HMM and CRF

jojonki

Mar. 2020

1 はじめに

言語処理において重要な系列ラベリングにおける HMM と CRF に関して勉強したメモ．使用した教材は名著と名高い高村先生による「言語処理のための機械学習入門」である．この教科書で十分に説明がなされているのであるが，せっかく勉強したので自分なりに途中式などのメモと合わせてこの文書を書いている．誤り箇所があれば Twitter 等で@jojonki までメンションを投げてもらえると助かる．

そもそものモチベーションとして，HMM と CRF はどちらも時系列データのラベル予測に利用可能なモデルであるが，どのように違うのか，それをしっかりと理解するためにこの文書を作った．解説の前提として，系列ラベリングに対して，x を入力系列データ，y を予測すべきラベル系列とする．

2 HMM

HMM (Hidden Markov Model) は，言語処理界隈においてはラベル付き（教師あり）データを与えられた際の学習データを扱う（ラベルが観測できているが，慣例として HMM と言語処理界隈では呼ぶ模様）．

HMM は生成モデルとして捉えることができ，入力系列 x 及びラベル列 y の同時確率 $P(x, y)$ を解く問題である．マルコフ性を仮定するため，(x_i, y_i) は (x_{i-1}, y_{i-1}) にのみ依存する．更に詳細に仮定すると，x_i は y_i のみに依存し，y_i は y_{i-1} のみに依存する．そのため x と y の同時確率分布は下記のように展開できる．なお，入力データの左端と右端にダミーシンボル（$\mathbf{B}, \mathbf{E}$）を入れて，文の開始と終了を表現できることを前提にしている．

$$\begin{aligned} P(\mathbf{x}, \mathbf{y}) &= \prod_i P(x_i, y_i | x_{i-1}, y_{i-1}) \\ &= \prod_i P(x_i | y_i) P(y_i | y_{i-1}) \end{aligned} \tag{1}$$

推論時には，与えられた x に対して，書きを解けば良い．掛け算は扱いにくいので log を取って最大化を考える．

$$\hat{y} = \arg\max_{y \in Y} \log P(\mathbf{x}, y) \tag{2}$$

学習時には，この同時確率分布を教師ありデータによって最大化することを考える．また学習データ $D = \{(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), ..., (\mathbf{x}^{(|D|)}, \mathbf{y}^{(|D|)})\}$ は $|D|$ 個のラベル付き系列データである．

$$\begin{aligned} \log P(D) &= \sum_{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \in D} \log P(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \\ &= \sum_{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \in D} \Big(\sum_j \log P(x_j^{(i)} | y_j^{(i)}) + \sum_j \log P(y_j^{(i)} | y_{j-1}^{(i)}) \Big) \\ &\quad n((x, y), D) \text{ を学習データ } D \text{ において，} x \text{ のラベルが } y \text{ の回数，} \\ &\quad n((y', y), D) \text{ を学習データ } D \text{ において，ラベル } y' \text{ のあとにラベル } y \text{ が続く回数とすると，} \\ &= \sum_{y \in \sum_y} \sum_{x \in \sum_x} n((x, y), D) \log P(x|y) + \sum_{y' \in \sum_y} \sum_{y \in \sum_y} n((y', y), D) \log P(y|y') \end{aligned} \tag{3}$$

$n((x,y),D)$ と $n((y',y),D)$ は学習データより単純に頻度をカウントするだけで良い．問題は条件付き確率の箇所であるが，これはラグランジュ法で解ける．制約として，$\sum_x P(x|y)=1$ 及び $\sum_y P(y|y')=1$ が使える．

$$
\begin{aligned}
L(\theta,\alpha,\beta) = &\sum_{y\in\sum_y}\sum_{x\in\sum_x} n((x,y),D)\log P(x|y) + \sum_{y'\in\sum_y}\sum_{y\in\sum_y} n((y',y),D)\log P(y|y') \\
&+ \sum_y \alpha_y(\sum_x P(x|y)-1) + \sum_{y'}\beta_{y'}(\sum_y P(y|y')-1)
\end{aligned}
\tag{4}
$$

$P(x|y)$ 及び $P(y|y')$ でそれぞれ微分して 0 となるようにする

$$
\begin{aligned}
\frac{\partial L(\theta,\alpha,\beta)}{\partial P(x|y)} &= \frac{n((x,y),D)}{P(x|y)} + \alpha_y = 0 \\
\frac{\partial L(\theta,\alpha,\beta)}{\partial P(y|y')} &= \frac{n((y',y),D)}{P(y|y')} + \beta_{y'} = 0
\end{aligned}
\tag{5}
$$

ここで $\sum_x P(x|y)=1$ 及び $\sum_y P(y|y')=1$ を利用して，α_y と $\beta_{y'}$ を消すと，下記が解析的に求められる．要は，学習データに対してそれぞれの頻度を求めていることになる．

$$
\begin{aligned}
P(x|y) &= \frac{n((x,y),D)}{\sum_x n((x,y),D)} \\
P(y|y') &= \frac{n((y,y'),D)}{\sum_y n((y,y'),D)}
\end{aligned}
\tag{6}
$$

推論時はすでに述べたが x と y の同時確率を最大化するものを選べば良い．その探索のために，ビタビアルゴリズムが主に使われるようだ．ビタビアルゴリズムは前向き計算及び後ろ向き計算からなる手法で，最小コスト経路を効率的に探索できる．`https://www.jonki.net/entry/2019/12/01/000807` などを参照してほしい．

3　対数線形モデル (最大エントロピーモデル)

SVM と並んでよく使われる分類器として対数線形モデル（言語処理においては，最大エントロピーモデルと呼ばれることも多いらしい）を CRF の前に説明しておく．というのも対数線形モデルを系列データに適用するのが CRF だからである．対数線形モデルでは $P(y|d)$ を直接する分類モデルである（d は例えば文，y はその文のクラスラベルである）．

対数線形モデルは下記で定式化される．

$$
\begin{aligned}
P(y|d) &= \frac{\exp(\mathbf{w}\cdot\phi(d,y))}{\sum_y \exp(\mathbf{w}\cdot\phi(d,y))} \\
&= \frac{\exp(\mathbf{w}\cdot\phi(d,y))}{Z_{d,\mathbf{w}}}
\end{aligned}
\tag{7}
$$

ϕ は素性ベクトル，$\mathbf{w}$ は各素性に対する重みである．これを最大化する y が推論結果になる（推論する際には，分母は正規化項であるので無視してよいが，確率値として出すなら計算が必要）．

対数線形モデルの学習は，下記のように定義できる．重みが無限大に大きくなることを避けるための正則化

項が入っている．解析的に重み $\mathbf{w}$ を求めることはできないので，最急降下法を利用する．

$$
\begin{aligned}
L(w) &= \sum_{(d^{(i)},y^{(i)})\in D} \log P(y^{(i)}|d^{(i)}) - \frac{C}{2}|\mathbf{w}|^2 \\
&= \sum_{(d^{(i)},y^{(i)})\in D} \Big(\mathbf{w}\cdot\phi(d,y) - \log Z_{d,\mathbf{w}}\Big) - \frac{C}{2}|\mathbf{w}|^2 \\
\nabla_{\mathbf{w}} L(\mathbf{w}) &= \sum_{(d^{(i)},y^{(i)})\in D} \Big(\phi(d^{(i)},y^{(i)}) - \frac{\sum_y \phi(d^{(i)},y)\exp(\mathbf{w}\cdot\phi(d^{(i)},y))}{Z_{d^{(i)},\mathbf{w}}}\Big) - C\mathbf{w} \qquad (8)\\
&= \sum_{(d^{(i)},y^{(i)})\in D} \Big(\phi(d^{(i)},y^{(i)}) - \sum_y P(y|d^{(i)})\phi(d^{(i)},y)\Big) - C\mathbf{w}
\end{aligned}
$$

パラメタ更新は下記のようにすれば良い，

$$\mathbf{w}^{new} = \mathbf{w}^{old} + \epsilon\nabla_{\mathbf{w}} L(\mathbf{w}^{old})$$

4 CRF

CRF (Conditional Random Fields, 条件付き確率場) の説明に入る．上述したとおり，対数線形モデルを系列ラベリングに割り当てたのが CRF である．つまり CRF は分類モデルである．対数線形モデルの式から CRF のモデルの式を比較するために，下記のように記述しよう．

$$\hat{y} = \arg\max_y \frac{\exp(\mathbf{w}\cdot\phi(d,y))}{Z_{d,\mathbf{w}}}$$

最大化は分子だけ見れば良い，

$$= \arg\max_y \mathbf{w}\cdot\phi(d,y)$$

入力データの記号は文書 d を使っていたが，より汎用的な入力系列データを表す $\mathbf{x}$ に変えておく，

$$= \arg\max_y \mathbf{w}\cdot\phi(\mathbf{x},y)$$

対数線形モデルではある 1 つのクラスをラベルとして求めるが，CRF では時系列ラベルなので y は $\mathbf{y}$ と表現する，

$$= \arg\max_{\mathbf{y}} \mathbf{w}\cdot\phi(\mathbf{x},\mathbf{y})$$

CRF においては，ステップ t とそれ以前のステップ（ここでは 1 ステップ前）だけを考える，

$$= \arg\max_{\mathbf{y}} \sum_t \mathbf{w}\cdot\phi(\mathbf{x},y_t,y_{t-1}) \qquad (9)$$

ここでは t と $t-1$ のラベルということで，マルコフ性のような捉え方ができる．HMM のときと同様に現在と直前のみ，などの制約をいれることによって，探索空間を限定して扱いやすくしている．またここまでくれば，ビタビアルゴリズムなどを用いて，過去ステップからの現在ステップのラベルは何か，と考えるので HMM のときとやっていることは似ている．

重みは対数線形モデルのときと同様に最急降下法を使うとすると，式 8 を思い出して，

$$
\begin{aligned}
\nabla_{\mathbf{w}} L(\mathbf{w}) &= \sum_{(\mathbf{x}^{(i)},\mathbf{y}^{(i)})\in D} \Big(\phi(\mathbf{x}^{(i)},\mathbf{y}^{(i)}) - \sum_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}^{(i)})\phi(\mathbf{x}^{(i)},\mathbf{y})\Big) - C\mathbf{w} \\
&\quad \text{ラベルはステップ } t \text{ と } t-1 \text{ だけ見るのでカッコの中の第 2 項を下記のように変形できる，} \qquad (10)\\
&= \sum_{(\mathbf{x}^{(i)},\mathbf{y}^{(i)})\in D} \Big(\phi(\mathbf{x}^{(i)},\mathbf{y}^{(i)}) - \sum_t \sum_{y_t,y_{t-1}} P(y_t,y_{t-1}|\mathbf{x}^{(i)})\phi(\mathbf{x}^{(i)},y_t,y_{t-1})\Big) - C\mathbf{w}
\end{aligned}
$$

この式の中の $P(y_t, y_{t-1}|\mathbf{x})$ がまだ良くわからないので解説すると，

$$P(y_t, y_{t-1}|\mathbf{x}) = \sum_{\mathbf{y}_{0:t-2}} \sum_{\mathbf{y}_{t+1:T+1}} P(\mathbf{y}|\mathbf{x})$$

各ステップの積で $P(\mathbf{y}|\mathbf{x})$ を表せるので，

$$= \sum_{\mathbf{y}_{0:t-2}} \sum_{\mathbf{y}_{t+1:T+1}} \prod_{t'} \frac{\exp(\mathbf{w} \cdot \phi(\mathbf{x}, y_{t'}, y_{t'-1}))}{Z_{\mathbf{x},\mathbf{w}}}$$

分子の記述簡略化（$\mathbf{x}, \mathbf{w}$ が消えているが，ここでは定数扱いして良いので省略），

$$= \sum_{\mathbf{y}_{0:t-2}} \sum_{\mathbf{y}_{t+1:T+1}} \prod_{t'} \frac{\psi_{t'}(y_{t'}, y_{t'-1})}{Z_{\mathbf{x},\mathbf{w}}}$$

$$= \frac{\psi_t(y_t, y_{t-1})}{Z_{\mathbf{x},\mathbf{w}}} \sum_{\mathbf{y}_{0:t-2}} \sum_{\mathbf{y}_{t+1:T+1}} \prod_{t' \neq t} \psi_{t'}(y_{t'}, y_{t'-1}) \tag{11}$$

t を境目に分離すると，

$$= \frac{\psi_t(y_t, y_{t-1})}{Z_{\mathbf{x},\mathbf{w}}} \Big(\sum_{\mathbf{y}_{0:t-2}} \prod_{t'=1}^{t-1} \psi_{t'}(y_{t'}, y_{t'-1}) \Big) \Big(\sum_{\mathbf{y}_{t+1:T+1}} \prod_{t'=t+1}^{T+1} \psi_{t'}(y_{t'}, y_{t'-1}) \Big)$$

記号 α, β を用いて記述を簡略化，

$$= \frac{\psi_t(y_t, y_{t-1})}{Z_{\mathbf{x},\mathbf{w}}} \alpha(y_{t-1}, t-1) \beta(y_t, t)$$

ところで，α, β だが，これは下式が成り立つため，動的計画法的に段階的に解ける．

$$\begin{aligned}
\alpha(y_t, t) &= \sum_{\mathbf{y}_{0:t-1}} \prod_{t'=1}^{t} \psi_{t'}(y_{t'}, y_{t'-1}) \\
&= \sum_{y_{t-1}} \psi_t(y_t, y_{t-1}) \sum_{\mathbf{y}_{0:t-2}} \prod_{t'=1}^{t-1} \psi_{t'}(y_{t'}, y_{t'-1}) \\
&= \sum_{y_{t-1}} \psi_t(y_t, y_{t-1}) \alpha(y_{t-1}, t-1) \\
\beta(y_t, t) &= \sum_{\mathbf{y}_{t+1:T+1}} \prod_{t'=t+1}^{T+1} \psi_{t'}(y_{t'}, y_{t'-1}) \\
&= \sum_{y_{t+1}} \psi_{t+1}(y_{t+1}, t_t) \sum_{\mathbf{y}_{t+2:T+1}} \prod_{t'=t+2}^{T+1} \psi_{t'}(y_{t'}, y_{t'-1}) \\
&= \sum_{y_{t+1}} \psi_{t+1}(y_{t+1}, t_t) \beta(y_{t+1}, t+1)
\end{aligned} \tag{12}$$

また $Z_{\mathbf{x},\mathbf{w}}$ であるが，実はこれは α で表現できる．

$$
\begin{aligned}
Z_{\mathbf{x},\mathbf{w}} &= \sum_{\mathbf{y_{0:T}}} \exp(\mathbf{w}\cdot\phi(\mathbf{x},\mathbf{y})) \\
&= \sum_{\mathbf{y_{0:T}}} \exp(\mathbf{w}\cdot\sum_{t'}\phi(\mathbf{x},y_{t'},y_{t'-1})) \\
&= \sum_{\mathbf{y_{0:T}}} \prod_{t'=1}^{T} \exp(\mathbf{w}\cdot\phi(\mathbf{x},y_{t'},y_{t'-1})) \\
&= \sum_{\mathbf{y}_T} \sum_{\mathbf{y_{0:T-1}}} \prod_{t'=1}^{T} \exp(\mathbf{w}\cdot\phi(\mathbf{x},y_{t'},y_{t'-1})) \\
&= \sum_{\mathbf{y}_T} \alpha(y_T,T) \\
&\quad \text{文末にダミーシンボル } \mathbf{E} \text{ があることを思い出すと，} \\
&= \alpha(\mathbf{E},T+1)
\end{aligned}
\tag{13}
$$

となり，α を計算しておけば，それがそのまま使えることを意味する．

5 まとめ

CRF は数式がかなり複雑になっているが，基本的には分類モデルである対数線形モデルを時系列に伸ばしているだけなので，意外と単純である．また計算もこれまでの結果を α, β では保持しているので，直前のタイムステップを使った高速な計算が可能なことも何となく分かったかと思う．

6 参考情報

- 言語処理のための機械学習入門,
  https://www.coronasha.co.jp/np/isbn/9784339027518/