Mean Reciprocal Rank (MRR) in Temporal Graph Benchmark (TGB) Applications

TARGET

28/01/2025

Definition and Calculation

For a set of Q queries, where each query i has a ranked list of predictions, the Reciprocal Rank (RR) for query i is defined as the inverse of the rank of the first correct answer:

$$\text{RR}_i = \frac{1}{\text{rank}_i}.$$

The Mean Reciprocal Rank is the average of RR values across all queries:

$$\text{MRR} = \frac{1}{Q} \sum_{i=1}^{Q} \frac{1}{\text{rank}_i}.$$

In TGB, predictions are evaluated using a **filtered ranking protocol** to avoid penalizing models for predicting plausible but non-ground-truth answers. Specifically: - A candidate set of predictions is ranked based on scores output by the model. - Ground-truth candidates are identified, and non-plausible options are excluded from consideration.

For example, if the ground-truth entity for a query is ranked k after filtering, the reciprocal rank is $\text{RR}_i = \frac{1}{k}$. The MRR aggregates this value across all queries.

Limitations

1. **Sensitivity to Top Rankings**: MRR only considers the rank of the first correct answer, ignoring subsequent correct answers. This may lead to incomplete evaluations in multi-label or multi-answer settings. 2. **Dependence on Filtering**: In tasks like temporal link prediction, the filtering protocol requires domain-specific knowledge, introducing variability in reported results. 3. **Bias in Long-Tail Predictions**: MRR tends to favor queries with high-ranked answers, potentially biasing against tasks where correct answers appear later in the ranking.

Potential Alternatives

1. **Normalized Discounted Cumulative Gain (nDCG)**: - Captures the relevance of all ranked candidates, not just the top-ranked one. - Uses a logarithmic discount for lower ranks, ensuring smoother relevance decay.

$$\text{nDCG} = \frac{\text{DCG}}{\text{IDCG}}, \quad \text{DCG} = \sum_{i=1}^{n} \frac{2^{\text{rel}_i} - 1}{\log_2(i+1)},$$

   where rel_i is the relevance score of the i^{th} prediction.

2. **Precision@k**: - Measures the fraction of correct answers in the top k predictions.

$$\text{Precision@k} = \frac{\text{Correct Predictions in Top } k}{k}.$$

3. **Hit Ratio (HR)**: - Evaluates whether at least one correct answer appears in the top k predictions:

$$\text{HR@k} = \begin{cases} 1, & \text{if a correct answer is in the top } k, \\ 0, & \text{otherwise.} \end{cases}$$