Log Probability and Log Softmax

Wednesday, January 5, 2022 11:00 PM

TOC:

Log probability & Log softmax

To represent probabilities in a logarithmic scale instead of the standard [0,1] unit interval.

Instead of multiplying the probabilities of independent events, logarithm converts multiplication to addition.

Interpretation of Log probability in Information theory:

- the negative of the average log probability is the information entropy of an event.
  From <https://en.wikipedia.org/wiki/Log_probability>

- likelihoods are often transformed to the log scale, and the corresponding log-likelihood can be interpreted as the degree to which an event supports a statistical model.
  From <https://en.wikipedia.org/wiki/Log_probability>

Log softmax function versus softmax

$$\sigma(z_i) = {e^{z_i}}\Big/{\sum_{j=1}^{K} e^{z_j}} \qquad \rightarrow Eq\ 1$$

$$for\ i = 1,2,.,.,K\ and\ z = z_1, z_2, ., ., z_K$$

In a multiclass classification problem, the network outputs K values, one for each class. Then, softmax is used to convert logits into probabilities which sums up to 1.

Softmax can convert values from [-∞, +∞] to [0,1].

But,

for $x = 19$,

$$e^{19} = 178482300.96$$

For x = 20,

$$e^{20} = 485165195.41.$$

The exponential function's output can blow up easily. Also, dividing large numbers in the softmax formula may cause instabilities.

Here is the log softmax: (log of softmax)

$$\log\left(e^{z_i} \Big/ \sum_{j=1}^{K} e^{z_j}\right) \rightarrow Eq\ 2$$

$\log\left(\frac{a}{b}\right) = \log a - \log b$, using this on equation 2,

$$\rightarrow z_i - \log\left(\sum_{j=1}^{K} e^{z_j}\right)$$

$$\approx z_i - \max(z) \text{ since sum is dominated by the largest number.}$$

We see that log softmax is nearly just $x - max(x)$ which is naturally much faster to compute than anything involving logarithms and exponentials. We are also guaranteed that the output won't be of a vastly different scale than the input.
From <https://datascience.stackexchange.com/questions/40714/what-is-the-advantage-of-using-log-softmax-instead-of-softmax>

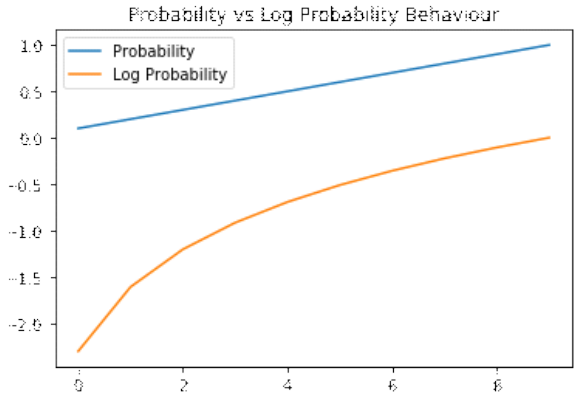
Stackoverflow: How is log_softmax() implemented to compute its value (and gradient) with better speed and numerical stability?

Log Softmax Advantages:

Representing probabilities in this way has several practical advantages:
From <https://en.wikipedia.org/wiki/Log_probability>

1. **Speed**: Since multiplication is more expensive than addition, taking the product of a high number of probabilities is often faster if they are represented in log form. (The conversion to log form is expensive, but is only incurred once.)
   - Multiplication arises from calculating the probability that multiple independent events occur: the probability that all independent events of interest occur is the product of all these events' probabilities.

2. **Accuracy**: The use of log probabilities improves numerical stability, when the probabilities are very small, because of the way in which computers approximate real numbers. [multiplication of very small numbers]



3. **Simplicity**: Many probability distributions have an exponential form. Taking the log of these distributions eliminates the exponential function, unwrapping the exponent.
   - For example, the log probability of the normal distribution's probability density function is
     $-((x - m_x)/\sigma_m)^2 + C$
     instead of

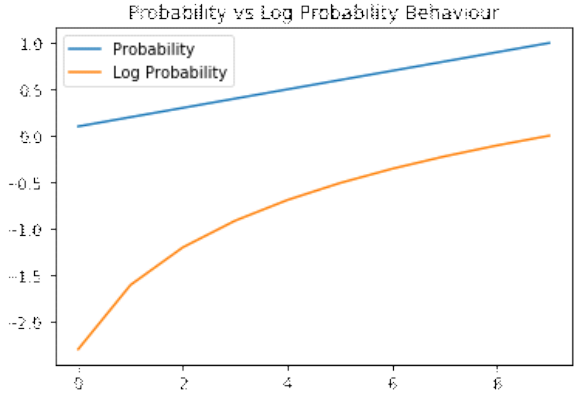
$$C_2 \exp\left(-((x - m_x)/\sigma_m)^2\right)$$

Log softmax can penalize errors larger than softmax

Probability
[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]
Log Prob :
[-2.3 -1.61 -1.2 -0.92 -0.69 -0.51 -0.36 -0.22 -0.11 0.]
From <https://deepdatascience.wordpress.com/2020/02/27/log-softmax-vs-softmax/>

Computing the gradient of log-softmax is cheaper than softmax
https://deepdatascience.wordpress.com/2020/02/27/log-softmax-vs-softmax/

It has an impact on the training time.



Example: suppose that for one example, the true class is 1. but the model outputs 0.001.
The above plots show that the number of steps needed to correct 0.001 to 1 is linear. But, log probability is exponential curve with large slope at this small probability.

Laws of logarithm

$\log A + \log B = \log AB$

$log\ A\ -\ log\ B\ =\ log \frac{A}{B}$

$\log A^n = n \log A$

$log\ 1\ =\ 0$
$\log_m m\ = 1$

$\log 0 =$ It's not a real number, because you can never get zero by raising anything to the power of anything else.

You can never reach zero, you can only approach it using an infinitely large and negative power