ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC VÀ KỸ THUẬT MÁY TÍNH

HỌC MÁY - CO3117

Bài tập mở rộng

Gãn nhãn từ loại sử dụng mô hình Markov ẩn

POS tagging using Hidden Markov Models

Giảng viên hướng dẫn: TS. Lê Thành Sách
Lớp: TN01
Nhóm: DNA05

THÀNH PHỐ HỒ CHÍ MINH, THÁNG 11 - 2025

BÁO CÁO KẾT QUẢ LÀM VIỆC NHÓM

STT	MSSV	Họ và Tên	Nhiệm vụ phụ trách	Mức độ hoàn thành
1	2310167	Tăng Hồng Ái	Thu thập và phân tích tập dữ liệu **Brown Corpus**. Thực hiện EDA: thống kê số lượng câu, số lượng từ, phân bố nhãn từ loại; mô tả đặc điểm ngữ liệu và viết phần báo cáo tương ứng.	100%
2	2310510	Phạm Khánh Duy	Xây dựng và cài đặt mô hình **Hidden Markov Model (HMM)** cho bài toán POS Tagging. Hiện thực thuật toán Viterbi, xử lý vấn đề từ vựng chưa biết bằng phương pháp **pseudo-word**, và viết phần báo cáo mô hình.	100%
3	2312506	Nguyễn Trần Yến Nhi	Thiết kế quy trình huấn luyện – kiểm thử, thực nghiệm mô hình, đánh giá hiệu quả với các chỉ số Accuracy, phân tích lỗi. Tổng hợp kết quả và viết phần báo cáo đánh giá mô hình.	100%

Bảng 1: Bảng phân công công việc và mức độ hoàn thành

Mục lục

1 Giới thiệu chung **3**
1.1 Mô hình Markov ẩn (Hidden Markov Model) . . . 3
1.1.1 Ứng dụng thực tiễn . . . 3
1.2 Parts of Speech (POS) tagging . . . 4

2 Xây dựng mô hình Hidden Markov Model (HMM) cho POS Tagging **5**
2.1 Định nghĩa bài toán . . . 5
2.2 Thuật toán huấn luyện . . . 5
2.3 Vấn đề từ chưa xuất hiện (Unknown Vocabulary) . . . 5
2.4 Hiện thực tổng quan . . . 6

3 Huấn luyện và đánh giá **7**
3.1 Tập dữ liệu . . . 7
3.2 Quá trình huấn luyện . . . 8
3.3 Kết quả huấn luyện . . . 8
3.4 Đánh giá kết quả . . . 9

4 Kết luận **10**

5 Phụ lục **11**

Tài liệu tham khảo **12**

1 Giới thiệu chung

1.1 Mô hình Markov ẩn (Hidden Markov Model)

Định nghĩa

Mô hình Markov ẩn (Hidden Markov Model - HMM) là một mô hình xác suất dùng để mô tả một hệ thống có các trạng thái ẩn (hidden states) mà không quan sát trực tiếp được, nhưng sinh ra các quan sát (observations) phụ thuộc vào trạng thái đó. HMM dựa trên giả thiết trạng thái hiện tại chỉ phụ thuộc vào trạng thái ngay trước đó.

Cấu trúc cơ bản

Một HMM điển hình gồm:

- Tập trạng thái ẩn $S = \{s_1, s_2, \ldots, s_N\}$.
- Tập quan sát $O = \{o_1, o_2, \ldots, o_M\}$.
- Ma trận xác suất chuyển trạng thái $A = [a_{ij}]$, với a_{ij} là xác suất chuyển tự trạng thái s_i sang s_j.
- Ma trận xác suất phát xạ $B = [b_{jk}]$, với b_{jk} là xác suất sinh ra quan sát o_k từ trạng thái ẩn s_j.
- Xác suất trạng thái ban đầu $\pi = [\pi_i]$, với π_i là xác suất trạng thái ẩn bắt đầu là s_i.

Thuật toán tiêu biểu

HMM sử dụng ba thuật toán cơ bản:

1. **Forward Algorithm:** Tính xác suất quan sát được chuỗi dữ liệu, dùng để đánh giá mô hình.
2. **Viterbi Algorithm:** Tìm chuỗi trạng thái ẩn tối ưu có xác suất lớn nhất với một chuỗi quan sát.
3. **Baum-Welch Algorithm:** Huấn luyện HMM qua Expectation-Maximization (EM) để ước lượng các tham số từ dữ liệu quan sát.

1.1.1 Ứng dụng thực tiễn

HMM được ứng dụng rộng rãi trong nhiều lĩnh vực, ví dụ:

- Nhận dạng giọng nói (speech recognition)
- Nhận dạng chữ viết tay (handwriting recognition)
- Dự đoán chuỗi sinh học (gene prediction)

- Phân tích chuỗi thời gian (time series analysis)

Trong nghiên cứu này, chúng ta tập trung vào một bài toán tiêu biểu trong NLP là **POS tagging**.

1.2 Parts of Speech (POS) tagging

Định nghĩa

POS tagging là bài toán gán nhãn từ loại cho từng từ trong câu, ví dụ: danh từ (NOUN), động từ (VERB), tính từ (ADJ), trạng từ (ADV), giới từ (ADP), mạo từ (DET), liên từ (CONJ), đại từ (PRON), v.v. Đây là bước tiền xử lý quan trọng trong NLP để hệ thống hiểu được cấu trúc cú pháp và ngữ nghĩa của văn bản.

Ví dụ minh họa:

Cho câu tiếng Anh: "The quick brown fox jumps over the lazy dog."
POS tagging sẽ gán nhãn như sau:

- The: DET
- quick: ADJ
- brown: ADJ
- fox: NOUN
- jumps: VERB
- over: ADP
- the: DET
- lazy: ADJ
- dog: NOUN
- . : PUNCT

Mục tiêu

Từ chuỗi câu, mô hình POS tagging dự đoán nhãn POS cho từng từ, hỗ trợ:

- Phân tích cú pháp (syntactic parsing)
- Trích xuất thông tin (information extraction)
- Dịch máy (machine translation)
- Phân tích ngôn ngữ học (linguistic analysis)

2 Xây dựng mô hình Hidden Markov Model (HMM) cho POS Tagging

2.1 Định nghĩa bài toán

POS Tagging là bài toán gán nhãn từ loại cho từng từ trong câu. Trong HMM:

- **Observations (quan sát)**: các từ trong câu.
- **Hidden states (trạng thái ẩn)**: các nhãn từ loại như `NOUN`, `VERB`, ...

Mục tiêu là tìm chuỗi trạng thái ẩn $s_1, s_2, \ldots, s_T$ sao cho xác suất sinh ra chuỗi quan sát $o_1, \ldots, o_T$ là lớn nhất. Nói cách khác, ta tìm chuỗi nhãn sao cho xác suất sinh ra được chuỗi từ trong câu văn đầu vào là lớn nhất. Điều này được thực hiện bằng **thuật toán Viterbi**.

2.2 Thuật toán huấn luyện

Với dữ liệu có nhãn, HMM có thể được huấn luyện bằng **Maximum Likelihood Estimation (MLE)**:

- Xác suất khởi đầu: $\pi_i = \frac{\text{Số lần trạng thái } s_i \text{ xuất hiện đầu câu}}{\text{Tổng số câu}}$
- Ma trận chuyển tiếp: $A_{ij} = \frac{c(s_i \to s_j)}{\sum_k c(s_i \to s_k)}$ với $c(s_i \to s_j)$ là số lần trạng thái ẩn s_j xuất hiện ngay sau s_i.
- Ma trận phát xạ: $B_{ik} = \frac{c(s_i \to o_k)}{\sum_j c(s_i \to o_j)}$ với $c(s_i \to o_j)$ là số lần trạng thái ẩn s_i sinh ra quan sát o_j.

2.3 Vấn đề từ chưa xuất hiện (Unknown Vocabulary)

Trong HMM, xác suất phát xạ được ước lượng từ dữ liệu huấn luyện. Do đó, một vấn đề lớn là nếu từ x chưa xuất hiện trong dữ liệu huấn luyện, thì xác suất sinh ra x từ nhãn s sẽ bằng 0 với mọi nhãn s. Tức là, xác suất sinh ra chuỗi từ trong câu văn đầu vào sẽ bằng 0 với mọi chuỗi nhãn $s_1 \ldots s_T$, khiến thuật toán Viterbi thất bại trên câu chứa từ mới.

Giải pháp: Pseudo-word

Để khắc phục, các từ xuất hiện ít lần hoặc chưa xuất hiện trong tập huấn luyện được ánh xạ sang một tập **pseudo-word** hạn chế. Các lớp pseudo-word được định nghĩa như sau:

- `<PUNCT>`: các dấu câu `.,;:!?""'()[]{}`.
- `<fourDigitNum>`, `<twoDigitNum>`, `<othernum>`: số có 4 chữ số, 2 chữ số, hoặc số khác.
- `<containsDigitAndSlash>`, `<containsDigitAndDash>`, `<containsDigitAndComma>`, `<containsDigitAndPeriod>`, `<containsDigitAndAlpha>`: các từ chứa ký tự số kết hợp `/`, `-`, `,`, `.`, hoặc chữ cái.

- `<ALLCAPS>`: chữ hoa toàn bộ.
- `<capPeriod>`: chữ hoa theo kiểu viết tắt (ví dụ: `M.`).
- `<initCap>`: chữ cái đầu viết hoa, các chữ sau viết thường.
- `<suffix_ing>`, `<suffix_ed>`, `<suffix_ly>`: các từ có hậu tố `-ing`, `-ed`, `-ly`.
- `<lowercase>`: chữ thường hoàn toàn.
- `<other>`: các từ còn lại.

Cách áp dụng:

1. Chọn ngưỡng γ (ví dụ $\gamma = 5$).
2. Mọi từ xuất hiện ít hơn γ lần trong huấn luyện được thay bằng pseudo-word tương ứng.
3. Khi dự đoán, từ chưa xuất hiện trong huấn luyện cũng ánh xạ sang pseudo-word tương ứng.

Việc ánh xạ này giúp "đóng" từ vựng: mọi từ trong dữ liệu kiểm tra sẽ xuất hiện ít nhất một lần trong huấn luyện (dưới dạng pseudo-word), tránh xác suất phát xạ bằng 0 và giữ thông tin về dạng từ (chữ hoa, số, hậu tố...).

2.4 Hiện thực tổng quan

Mô hình POS tagging được xây dựng dựa trên HMM (Hidden Markov Model) với các thành phần chính:

- **HiddenMarkovModel**: lớp cơ bản triển khai HMM sử dụng `PyTorch`, bao gồm:
  - Ma trận chuyển tiếp (A), ma trận phát xạ (B) và xác suất khởi đầu (π).
  - Thuật toán `forward`, `backward` tính xác suất chuỗi quan sát.
  - Thuật toán `Viterbi` tìm chuỗi trạng thái ẩn có xác suất cao nhất.
  - Thuật toán `Baum-Welch` và huấn luyện có giám sát (MLE) để cập nhật tham số HMM.
- **HMMUtils**: tiện ích xử lý token, ánh xạ từ các từ hiếm hoặc từ đặc biệt sang `pseudo-word` dựa trên các quy tắc regex, ví dụ: `<PUNCT>`, `<fourDigitNum>`, `<initCap>`, `<suffix_ing>`...
- **POS_HMM**: wrapper cho tác vụ POS tagging, chịu trách nhiệm:
  - Tiền xử lý dữ liệu train và mapping pseudo-word.
  - Khởi tạo HMM với số trạng thái (tags) và số quan sát (từ/vocab).
  - Gọi hàm huấn luyện `train_supervised_MLE`.
  - Dự đoán nhãn cho câu hoặc batch câu (`predict_sentence`, `predict_batch`).
  - Tính toán các chỉ số đánh giá như `accuracy`, `precision`, `recall`, `F1`.

3 Huấn luyện và đánh giá

3.1 Tập dữ liệu

Trong quá trình huấn luyện mô hình HMM cho bài toán *Parts-of-Speech Tagging*, chúng ta sử dụng **Brown Corpus** — một trong những bộ dữ liệu ngôn ngữ cổ điển và phổ biến nhất được cung cấp sẵn trong thư viện `NLTK`. Brown corpus bao gồm hơn 1 triệu từ tiếng Anh được gán nhãn ngữ pháp đầy đủ.

- Tổng số câu: **57.340 câu**.
- Tổng số từ: khoảng **1161192 từ**.
- Số từ vựng duy nhất: khoảng **49815 từ**.
- Số nhãn POS: **12 nhãn** theo chuẩn `Universal Tagset`.

Bộ dữ liệu được chuẩn hóa bằng `universal_tagset` để giảm độ phức tạp, chỉ giữ lại 12 nhãn POS phổ biến, gồm có: `NOUN`, `VERB`, `ADJ`, `ADV`, `PRON`, `DET`, `ADP`, `CONJ`, `PRT`, `NUM`, `.`, `X`.

Dữ liệu sau khi tải được phân tích để hiểu rõ hơn về phân bố nhãn, độ dài câu và tần suất xuất hiện từ.

Kết quả cho thấy:

- **NOUN** (danh từ) chiếm tỷ lệ lớn nhất: khoảng 23.7%.
- **VERB** (động từ) chiếm khoảng 15.7%.
- **ADP**, **DET**, **ADJ** cũng chiếm tỷ lệ cao.

Bảng 2 trình bày phân bố nhãn POS trong toàn bộ tập dữ liệu.

Bảng 2: Phân bố các nhãn POS trong Brown Corpus

Nhãn POS	Số lượng	Tỷ lệ (%)
NOUN	275,558	23.73
VERB	182,750	15.74
.	147,565	12.71
ADP	144,766	12.47
DET	137,019	11.80
ADJ	83,721	7.21
ADV	56,239	4.84
PRON	49,334	4.25
CONJ	38,151	3.29
PRT	29,829	2.57
NUM	14,874	1.28
X	1,386	0.12

Ngoài ra, độ dài câu trung bình trong Brown Corpus là khoảng **20 từ**, với câu dài nhất lên tới **180 từ**. Những thống kê này giúp điều chỉnh kích thước bộ nhớ đệm và giới hạn độ dài khi xử lý huấn luyện mô hình HMM.

3.2 Quá trình huấn luyện

Tập dữ liệu được chia theo tỷ lệ:

$$\text{Train:Test} = 80 : 20$$

Tập huấn luyện được sử dụng để ước lượng các ma trận:

- Ma trận chuyển trạng thái A giữa các nhãn POS.
- Ma trận phát xạ B từ nhãn POS sang các từ quan sát.
- Phân phối khởi tạo π.

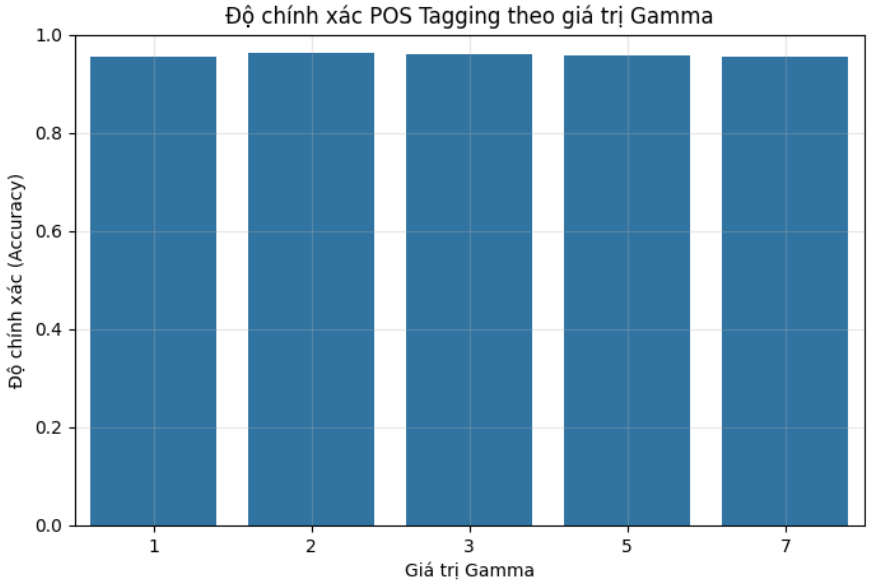
Mô hình được huấn luyện với nhiều giá trị siêu tham số `gamma` khác nhau để tìm ra giá trị tối ưu, nhằm điều chỉnh mức độ làm trơn (smoothing) cho các xác suất nhỏ.

3.3 Kết quả huấn luyện

Kết quả tổng hợp thể hiện trong **Bảng 3**. Mô hình đạt độ chính xác cao nhất tại **gamma = 2**, tương ứng với độ chính xác khoảng **96.4%**.

Bảng 3: Kết quả huấn luyện và đánh giá POS_HMM theo giá trị Gamma

Gamma	Accuracy	Precision	Recall	F1
1	0.9539	0.9572	0.9539	0.9551
2	**0.9641**	**0.9643**	**0.9641**	**0.9642**
3	0.9613	0.9617	0.9613	0.9615
5	0.9576	0.9579	0.9576	0.9577
7	0.9549	0.9551	0.9549	0.9549



Hình 3.1: Độ chính xác POS Tagging theo giá trị Gamma

3.4 Đánh giá kết quả

Kết quả thực nghiệm cho thấy mô hình POS_HMM hoạt động ổn định trên tập dữ liệu Brown Corpus với độ chính xác tổng thể cao. Khi thay đổi giá trị `gamma`, có thể quan sát sự khác biệt rõ rệt trong hiệu năng mô hình:

- **Gamma = 2** cho kết quả tốt nhất với độ chính xác khoảng **96.4%**. Đây là giá trị cân bằng giúp mô hình cập nhật xác suất hợp lý khi phân phối quan sát hoặc chuyển trạng thái có tần suất thấp.
- Khi `gamma` nhỏ hơn (ví dụ: 1), mô hình trở nên nhạy cảm với các phân phối hiếm, dẫn đến sai số cao hơn.
- Ngược lại, với `gamma` lớn (5 hoặc 7), việc bổ sung pseudo-count quá nhiều làm phân phối xác suất trở nên "phẳng" hơn, khiến mô hình giảm khả năng phân biệt giữa các nhãn.

Ngoài ra, phần lớn sai số xuất hiện ở các nhãn ít phổ biến như `NUM`, `X` hoặc những từ có tần suất rất thấp trong tập huấn luyện. Đây là hạn chế thường gặp của các mô hình thống kê dựa trên tần suất.

Tổng thể, mô hình HMM cho thấy khả năng gán nhãn POS hiệu quả, đồng thời duy trì độ chính xác cao trên tập dữ liệu lớn và đa dạng như Brown Corpus. Điều này khẳng định rằng cấu trúc chuỗi ẩn và các giả định Markov vẫn phù hợp để mô hình hóa quy luật ngữ pháp cơ bản trong tiếng Anh.

4 Kết luận

Trong bài nghiên cứu này, chúng tôi đã xây dựng và huấn luyện mô hình **Hidden Markov Model (HMM)** cho bài toán **gán nhãn từ loại (POS Tagging)** trên tập dữ liệu **Brown Corpus**. Quá trình thực nghiệm cho thấy HMM là một phương pháp xác suất hiệu quả, có khả năng mô hình hóa tốt mối quan hệ chuỗi giữa các nhãn ngữ pháp trong câu tiếng Anh.

Mô hình được cài đặt hoàn chỉnh với các thành phần:

- Ma trận chuyển trạng thái (*Transition matrix*) giữa các nhãn từ loại.
- Ma trận phát xạ (*Emission matrix*) giữa nhãn và từ quan sát.
- Phân phối khởi tạo (*Initial probabilities*) cho trạng thái đầu tiên.
- Giải thuật **Viterbi** để tìm chuỗi nhãn ẩn có xác suất cao nhất.

Bên cạnh đó, mô hình còn được mở rộng để xử lý vấn đề **"Unknown Vocabulary"** bằng cách áp dụng kỹ thuật **pseudo-word mapping**, giúp mô hình nhận diện được các từ hiếm hoặc chưa từng xuất hiện trong tập huấn luyện. Phương pháp này giúp đóng kín không gian từ vựng và duy trì độ chính xác của hệ thống trong thực tế.

Kết quả thực nghiệm đạt độ chính xác trung bình **96,4%** trên tập kiểm thử, vượt trội so với phương pháp gán nhãn theo tần suất cao nhất (*Most Frequent Tag per word baseline*). Điều này chứng tỏ rằng HMM vẫn là một mô hình cơ bản đủ tốt cho các bài toán gán nhãn chuỗi, đặc biệt trong bối cảnh dữ liệu hạn chế và yêu cầu mô hình có khả năng diễn giải cao.

Tổng kết lại, mô hình HMM cho POS Tagging vừa đảm bảo tính đơn giản, dễ huấn luyện, vừa đạt hiệu quả cao, là một nền tảng vững chắc để nghiên cứu và mở rộng sang các bài toán xử lý ngôn ngữ tự nhiên phức tạp hơn.

5 Phụ lục

Dataset: Brown Corpus từ thư viện NLTK, nhãn từ loại được chuẩn hóa theo Universal Tagset.

Github: DNA05's github page

Colab Notebook: DNA05-BTMR

Tài liệu tham khảo

[1] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R and Python*. Springer, 2nd edition, 2021.

[2] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[3] Scikit-learn developers. Classification metrics — scikit-learn documentation. https://scikit-learn.org/stable/modules/model_evaluation.html, 2025. Accessed: 2025-10-25.

[4] Scikit-learn developers. Preprocessing data — scikit-learn documentation. https://scikit-learn.org/stable/modules/preprocessing.html, 2025. Accessed: 2025-10-25.