(a,b) = \sqrt{\sum_k (a_k - b_k)^2}$. The smaller this distance, the larger should be the weight, so we must have $w_{ij} \sim f(d(x_i, x_j))$ for some decreasing function $f(v)$. The sum of the weights should be equal to 1, so we must have $w_{ij} = f(d(x_i, x_j))/\left(\sum_\ell f(d(x_i, x_\ell)\right)$. This expression can be simplified if we take into account that overall, the values x_{ij} are reasonably random, in which case the value $x_i^2 = \sum_k x_{i,k}^2$ is close to some constant C (N time average of $x_{i,j}^2$). Then, $d^2(x_i, x_j) = x_i^2 + x_j^2 - 2x_i \cdot x_j \approx 2C - 2x_i \cdot x_j$. So, a decreasing function of $d(x_i, x_j)$ can be described as an increasing function of the dot product $x_i \cdot x_j$. Thus, $w_{ij} = F(x_i \cdot x_j)/\left(\sum_\ell F(x_i \cdot x_\ell)\right)$.

Empirical evidence shows that out of all increasing functions $F(v)$, functions $F(v) = \exp(\alpha \cdot v)$ work the best. How can we explain this empirical fact?

Our explanation is based on the fact that measurements are noisy. So, a natural requirement is that the resulting values y_i should be affected by the noise as little as possible. If we replace the original values $x_{i,j}$ with noisy values $\widetilde{x}_{i,k} = x_{i,k} + n_{i,k}$ for some noise $n_{i,k}$ with 0 mean, then the dot product $\widetilde{x}_i \cdot \widetilde{x}_j$ becomes $x_i \cdot x_j + x_i \cdot n_i + n_i \cdot x_j + n_i \cdot n_j$. The expected value of terms $x_i \cdot n_j$ is 0, so the only non-zero addition to the dot product is $E[n_i \cdot n_j]$. For local noise, this expected value is 0, but if the noise had a global component with mean square value m, then, on average, all dot products are increased by the same constant m.

So, we want to find the function $F(v)$ for which adding a constant m to all dot product would not change the weights. In particular, for two objects, this means that $\dfrac{F(a+m)}{F(a+m) + F(b+m)} = \dfrac{F(a)}{F(a) + F(b)}$ for all a, b, and m. If we apply $1/z$ to both sides of this equality and subtract 1 from both sides, we get $F(b+m)/F(a+m) = F(b)/F(a)$. Multiplying both sides by $F(a+m)/F(b)$, we get $F(b+m)/F(b) = F(a+m)/F(a)$ for all a and b, i.e., that the ratio $F(a+m)/F(a)$ does not depend on a, it only depends on m: $F(a+m)/F(a) = g(m)$ for some function $g(m)$. Thus, $F(a+m) = g(m) \cdot F(a)$. It is known that the only increasing solution to this functional equation is $F(a) = c \cdot \exp(\alpha \cdot a)$ which is, from the viewpoint of the weights $w_{i,j}$, equivalent to $F(a) = \exp(\alpha \cdot a)$. This is exactly what we needed to explain.

(To solve the functional equation, differentiate both sides by m and take $m = 0$. Then $F'(a) = g'(0) \cdot F(a)$, with $\alpha \overset{\text{def}}{=} g'(0)$, i.e., $dF/da = \alpha \cdot F$ and $dF/F = \alpha \cdot da$. Integrating, we get $\ln(F) = \alpha \cdot a + \text{const}$, so $F(a) = \text{const} \cdot \exp(\alpha \cdot a)$.)