Perplexity

Spring Semester Capstone Study

TEAM Kai.Lib

발표자 : 원철황

2020.04.15 (WED)

What is perplexity?

Perplexity

- In Information Theory, perplexity is measurement of How well a probability distribution or probability model predicts a sample.
- It may be used to compare probability models.
- A low perplexity indicates the probability distribution is good at predicting the sample.

An Estimate of an Upper Bound for the Entropy of English

Peter F. Brown,Stephen A. Della Pietra,
Vincent J. Della Pietra,Jennifer C. Lai, Robert L. Mercer

We present an estimate of an upper bound of 1.75 bits for the entropy of characters in printed English, obtained by constructing a word trigram model and then computing the cross-entropy between this model and a balanced sample of English text. We suggest the well-known and widely available Brown Corpus of printed English as a standard against which to measure progress in language modelling and offer our bound as the first of what we hope will be a series of steadily decreasing bounds.

What is perplexity?

Perplexity

- In Language Model, perplexity is used to measure the performance of Models
- Extirnsic Evaluation: 실제 작업으로 입력에 대한 결과를 확인하는 외부 평가 방식
- Intrinsic Evaluation: 모델 내에서 자신의 성능을 수치화하여 결과를 내놓는 내부 평가 방식
- Perplexity is usually used in words, PPL

$$PPL(W) = P(w_1, w_2, w_3, \ldots, w_N)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P(w_1, w_2, w_3, \ldots, w_N)}}$$

- PPL은 단어의 수로 정규화 (normalization)된 테스트 데이터에 대한 역수이다.
- PPL을 최소화한다는 것은 문장의 확률을 최대화 하는 것과 같은 뜻

The categorical cross entropy loss measures the dissimilarity between the true label distribution y and the predicted label distribution ˆy, and is defined as cross entropy:

$$L_{\text{cross-entropy}}(\hat{\mathbf{y}}, \mathbf{y}) = -\sum_i y_i \log(\hat{y}_i)$$

Examples

셰익스피어 Training Set 이용한 셰익스피어의 작품 추론

Unigram

To him swallowed confess hear both. Which. Of save on trail for are ay device and rote life have
Every enter now severally so, let
Hill he late speaks; or! a more to leg less first you enter
Are where exeunt and sighs have rise excellency took of.. Sleep knave we. near; vile like

Bigram

What means, sir. I confess she? then all sorts, he is trim, captain.
Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry. Live king. Follow.
What we, hath got so she that I rest and sent to scold and nature bankrupt, nor the first gentleman?

Trigram

Sweet prince, Falstaff shall die. Harry of Monmouth's grave.
This shall forbid it should be branded, if renown made it empty.
Indeed the duke; and had a very good friend.
Fly, and will rid me these news of price. Therefore the sadness of parting, as they say, 'tis done.

Quadrigram

King Henry.What! I will go seek the traitor Gloucester. Exeunt some of the watch. A great banquet serv'd in;
Will you not tell me who I am?
It cannot be but so.
Indeed the short and the long. Marry, 'tis a noble Lepidus.

Examples

WSJ Training Set 이용한 셰익스피어의 작품 추론

Unigram

Months the my and issue of year foreign new exchange's september were recession exchange new endorsed a acquire to six executives

Bigram

Last December through the way to preserve the Hudson corporation N. B. E. C. Taylor would seem to complete the major central planners one point five percent of U. S. E. has already old M. X. corporation of living on information such as more frequently fishing to keep her

Trigram

They also point to ninety nine point six billion dollars from two hundred four oh six three percent of the rates of interest stores as Mexico and Brazil on market conditions

What is perplexity?

Perplexity

- In Categorical Classification, Cross Entrophy is Summation of perdiction entrophy based label
- Chain Rule을 적용하면 다음과 같이 표현 가능

$$PPL(W) = \sqrt[N]{\frac{1}{P(w_1, w_2, w_3, \ldots, w_N)}} = \sqrt[N]{\frac{1}{\prod_{i=1}^{N} P(w_i|w_1, w_2, \ldots, w_{i-1})}}$$

- 문장의 길이가 길어질 수록 확률이 낮아지기 때문에 문장 길이 N 제곱근으로 정규화한다.
- PPL 해석을 위한 예시) **N번의 주사위를 던져 수열을 만드는 상황**

$$\mathrm{PPL}(x) = (\tfrac{1}{6}^{N})^{-\frac{1}{N}} = 6$$

- **Time Step 마다 6가지 branch로 확장 가능함을 뜻함**
- **20,000 개의 어휘로 이루어진 뉴스 기사의 단순 PPL을 추론하면 값은 20,000 이 나올 것.**
- 이를 수식적으로 전개하면 얻어지는 Cross Entrophy와 PPL의 관계는 다음과 같다.

$$\mathrm{PPL} = \exp(\text{Cross Entropy})$$

Conclusion

The Shannon Game:

- How well can we predict the next word?

  I always order pizza with cheese and ____

  The 33rd President of the US was ____

  I saw a ____

- Unigrams are terrible at this game. (Why?)

mushrooms 0.1
pepperoni 0.1
anchovies 0.01
....
fried rice 0.0001
....
and 1e-100

- 예시와 같은 추론에서 다음 단어를 예측할 확률은 항상 균일분포가 아닌 Distributed
- 위의 주사위 경우처럼 단순 균일분포 가정을 통해 해석한 Branch 개수가 나타나는 것이 아님

Training 38 million words, test 1.5 million words, WSJ

N-gram Order	Unigram	Bigram	Trigram
Perplexity	962	170	109

Conclusion

Perplexity

Language Model 성능 평가에 Loss 대신 사용해도 되느냐?

Yes