Sketching Algorithm for Kendall Tau's Rank Correlation

Gromit Chan

July 27, 2020

Here is the sketching algorithm inspired by [BHP09] to approximate Kendall's Tau, a known rank correlation, with error ϵ and confidence $1-\delta$. The original paper's sketches also cater to sparse settings (i.e., rank correlations between a list of sparsely ranked items) but here I simplify and modify it for the original Kendall's Tau calculation, where all items are ranked.

1 Kendall's Tau Rank Correlation

Kendall's Tau measures the correlation between two rankings (e.g. similarity between two users' preferences to a set of items). Suppose two rankings, x and y, are two lists of integers with n items. Then, the Kendall's Tau coefficient τ can be calculated as follows:

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} sgn(x_i - x_j) sgn(y_i - y_j) \tag{1}$$

, where sgn denotes the sign of the integer.

The intuition is that, if we choose any two items from n items, how many of these item pairs have the same sorted order between the rankings? (i.e. $sgn(x_i - x_j) = sgn(y_i - y_j)$) Such pairs are said to be *concordant* else they are said to be *discordant*. In other words, Equation 1 can also be expressed as follows:

$$\tau = \frac{\text{(number of concordant pairs) - (number of discordant pairs)}}{\binom{n}{2}} \tag{2}$$

Interestingly, Kendall's Tau can also be viewed as the probability that the pair you choose from the $\binom{n}{2}$ pairs of items is concordant. Denote such an event as C and the event that the pair is discordant as D. $P(C) = \frac{n_c}{\binom{n}{2}}$ where n_c is the number of concordant pairs and $P(D) = 1 - P(C)$. Thus, $\tau = P(C) - P(D) = 2P(C) - 1$. This allows us to approximate Kendall's Tau by approximating $P(C)$.

2 Approximating Kendall's Tau

Denote the set of $\binom{n}{2}$ pairs of items as I. If we choose a pair uniformly at random from I and see whether it is a concordant pair, this becomes a Bernoulli trial with a success probability of $P(C)$. Thus, we can perform a series of independent Bernoulli trials and measure the percentage of successes, which approximates $P(C)$ and eventually, Kendall's Tau. Let X be the number of successes of k Bernoulli trials so that the maximum likelihood estimator of $P(C)$ is $P(\hat{C}) = \frac{X}{k}$. The problem now is to find the appropriate k to have a desirable error and confidence. This can be done by Hoeffding's inequality[1].

The error ϵ_c for $P(\hat{C})$ is defined as follows: given a Bernoulli random variable X_i and the number of successes in k_c trials $X = \sum_{i=1}^{k_c} X_i$, the following inequality holds: $P(|X - E(X)| \geq k_c \epsilon_c) \leq 2exp(-2k_c\epsilon_c^2)$. Substituting $X = k_c P(\hat{C})$ and $E(X) = k_c P(C)$, we have $P(|P(C) - P(\hat{C})| \geq \epsilon_c) \leq 2exp(-2k_c\epsilon_c^2)$. Therefore,

[1] `https://en.wikipedia.org/wiki/Hoeffding's_inequality`

in order to make sure the true $P(C)$ falls within the range of $[P(\hat{C}) - \epsilon_c, P(\hat{C}) + \epsilon_c]$ with high probability $1 - \delta$ (i.e. confidence), we need to have $2exp(-2k_c\epsilon_c^2) \leq \delta$, which we get $k_c \geq \frac{ln\frac{2}{\delta}}{2\epsilon_c^2}$.

Notice that k_c is the number of trails needed to have $P(\hat{C})$ with error ϵ_c and confidence $1 - \delta$ but not the final k to get $\hat{\tau} = 2P(\hat{C}) - 1$. Since $|\tau - \tau_c| = |(2P(C) - 1) - (2P(\hat{C}) - 1)| = 2|P(C) - P(\hat{C})|$, we can replace ϵ_c with 2ϵ and arrive the final number of trials needed:

$$k \geq \frac{2\ln\frac{2}{\delta}}{\epsilon^2} \tag{3}$$

3 Implementing the Sketches

The sketch is basically an integer list with k elements, of which each element is a uniform random sample from the range $[0, \binom{n}{2})$. These elements are indexes of the condensed matrix for $\binom{n}{2}$ pairs (i.e. $[(0,1), (0,2), ..., (0, n-1), (1,2), (1,3), ..., (n-2, n-1)]$). There is a closed form expression to find (i, j) from the indexes[2] so that it is not necessary to create a long list of condensed matrix that consumes runtime and memory. Once the sketch is created, it can be used with any pairs of ranks.

References

[BHP09] Yoram Bachrach, Ralf Herbrich, and Ely Porat. "Sketching algorithms for approximating rank correlations in collaborative filtering systems". In: *International Symposium on String Processing and Information Retrieval*. Springer. 2009, pp. 344–352 (cit. on p. 1).

[2] `https://stackoverflow.com/a/36867493`