Detailed Deviation of the trigger threshold equation

Our system automatically generates a threshold that triggers queries at the desired frequency. If a teacher expects to answer a question out of 1,000 data, the system queries the top 0.1% most *significant* subtrees.

Recall in section 3.4 in the paper, we showed that we measure the tightness of a subtree with the *significance*. This *significance* is used to define the trigger. Given the desired frequency, expressed as the number of queries desired out of the number of subsequences that reach the dendrogram, the trigger is computed using equation (1), which is simply the negative of the probit function [1] of the desired frequency.

$$trigger\ threshold = -probit(desired\ frequency) \qquad (1)$$

The readers may appreciate from the equation that we assume the distribution of the negative of the density scores is the standard normal distribution. The threshold is the point from which the integral of the probability of density scores to $+\infty$ is the desired frequency. The assumption of normal distribution is reasonable as can be inferred from Figure7.*right*.

In the following, we offer a detailed deviation of equation (1).

Suppose the negative of the density scores follow a normal distribution, then the trigger threshold computed using equation(1) is the point at which the cumulative probability equals the desired frequency, where the cumulative probability is computed as the integral of the probability of the density score from a given point to $+\infty$. Clearly, this is the point that we want for the threshold. In other words, if the assumption of normal distribution of the negative of the density scores is correct, the threshold computed using equation (1) will trigger queries at desired frequency. In the following, we will explain why the distribution of the negative of the density score is the normal distribution.

As can be seen in Figure 7.*right* in the paper, the heights of subtrees of size four follow approximately a Gaussian distribution. If we assume this distribution is a Gaussian, then the Z-scores of the heights of subtrees of size four, which are computed using equation (2), follow the normal distribution. More formally, it is

$$P[\text{Zscores of the subtree heights} \mid \text{subtree size} = 4] \sim \text{N}(0,1)$$

Similarly, we have

$$P[\text{Zscores of the subtree heights} \mid \text{subtree size} = \text{n}] \sim \text{N}(0,1),$$

where n are the possible sizes of subtrees, ranging from two to MaxSubtreeSize. This is true because the heights of subtrees of each size follow approximately a Gaussian distribution similar to Figure *7.right*, but with a different mean and standard deviation.

$$Zscore\ of\ heights = \frac{subtree\ heights - Mean_{subtree\ size}}{std_{subtree\ size}} \qquad (2)$$

Now, let's consider the general case of the distribution of the Z-scores of heights of subtrees of all possible sizes. According to the total probability law, we have

$$\begin{aligned} &P[\text{Zscores of the heights of subtrees } = \text{v}] \\ &= P[\text{subtree size} = 1] * P[\text{Zscores of the subtree heights} = \text{v} \mid \text{size} = 1] \\ &\quad + P[\text{subtree size} = 2] * P[\text{Zscores of the subtree heights} = \text{v} \mid \text{size} = 2] \\ &\quad + \ldots \\ &\quad + P[\text{subtree size maxSubtreeSize}] * P[\text{Zscores of the subtree heights} = \text{v} \mid \text{size} \\ &\qquad = \text{maxSubtreeSize}] \end{aligned}$$

Since the conditional distributions for all sizes of subtrees are the same, the above equation can be simplified as

$$\begin{aligned} &P[\text{Zscores of the heights of subtrees } = \text{v}] \\ &= P[\text{Zscores of the subtree heights} = \text{v} \mid \text{size} = 1] \\ &* (P[\text{subtree size} = 1] + P[\text{subtree size} = 2] + \cdots + P[\text{subtree size maxSubtreeSize}]) \\ &= P[\text{Zscores of the subtree heights} = \text{v} \mid \text{size} = 1] * 1 \\ &= P[\text{Zscores of the subtree heights} = \text{v} \mid \text{size} = 1] \end{aligned}$$

The equation implies the Z-scores of the heights of subtrees of all sizes follow the same distribution as the Zscores of the heights of subtrees of a given size, which, as mentioned above, is a normal distribution.

If we compare equation of *significance* (c.f. section 3.4 in paper) with equation(2), it's easy to see that the Z-scores of the heights of the subtrees is the negative number of the density score, which validates our assumption that the negative number of the density score follows a normal distribution and makes our computation of the trigger threshold reasonable.

Reference:

[1] Wichura, M.J. (1988). "Algorithm AS241: The Percentage Points of the Normal Distribution". Applied Statistics (Blackwell Publishing) 37 (3): 477–484.