

ECE 515

Information Theory

Perplexity

Language Models

- A language model estimates the probability of the next word given all previous words
- Suppose the current context is
 - The weather today is
- The model predicts a probability distribution over every possible next word

Next Word Distribution

Next Word	Probability
sunny	0.45
cloudy	0.25
rainy	0.15
cold	0.05
beautiful	0.04
others	0.06

- All probabilities sum to one
- The model assigns a probability to every possible continuation

Model Learning

- During training, the model is exposed to billions of sentences such as
 - The weather today is sunny.
 - The weather today is cloudy.
 - The weather today is rainy.

- Over time, it learns conditional probability distributions such as

$$P(\text{sunny} \mid \text{The weather today is})$$

$$P(\text{rainy} \mid \text{The weather today is})$$

- The goal of training is to learn these conditional probabilities accurately

Evaluating a Language Model

- Evaluate the model on the sentence
 - The weather today is sunny.
- Instead of predicting the whole sentence at once, the model predicts one word at a time

⟨START⟩ → The → weather → today → is → sunny

- Each arrow represents one prediction

Predictions 1 and 2

Prediction 1

Context: ⟨START⟩

Word	Probability
The	0.60
A	0.20
It	0.10
others	0.10

Correct word: The
Record 0.60

Prediction 2

Context: The

Word	Probability
weather	0.70
dog	0.10
man	0.08
others	0.12

Correct word: weather
Record 0.70

Recorded Probabilities

- Continuing in the same way produces

The	weather	today	is	sunny
0.60	0.70	0.50	0.45	0.80

- These are the probabilities assigned to the correct words

Chain Rule

- For three events A, B, C

$$P(A, B, C) = P(A)P(B|A)P(C|A, B)$$

- Language follows this principle
- For a sentence $w_1, w_2, \dots, w_N$

$$P(w_1, w_2, \dots, w_N) = \prod_{i=1}^N P(w_i | w_1, \dots, w_{i-1})$$

Probability of a Sentence

$$\begin{aligned} P(\text{The weather today is sunny}) = & \\ P(\text{The}) \times P(\text{weather} \mid \text{The}) \times P(\text{today} \mid \text{The, weather}) & \\ \times P(\text{is} \mid \text{The, weather, today}) & \\ \times P(\text{sunny} \mid \text{The, weather, today, is}) & \end{aligned}$$

Each factor is the probability assigned to the correct word given the words before it

Example

- Assume the probabilities of the correct words are
0.60, 0.70, 0.50, 0.45, 0.80
- Then
$$P(\text{sentence}) = 0.60 \times 0.70 \times 0.50 \times 0.45 \times 0.80$$
$$= 0.0756$$

Sentence Probabilities

- Every probability lies between zero and one
- Multiplying many numbers less than one will produce a very small number
- Suppose every prediction is excellent, so $p = 0.95$

$$0.95^5 = 0.77$$

$$0.95^{20} = 0.36$$

$$0.95^{100} = 0.0059$$

- Sentence probabilities cannot be compared across sentences of different lengths

Sentence Probabilities

- Solution:
 - Use the average model confidence per prediction
- Since probabilities multiply, the appropriate average is the geometric mean

Geometric Mean

- For probabilities $p_1, p_2, \dots, p_N$ the geometric mean is

$$\left(\prod_{i=1}^N p_i \right)^{1/N}$$

- For the example

$$(0.0756)^{1/5} = 0.596$$

- On average, the model assigns 59.6% probability to the correct next word

Perplexity

Perplexity is defined as

$$PP = \frac{1}{\left(\prod_{i=1}^N P(w_i)\right)^{1/N}} = \left(\frac{1}{P(\text{sentence})}\right)^{1/N}$$

For the example

$$PP = \frac{1}{0.596} = 1.68$$

Interpreting Perplexity

- Suppose every prediction assigned exactly 25% probability to the correct word

$$PP = \frac{1}{0.25} = 4$$

- The model behaves as though it has about four equally likely choices at every prediction

Average Probability and Perplexity

Average Probability	Perplexity
1.00	1
0.50	2
0.25	4
0.10	10
0.05	20

Lower perplexity means the model consistently assigns higher probabilities to the correct words

Why Is It Called Perplexity?

- Student A knows every test answer immediately
 - Effective number of choices: 1, perplexity: 1
- Student B usually narrows each question to two possible answers
 - Effective number of choices: 2, perplexity: 2
- Student C feels every question has about ten equally plausible answers
 - Effective number of choices: 10, perplexity: 10
- Higher perplexity means greater uncertainty

Relationship to Cross Entropy

- Models minimize the average negative log-likelihood (cross entropy)

$$H = -\frac{1}{N} \sum_{i=1}^N \log P(w_i)$$

- Perplexity is obtained by exponentiating this quantity

$$PP = e^H = \exp \left(-\frac{1}{N} \sum_{i=1}^N \log P(w_i) \right)$$

- Lower cross entropy $\Leftrightarrow$ lower perplexity

Conclusion

- Perplexity is one of the most widely used evaluation metrics for language models
- Perplexity measures the average uncertainty of a language model when predicting the next word
 - Low perplexity: the model consistently assigns high probability to the correct next word
 - High perplexity: the model is uncertain and spreads probability over many possible words
- Interpretation: the effective number of equally likely choices at each step
 - $PP = 4$ means about four equally plausible next words