N-gram Language Models: Smoothing, Interpolation, and Backoff

Nhóm 4

Ngày 23 tháng 2 năm 2026

Thành viên nhóm

Trần Lê Anh Pha — MSSV: 24521287

Trịnh Duy Hưng — MSSV: 24520610

Mục lục

Vấn đề khi sử dụng ước lượng khả năng cực đại

Vấn đề: bất kỳ tập hợp huấn luyện hữu hạn nào cũng sẽ thiếu một số chuỗi từ tiếng Anh

$\Rightarrow$ Khi dùng Maximum Likelihood Estimation (MLE), nếu một chuỗi từ **không xuất hiện trong tập huấn luyện**, thì xác suất của nó sẽ bằng 0

Vấn đề khi sử dụng ước lượng khả năng cực đại

Hệ quả: khi tính xác suất cả câu bằng cách nhân các xác suất: chỉ cần một n-gram có xác suất 0 $\rightarrow$ Cả câu có xác suất 0.

Điều này khiến:

- Perplexity bị vô hạn
- Mô hình đánh giá sai những câu hợp lệ
- Không thể tổng quát hóa tốt

$\Rightarrow$ Cần một phương pháp để giải quyết vấn đề này

Mục lục

Laplace smoothing

Phương pháp: cộng thêm 1 vào tất cả các tần số n-gram trước khi chúng ta chuẩn hóa chúng thành xác suất

Laplace smoothing không hiệu quả cho n-gram hiện đại, nhưng hữu ích để minh họa các khái niệm smoothing và làm mốc tham chiếu; đồng thời vẫn phù hợp cho một số bài toán như phân loại văn bản.

Laplace Smoothing

unigram probabilities trước Laplace smoothing:

$$P(w_i) = \frac{c_i}{N}$$

unigram probabilities sau Laplace smoothing:

$$P_{Laplace}(w_i) = \frac{c_i + 1}{N + V}$$

với V là số từ trong vocabulary

Laplace smoothing

MLE trước Laplace smoothing:

$$P_{MLE}(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{C(w_{n-1})}$$

MLE sau aplace smoothing:

$$P_{Laplace}(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)+1}{C(w_{n-1})+V} = \frac{C^*(w_{n-1}w_n)}{C(w_{n-1})}$$

Laplace smoothing

Thay vì nhìn trực tiếp vào xác suất đã smooth, ta có thể quy đổi nó về một "count mới" C^*:

$$C^*(w_{n-1}w_n) = \frac{(C(w_{n-1}w_n) + 1) \cdot C(w_{n-1})}{C(w_{n-1}) + V}$$

Laplace Smoothing

discount: tỉ lệ giữa count sau smoothing và count ban đầu, phản ánh mức độ suy giảm của các n-gram đã quan sát.

Laplace smoothing có thể làm thay đổi phân bố xác suất một cách đáng kể, đặc biệt với các n-gram có tần suất cao.

Nguyên nhân là một lượng lớn xác suất được phân bổ lại cho các n-gram chưa từng xuất hiện (count bằng 0).

Mục lục

Add-k smoothing

add-k: tương tự như laplace smoothing nhưng thay vì thêm 1 thì thêm k kí tự

$$P^*_{Add-k}(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n) + k}{C(w_{n-1}) + kV}$$

cần 1 phương pháp để chọn k: vd chọn trên **devset**

Hữu ích cho 1 vài công việc nhưng không làm tốt với language modeling, tạo số liệu với phương sai kém và thường discount không phù hợp

Mục lục

Language Model Interpolation

Nếu ngữ cảnh dài quá mà không có dữ liệu, ta **giảm bớt ngữ cảnh** để có nhiều dữ liệu hơn

Nếu không có **trigram**:

$$P(w_n \mid w_{n-2}w_{n-1})$$

ta dùng **bigram**:

$$P(w_n \mid w_{n-1})$$

nếu **bigram** cũng không có, ta dùng **unigram**:

$$P(w_n)$$

Language Model Interpolation

interpolation: Tính xác suất mới bằng cách nội suy các xác suất tri-gram, bi-gram và unigram.

$$P(w_n \mid w_{n-2}w_{n-1}) = \lambda_1 P(w_n) + \lambda_2 P(w_n \mid w_{n-1}) + \lambda_3 P(w_n \mid w_{n-2}w_{n-1})$$

với $\sum \lambda_i = 1$

Language Model Interpolation

slightly more sophisticated version: Mỗi trọng số λ được tính dựa vào ngữ cảnh:

$$\begin{aligned} P(w_n|w_{n-2}w_{n-1}) = \lambda_1(w_{n-2:n-1})P(w_n)+ \\ \lambda_2(w_{n-2:n-1})P(w_n|w_{n-1})+ \\ \lambda_3(w_{n-2:n-1})P(w_n|w_{n-2}w_{n-1}) \end{aligned}$$

Language Model Interpolation

Chọn λ:

Có thể dùng **held-out** để chọn λ

held-out: là tập hợp dữ liệu huấn luyện bổ sung, sử dụng để xác lập các giá trị λ

Mục lục

Stupid Backoff

backoff: nếu n-gram có zero counts thì lùi về (n-1) gram, tiếp tục như vậy cho tới hết zero counts.

discount: giảm trọng số các n-gram bậc cao để giữ lại một phần xác suất cho các n-gram bậc thấp hơn.

Stupid Backoff

Từ bỏ ý tưởng cố gắng biến mô hình ngôn ngữ thành một phân phối xác suất thực sự.

Nếu một n-gram bậc cao có zero counts, chúng ta đơn giản quay về n-gram bậc thấp hơn, được cân bằng bằng một trọng số cố định

$$S(w_i \mid w_{i-N+1:i-1}) = \begin{cases} \frac{\text{count}(w_{i-N+1:i})}{\text{count}(w_{i-N+1:i-1})} & \text{if } \text{count}(w_{i-N+1:i}) > 0 \\ \lambda\, S(w_i \mid w_{i-N+2:i-1}) & \text{otherwise} \end{cases}$$

Phương pháp backoff kết thúc ở unigram, có điểm $S(w) = count(w)/N$

Mục lục

Câu hỏi

Tại sao smoothing là cần thiết trong n-gram?

Khi vocabulary lớn, Laplace smoothing có vấn đề gì?

Add-k khác Laplace ở điểm nào?

Backoff khác interpolation thế nào?

Vì sao phải discount trong backoff chuẩn?