TA6

Arun Kumar Rajasekaran

Project discussion and ppts

Evaluation metrics in NLP

Types

- Intrinsic Evaluation — Focuses on intermediary objectives (i.e. the performance of an NLP component on a defined subtask)
- Extrinsic Evaluation — Focuses on the performance of the final objective (i.e. the performance of the component on the complete application)

Metrics reference

https://medium.com/@mikeusru/common-metrics-for-evaluating-natural-language-processing-nlp-models-e84190063b5f

1. Accuracy

Denotes the fraction of times the model makes a correct prediction as compared to the total predictions it makes. Best used when the output variable is categorical or discrete. For example, how often a sentiment classification algorithm is correct.

2. Precision

Evaluates the percent of true positives identified given all positive cases. Particularly helpful when identifying positives are more important than overall accuracy. For example, if identifying a cancer that is prevalent 1% of the time, a model that always spits out “negative” will be 99% accurate, but 0% precise.

3. Recall

The percent of true positives versus combined true and false positives. In the example with a rare cancer that is prevalent 1% of the time, if a model creates totally random predictions (50/50), it will have 50% accuracy (50/100), 50% precision (0.5/1), and 1% recall (0.5/50)

4. F1 score

Combines precision and recall to give a single metric — both completeness and exactness. (2 * Precision * Recall) / (Precision + Recall). Used together with accuracy, and useful in sequence-labeling tasks, such as entity extraction, and retrieval-based question answering.

5. AUC

Area Under Curve; Combines true positives vs false positives as threshold for prediction is varied. Used to measure the quality of a model independent of prediction threshold, and to find the optimal prediction threshold for a classification task.

6. MRR

Mean Reciprocal Rank. Evaluate the responses retrieved given their probability of being correct. The mean of the reciprocal of the ranks of the retrieved results. Used heavily in all information-retrieval tasks, including article search and e-commerce search.

7. MAP

Mean average precision, calculated across each retrieved result. Used in information-retrieval tasks.

8. RMSE

Root mean squared error — very common way to capture a model's performance in a real-value prediction task. Good way to ask "How far off from the answer am I?" Calculates the square root of the mean of the squared errors for each data point. Used in numerical prediction — temperature, stock market price, position in euclidean space…

9. MAPE

Mean absolute percentage error. Used when the output variable is a continuous variable, and is the average of absolute percentage error for each data point. Often used in conjunction with RMSE and to test the performance of regression models.

10. BLUE

Captures the amount of n-gram overlap between the output sentence and the reference ground truth sentence. Has many variants, and mainly used in machine translation tasks. Has also been adapted to text to text tasks such as paraphrase generation and summarization.

11. METEOR

Precision-based metric to measure quality of generated text. Sort of a more robust BLEU. Allows synonyms and stemmed words to be matched with the reference word. Mainly used in machine translation.

12. ROGUE

Like BLEU and METEOR, compares quality of generated to reference text. Measures recall. Mainly used for summarization tasks where it’s important to evaluate how many words a model can recall (recall = % of true positives versus both true and false positives).

12. Perplexity

Measures how confused an NLP model is, derived from cross-entropy in a next word prediction task. Used to evaluate language models, and in language-generation tasks, such as dialog generation.

NLTK. Sample usage

https://www.nltk.org/howto/metrics.html

More on Similarity Distances

https://flavien-vidal.medium.com/similarity-distances-for-natural-language-processing-16f63cd5ba55