**Trường đại học Khoa học Tự nhiên**

**Khoa Công nghệ Thông tin**



**BÁO CÁO ĐỒ ÁN MÔ HÌNH MARKOV ẨN (Phần 1: Lý thuyết)**

**Môn**: Toán ứng dụng & thống kê

Nhóm sinh viên thực hiện:

20120018 – Trần Kiều Minh Lâm

20120067 – Nguyễn Phạm Bá Duy

20120316 – Nguyễn Hoàng Lâm

Giảng viên hướng dẫn: Nguyễn Đình Thúc, Nguyễn Văn Quang Huy, Võ Nam Thục Đoan

**Năm học:** 2021 – 2022

*1. Các thành phần của một mô hình Markov ẩn là gì? Chúng khác gì với mô hình Markov?*Các thành phần cơ bản của một mô hình Markov ẩn (Hidden Markov Model - HMM) bao gồm:

* Không gian mẫu của trạng thái ẩn (state space)
* Không gian mẫu của trạng thái quan sát được (observation space)
* Chuỗi trạng thái ẩn (state sequence)
* Chuỗi trạng thái quan sát được (observation sequence)
* Ma trận xác suất ban đầu của trạng thái ẩn (starting state probs)
* Ma trận xác suất chuyển trạng thái ẩn (transition probs)
* Ma trận xác suất đầu ra trạng thái quan sát được (emission probs)

So với mô hình Markov ẩn, mô hình Markov chỉ có 4 thành phần cơ bản và các tham số không bị ẩn, bao gồm:

* Không gian mẫu của trạng thái (state space)
* Chuỗi trạng thái (state sequence)
* Ma trận xác suất ban đầu của trạng thái (starting state probs)
* Ma trận xác suất chuyển trạng thái (transition probs)

*2. Các giả thiết (assumption) đặt ra cho mô hình Markov ẩn là gì? Tìm ví dụ các bài toán mà các giả thiết này hợp lý và bất hợp lý.*Mô hình Markov ẩn có đặt ra các giả thiết sau, bao gồm:

* Các trạng thái quan sát được chỉ nằm trong một không gian mẫu hữu hạn nhất định.
* Các trạng thái ẩn chỉ nằm trong một không gian mẫu hữu hạn nhất định.
* Xác suất của trạng thái ẩn tại một thời điểm t chỉ phụ thuộc vào trạng thái ẩn tại thời điểm t - 1 trước đó (Markov assumption)
* Mỗi trạng thái quan sát được tại thời điểm t chỉ phụ thuộc trực tiếp vào trạng thái ẩn tại thời điểm t đó (independence assumption).

Chính vì những giả thiết đó, dẫn đến sự bất hợp lý của bài toán dự đoán giá cổ phiếu. Bởi vì, giá của một cổ phiếu lên xuống phụ thuộc giá cổ phiếu của nhiều ngày trước đó.

Nhưng, những giả thiết đó lại hợp lý cho bài toán đánh nhãn các thành phần trong câu (POS tagging). Bởi vì, một từ trong thành phần trong câu chỉ phụ thuộc phần lớn vào từ trước nó.

Ví dụ, sau **tính từ** thường là sự xuất hiện của **danh từ** mà không phụ thuộc vào các từ trước đó.

*3. Cho một mô hình Markov ẩn với các tham số đã biết, thuật toán tiến trước (forward algorithm) được dùng để xác định độ hợp lý (likelihood) của một chuỗi quan sát (observation). Mô tả và đánh giá độ phức tạp của thuật toán tiến trước.*

* **Tham số đầu vào**: ma trận xác suất chuyển trạng thái ẩn , ma trận xác suất đầu ra trạng thái quan sát được và chuỗi trạng thái quan sát được .
* **Kết quả đầu ra:** xác suất để xuất hiện chuỗi trạng thái quan sát được khi biết trước ma trận xác suất chuyển trạng thái ẩn và ma trận xác suất đầu ra trạng thái quan sát được. Kí hiệu:
* **Kí hiệu:**
  + Chuỗi trạng thái quan sát được kí hiệu là .
  + Chuỗi trạng thái ẩn kí hiệu là .
  + nghĩa là xác suất để chuyển từ trạng thái ẩn sang trạng thái ẩn trong ma trận xác suất chuyển trạng thái ẩn
  + nghĩa là xác suất để thu được trạng thái quan sát được từ trạng thái ẩn trong ma trận xác suất đầu ra trạng thái quan sát được
  + chính là ma trận xác suất ban đầu của trạng thái ẩn.
* **Ý tưởng:** Bài toán sử dụng phương pháp quy hoạch động với:
* **Mã giả:**

**Bước 1:** Khởi tạo:

Duyệt từng trạng thái ẩn trong không gian mẫu của nó:

**Bước 2**: Duyệt

Duyệt từng trạng thái ẩn trong không gian mẫu của nó:

**Bước 3:** Kết quả

* **Đánh giá độ phức tạp:** với là kích thước của không gian mẫu tham số ẩn.

4. Cho một mô hình Markov ẩn với các tham số đã biết, thuật toán Viterbi được dùng để xác định chuỗi trạng thái (state) khả dĩ nhất. Mô tả và đánh giá độ phức tạp của thuật toán Viterbi.

**Tham số đầu vào:** ma trận xác suất chuyển trạng thái ẩn , ma trận xác suất đầu ra của trang thái quan sát được và chuỗi trạng thái quan sát được

**Kết quả đầu ra:** chuỗi trạng thái ẩn có xác suất tạo ra được chuỗi trạng thái quan sát được là cao nhất.

**Kí hiệu:**

* + Chuỗi trạng thái quan sát được kí hiệu là .
  + Chuỗi trạng thái ẩn kí hiệu là .
  + nghĩa là xác suất để chuyển từ trạng thái ẩn sang trạng thái ẩn trong ma trận xác suất chuyển trạng thái ẩn
  + nghĩa là xác suất để thu được trạng thái quan sát được từ trạng thái ẩn trong ma trận xác suất đầu ra trạng thái quan sát được
  + chính là ma trận xác suất ban đầu của trạng thái ẩn.

**Ý tưởng:** Bài toán cũng sử dụng phương pháp quy hoạch động giống với thuật toán tiến trước (forward algorithm) chỉ khác là thay vì tìm tổng ở thuật toán tiến trước, thuật toán Viterbi tìm giá trị lớn nhất với:

Ngoài ra, cần dùng một mảng để truy vết tìm ra các trạng thái ẩn mà đạt giá trị lớn nhất. Với:

**Mã giả:**

**Bước 1:** Khởi tạo:

Duyệt từng trạng thái ẩn trong không gian mẫu của nó:

**Bước 2**: Duyệt

Duyệt từng trạng thái ẩn trong không gian mẫu của nó:

**Bước 3:** Truy vết theo mảng để tìm

**Đánh giá độ phức tạp:** với là kích thước của không gian mẫu tham số ẩn.

5. Cho một chuỗi quan sát, giả sử ta cho rằng chuỗi quan sát này được sinh ra từ một mô hình Markov ẩn với tham số chưa biết, thuật toán Baum-Welch được dùng để ước lượng các tham số này. Thuật toán Baum-Welch là trường hợp đặc biệt của thuật toán Kỳ vọng-Tối ưu (Expectation-Maximization, hay EM). Thuật toán này gồm 2 bước: bước E (Expectation, hay kỳ vọng) và bước M (Maximization, hay tối ưu).

* Mô tả thuật toán Kỳ vọng-Tối ưu tổng quát.

**Tham số đầu vào:** Một mô hình thống kê có tập dữ liệu quan sát được, một tập dự liệu bị ẩn và một tập các tham số chưa biết.

**Kết quả đầu ra:** Một tập các tham số đã được hợp lý cực đại (maximum likelihood estimation)

**Mã giả:**

Bước 1: Khởi tạo các tham số chưa biết bằng các giá trị ngẫu nhiên

Bước 2: Bước E (Expectation step): sử dụng những dự liệu quan sát được, ước tính giá trị cho các dữ liệu bị thiếu

Bước 3: Bước M (Maximization step): sử dụng dữ liệu sinh ra ở bước E để cập nhật các tham số

Bước 4: Lặp lại bước 2 và 3 cho tới khi đạt được sự hội tụ.

* Mô tả và đánh giá độ phức tạp của bước E và bước M của thuật toán Baum-Welch.

**Tham số đầu vào:** chuỗi trạng thái quan sát được .

**Kết quả đầu ra:** Ước lượng ma trận xác suất chuyển tham số ẩn , ma trận xác suất đầu ra tham số quan sát được mà có khả năng tạo ra chuỗi quan sát được là lớn nhất.

**Kí hiệu:**

* + Chuỗi trạng thái quan sát được kí hiệu là .
  + Chuỗi trạng thái ẩn kí hiệu là .
  + nghĩa là xác suất để chuyển từ trạng thái ẩn sang trạng thái ẩn trong ma trận xác suất chuyển trạng thái ẩn
  + nghĩa là xác suất để thu được trạng thái quan sát được từ trạng thái ẩn trong ma trận xác suất đầu ra trạng thái quan sát được

**Ý tưởng:**

Khởi tạo các giá trị ngẫu nhiên cho 2 ma trận và , sau đó lặp lại việc cập nhật 2 ma trận đó. Mỗi lần lặp lại việc cập nhật gồm 2 bước: bước E (Expectation step) và bước M (Maximization step).

**Bước E:** Giả sử ta biết trước 2 ma trận A và B, cần tính các tham số sau:

Với: đếm số lần mà trạng thái ẩn tại thời điểm bằng với

đếm số lần mà và xuất hiện tại thời điểm thứ và

Chúng được gọi là số lượng giả . Để tính được số lượng giả, ta cần tính thông qua 2 tham số, gồm:

Có thể tính được và bằng cách sử dụng quy hoạch động.

Khi biết được và . Ta có thể tính được và bằng công thức:

**Bước M:** Sau khi biết được và ta có thể sử dụng hợp lý cực đại (maximum likelihood estimation) để cập nhật các giá trị cho và

Với

**Tài liệu tham khảo**

<https://www.cs.cmu.edu/~tbergkir/11711fa17/recitation4_notes.pdf>