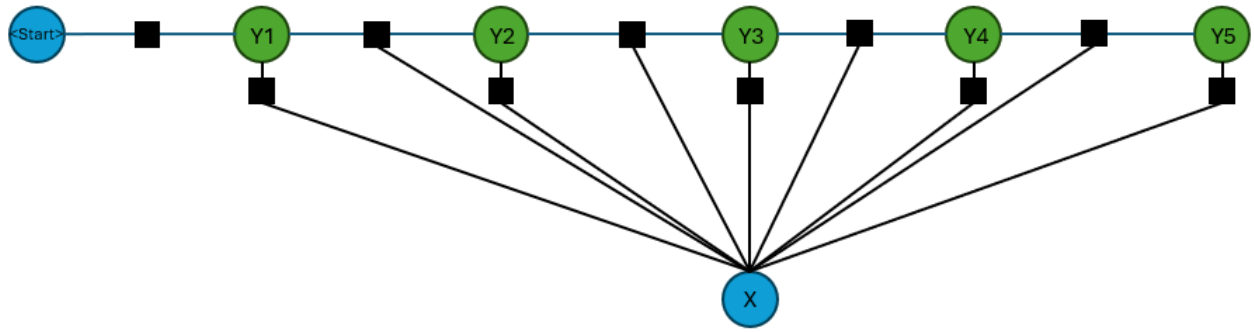
1. Định nghĩa



Hình 1. Linear-Chain CRFs dạng factor với các ô vuông là các hàm phụ thuộc giữa các nút

Gọi X là biến ngẫu nhiên đại diện cho chuỗi dữ liệu đầu vào cần được gán nhãn, Y là biến ngẫu nhiên đại diện cho chuỗi nhãn tương ứng với chuỗi dữ liệu X. Tất cả các thành phần Y_i của Y thuộc một tập nhãn hữu hạn $\mathcal{Y}$ (tập các nhãn có thể có). Ω_x là các trường hợp có thể có của chuỗi X, Ω_y là các trường hợp có thể có của chuỗi nhãn Y.

Giả định cả X và Y đều được coi là biến ngẫu nhiên phân phối chung (jointly distributed), nghĩa là chúng có mối liên hệ xác suất với nhau, và xác suất $P(X, Y)$ là dương nghiêm ngặt ($P(X = x, Y = y) > 0, \forall x, y$).

CRFs [8] là một mô hình phân biệt, tập trung vào việc xây dựng mô hình xác suất có điều kiện P(Y|X). CRFs dự đoán chuỗi nhãn Y dựa trên chuỗi dữ liệu X đã cho. CRFs không cố gắng mô hình hóa xác suất của X (tức là P(X)), mà chỉ quan tâm đến xác suất của Y khi biết X.

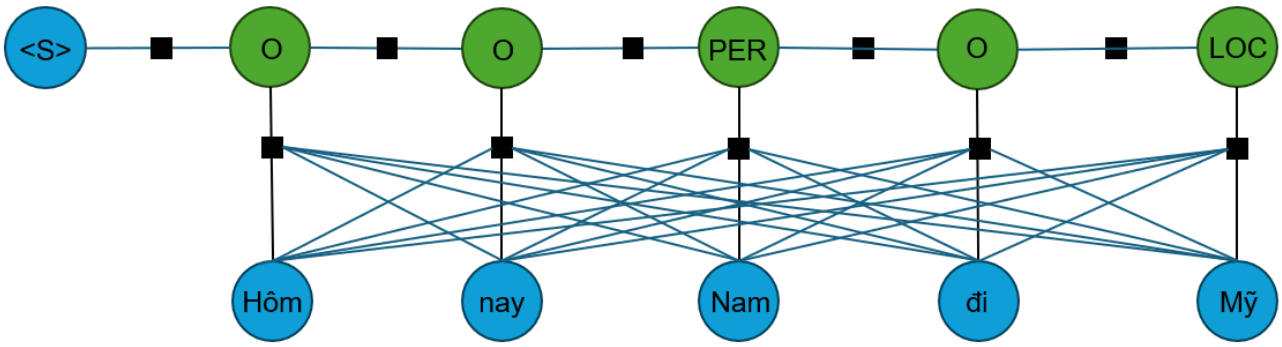
Định nghĩa: Cho đồ thị $G = (V, E)$ sao cho $Y = (Y_v)_{v \in V}$, nghĩa là Y được chỉ mục hóa theo các đỉnh của đồ thị G. Khi đó, cặp (X, Y) là một trường ngẫu nhiên điều kiện (conditional random fields - CRFs) trong trường hợp, khi biết X, các biến ngẫu nhiên Y_v thỏa mãn tính chất Markov đối với đồ thị:

$$P(Y_v|X, Y_w, w \neq v) = P(Y_v|X, Y_w, w \sim v)$$

trong đó $w \sim v$ có nghĩa là w và v là các đỉnh kề nhau trong đồ thị G. Hay nói cách khác trạng thái của các đỉnh trong đồ thị chỉ phụ thuộc vào các điểm lân cận.

=> CRFs là một trường hợp đặc biệt của MRF, trong có các nút có thể chia thành 2 tập riêng biệt X, Y. Và xác suất của chuỗi nhãn Y được xác định dựa trên toàn bộ chuỗi quan sát X. Do X là các biến quan sát lên cấu trúc đồ thị của X là tùy ý và Y và các biến y ∈ Y có thể phụ thuộc vào bất kì biến nào trong X.

Trong trường hợp CRFs có X, Y là các chuỗi X = (X1, …, Xn), Y = (Y1,…, Yn) và đồ thị G là cây mà các đỉnh có bậc không quá 2 (chuỗi tuyến tính) thì được gọi là trường ngẫu nhiên có điều kiện tuyến tính (Linear-Chain CRFs).



Hình 2. Ví dụ minh họa Linear-Chain CRFs trong bài toán gán nhãn thực thể có tên

Hình 2 là một ví dụ về Linear-Chain CRFs được sử dụng trong bài toán gán nhãn thực thể có tên (tìm xem từ nào là tên riêng – PER, từ nào là tên địa danh – LOC). Ở đây, các từ trong câu đầu vào cần được gán nhãn sẽ có vai trò là chuỗi X, các nhãn cần được gán cho từng từ trong câu đầu vào sẽ là chuỗi Y. Các nhãn này sẽ nhận một trong các giá trị: PER-Tên riêng, LOC-Địa điểm, O-Không xác định. Theo tích chất Markov thì nhãn của từ hiện tại chỉ phụ thuộc vào nhãn trước, nhãn sau và câu đầu vào.

2. Xây dựng mô hình xác suất P(Y|X)

Với giả định $P(X = x, Y = y)$ là dương nghiêm ngặt, theo định lý Hammersley–Clifford [9], ta có:

$$E(x,y) = -\sum_{c_i \in C} f_i\,(c_i)$$

$$P(X = x, Y = y) = \frac{1}{Z} e^{-E(x,y)}$$

$$Z = \sum_{x \in \Omega_x, y \in \Omega_y} e^{-E(x,y)}$$

Trong đó C là tập tất cả các nhóm đầy đủ của đồ thị G (một **nhóm đầy đủ** trong đồ thị vô hướng là một tập hợp các đỉnh mà giữa tất cả các cặp đỉnh trong tập hợp đó đều tồn tại một cạnh), f_i là hàm năng lượng của cụm c_i chỉ ra khả năng xảy ra các mối quan hệ trong cụm. Z là hằng số chuẩn hóa để tạo phân phối xác suất hợp lệ (<1). $E(x,y)$ là hàm năng lượng được sử dụng để đánh giá mức độ "tốt" của một cặp giá trị (x,y) cụ thể của các biến ngẫu nhiên X, Y. Cặp giá trị (x,y) có $E(x,y)$ thấp hơn được coi là tốt hơn.

Dựa vào công thức trên kết hợp định lý Bayes, ta suy ra phân phối của chuỗi nhãn Y khi biết X có dạng sau:

$$P(Y = y|X = x) = \frac{P(Y = y, X = x)}{P(X = x)} = \frac{\frac{e^{-E(x,y)}}{Z}}{\frac{\sum_{y' \in \Omega_y} e^{-E(x,y')}}{Z}}$$

$$= \frac{e^{-E(x,y)}}{Z(x)}$$

$$= \frac{exp\left(\sum_{c_i \in C} f_i\,(c_i)\right)}{Z(x)}$$

$$Z(x) = \sum_{y' \in \Omega_y} e^{-E(x,y')}$$

Với Linear-Chain CRFs, tập các cụm là 2 đỉnh của các cạnh và các đỉnh lẻ, khi đó, ta có:

$$E(x,y) = -\left(\sum_{(i-1,i)\in E} f\,(y_{i-1}, y_i, x, i) + \sum_{y_i \in y} g\,(y_i, x, i) \right)$$

Để đơn giản, ta thêm 2 nhãn vào đầu và cuối chuỗi nhãn: Y0 = <Start>. Trong Linear-Chain CRFs, hàm năng lượng cho các cạnh là tổng hợp các hàm đặc trưng cạnh f_k và hàm năng lượng cho đỉnh là tổng hợp các hàm đặc trưng của đỉnh g_k.

$$E(x,y) = -\left(\sum_{i=1}^{n} \sum_{k} \lambda_k\, f_k(y_{i-1}, y_i, x, i) + \sum_{i=1}^{n} \sum_{k} \mu_k\, g_k(y_i, x, i) \right)$$

$$p_\theta(Y = y|X = x) = \frac{exp(\sum_{i=1}^{n} \sum_k \lambda_k\, f_k(y_{i-1}, y_i, x, i) + \sum_{i=1}^{n} \sum_k \mu_k\, g_k(y_i, x, i))}{Z_\theta(x)}$$

$$Z_\theta(x) = \sum_{y' \in \Omega_y} exp\left(\sum_{i=1}^{n} \sum_{k} \lambda_k\, f_k(y'_{i-1}, y'_i, x, i) + \sum_{i=1}^{n} \sum_{k} \mu_k\, g_k(y'_i, x, i) \right)$$

Các hàm đặc trưng f_k và g_k được cho trước và cố định, thường là chỉ báo cho 1 đặc trưng ví dụ 1 hàm đặc trưng sẽ trả về giá trị 1 khi X_i viết hoa chữ cái đầu và Y_i có nhãn là "N" ngược lại sẽ trả về 0.

Trọng số λ_k, μ_k của hàm đặc trưng là một hệ số điều chỉnh mức độ ảnh hưởng của hàm đặc trưng đến năng lượng của cấu hình. Trọng số càng cao, hàm đặc trưng càng có ảnh hưởng lớn đến xác suất của chuỗi nhãn.

3. Linear-Chain CRFs dạng ma trận

Giả sử, $\mathcal{Y} = \{C_1, \dots, C_l\}$, $\mathcal{Y}' = \mathcal{Y} \cup \{< Start >\}$. Xác xuất có điều kiện của chuỗi Y có thể được biểu diễn dưới dạng ma trận. Tại mỗi vị trí i trong chuỗi quan sát x, ta định nghĩa một ma trận biến ngẫu nhiên kích thước $|\mathcal{Y}'| \times |\mathcal{Y}'|$, $M_i(x) = \left[M_i(C_j, C_k|x)\right],\ C_j, C_k \in \mathcal{Y}$.

Mỗi phần tử $M_i(C_j, C_k|x)$ đại diện cho một giá trị xác suất chưa chuẩn hóa. $M_i(x)$ là biến ngẫu nhiên mà giá trị phụ thuộc vào chuỗi quan sát X.

$$M_i(C_j, C_k|x) = exp\left(\sum_k \lambda_k f_k(Y_{i-1} = C_j, y_i = C_k, x, i) + \sum_k \mu_k g_k(Y_i = C_j, x, i)\right)$$

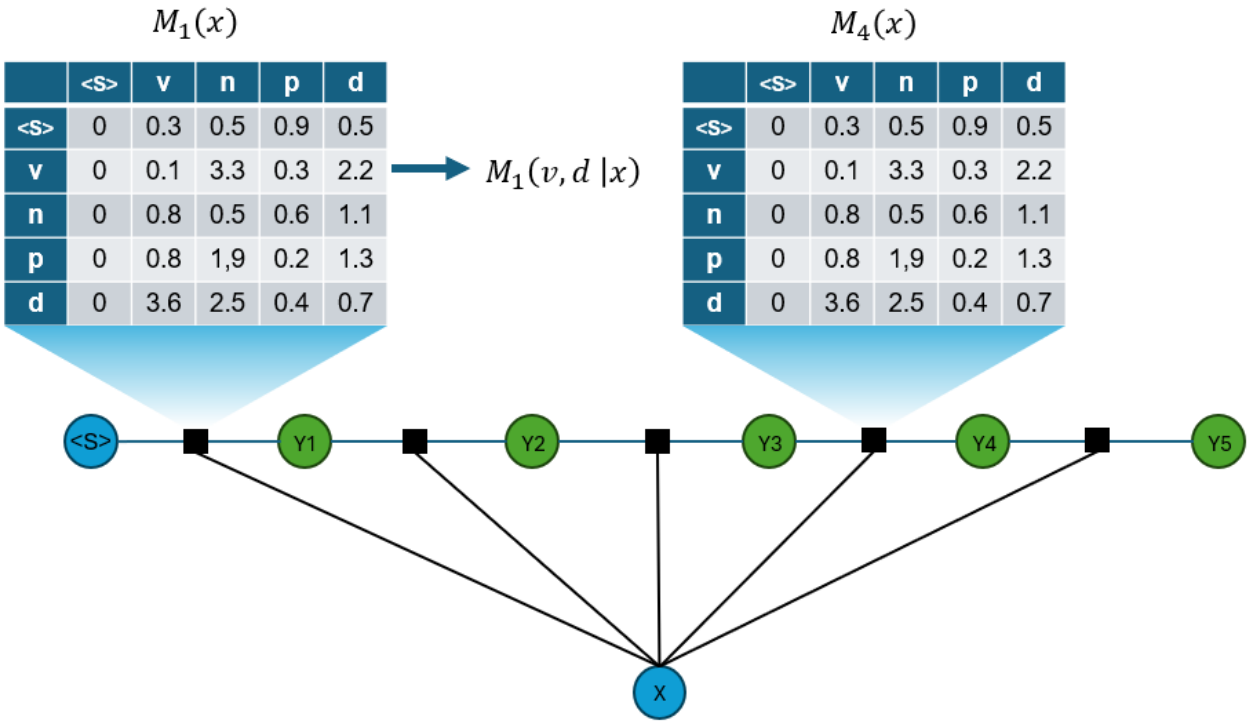
Với cách biểu diễn trên, $Z_\theta(x)$ có thể viết lại dưới dạng sau với $1_{|y'|\times 1}$ là ma trận kích thước $|\mathcal{Y}'|$ hàng và 1 cột có các giá trị bằng 1:

$$Z_\theta(x) = \left(M_1(x) \times M_2(x) \times \dots \times M_{n+1} \times 1_{|y'|\times 1}\right)_{0,0}$$

Công thức xác suất có điều kiện có thể biểu diễn dưới dạng ma trận:

$$p_\theta(Y = y|X = x) = \frac{\prod_{i=1}^{n} M_i(y_{i-1}, y_i|x)}{\left(\left(\prod_{i=1}^{n} M_i(x)\right) \times 1_{|y'|\times 1}\right)_{0,0}}$$

Biểu diễn này hữu ích trong việc huấn luyện và suy luận mô hình CRFs.



Hình 3. Linear-Chain CRFs biểu diễn dưới dạng factor với các factor được coi là ma trận chuyển đổi

Hình 3 là một ví dụ mình họa của linear-Chain CRFs biểu diễn dưới dạng factor cho bài toán POS tiếng Việt (gán nhãn động từ - v, danh từ - n, đại từ - p, trạng từ - d). Ở đây, câu đầu vào có 5 từ và mỗi 1 từ sẽ được gán nhãn từ loại tương ứng. Chuỗi từ loại chính là chuỗi Y. Giữa mỗi cặp nhãn cần gán kề nhau sẽ có một ma trận thể hiện khả năng mà giá trị nhán được gán khi biết nhãn của từ liền kề.