Mô hình Markov ẩn (Hidden Markov Model - HMM)

HMM là một mô hình thống kê được sử đụng dể mô tả quy trình Markov với các tham số không biết được ẩn đi. Mô hình xác định các tham số ẩn cho quy trình thông qua chuỗi các quan sát rồi sử dụng các tham số đó cho việc phân tích về sau [**rabiner1989tutorial**]

Hình 0.1: Cấu trúc của HMM

Hình 2.7 mô tả cấu trúc của HMM. X là biến ẩn mà bên quan sát không phát hiện ra. x_t biểu diễn trạng thái tại thời gian t. Mỗi biến quan sát được y chỉ phụ thuộc vào x_t, và x_t liên quan đến trạng thái trước đó là x_{t-1}

Nếu các trạng thái ẩn có N giá trị, thì tại thời điểm t sẽ có thể nhận N giá trị . Vì vậy, có tối đa N^2 khả năng từ một trạng thái ẩn chuyển sang trạng thái ẩn tiếp theo. Với biến quan sát y có M có thể nhận, mỗi giá trị trong trạng thái ẩn đều có xác suất chuyển tới tới mọi biến quan sát khác. Vì vậy, nếu chuỗi quan sát gọi là Y, chuỗi trạng thái ẩn là X, trong đó:

$$X = (x_0, x_1, ..., x_n)$$

$$Y = (y_0, y_1, ..., y_n)$$

Xác suất của chuỗi Y qua mô hình HMM có thể được biểu diễn bằng biểu thức sau:

$$P(Y) = \sum_X P(Y|X)P(X)$$

Với dữ liệu chưa được gán nhãn, phương pháp thống kê của các tham số không thể tính toán được trực tiếp do sự tồn tại của các biến ẩn. Khi đó thuật

toán Cực đại hóa kì vọng (Expectation Maximization - EM) được sử dụng để lặp cho đến khi hội tụ để ra được tham số cho mô hình. Thuật toán EM được chia làm 2 phần là E và M. Ở bước E, thuật toán sử dụng tham số đã biết để tính ra hậu phân phối của biến ẩn $P(T|S, \theta^{old})$. Ở bước M, thuật toán tính giá trị kì vọng cực đại của log-likelihood dựa vào hậu phân phối. Kì vọng ở đây là một hàm có tham số θ, và hàm sẽ cực đại hóa kì vọng của hàm $Q = (\theta, \theta^{old})$ để có thể có được nghiệm của θ rồi dùng nghiệm này đưa vào θ^{old} trong bước E. Sau đó việc này được lặp lại cho đến khi hội tụ. (zhang2001segmentation)

Sau khi quá trình huấn luyện kết thúc, sử dụng mô hình HMM có thể dự đoán chuỗi câu mới. Khi mô hình nhận vào một tập các chuỗi có được từ việc quan sát, mô hình sẽ tìm ra chuỗi ẩn phù hợp nhất sử dụng thuật toán Viterbi ([**ghahramani2001introduction**])

Mô hình HMM đã được sử dụng rộng rãi để giải quyết nhiều bài toán về xử lý ngôn ngữ tự nhiên như nhận diieenj giọng nói, dịch máy, gán nhãn, nhận diện tên thực thể ... [**ponomareva2007biomedical**]