

数理情報工学特論第一
【機械学習とデータマイニング】
3章：分類③

かしま ひさし
鹿島 久嗣
(数理 6 研)

kashima@mist.i.~

構造データのモデリング、とくに配列構造のラベリング問題について学びます

- 構造データのモデリング
- 配列データのラベル付け問題
- 条件付き確率場（CRF）
- 条件付き確率場における予測
- 条件付き確率場の事後確率最大化学習

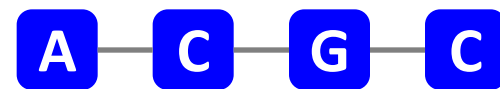
ポイントは「動的計画法」(dynamic programming)

構造データのモデリング

様々な応用分野において、配列や木、グラフなどの構造をもったデータが現れる場面があります

- 様々な分野において、入出力に配列、木、グラフなどの複雑な構造をもったデータを扱う必要がしばしば生じる

— 自然言語処理：文、構文木、文書間リンク



DNA

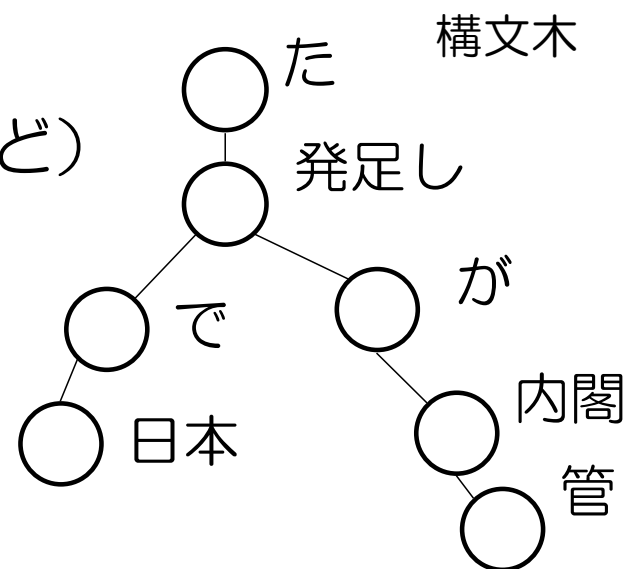
— バイオインフォマティクス：DNA、RNA、タンパク質

- しかし、これまでに紹介してきた方法で、これら構造データをどのように扱ったらよいかは自明ではない

- 構造データを扱う学習問題は、入力 x のもつ構造の種類（配列、木、グラフなど）の別のほか、出力 y における構造の有無によっても大別される

— 出力 y に構造が無い場合

— 出力 y に構造がある場合



構文木

発足し

が

で

日本

内閣

管

出力が構造をもたない場合の例

- 出力 y に構造が無く、入力 x のみが構造をもっている場合
- 出力 y はこれまでに考えてきた回帰や分類などと同じく、1次元（もしくは複数次元）の実数値もしくは離散値をもつ
- 応用としては：
 - HTMLやXMLなどの半構造文書を木もしくはグラフとして表現し文書の構造やレイアウトなどをもとに、文書の性質を予測
 - DNA配列やタンパク質のアミノ酸配列をもとに、それらの機能を予測
 - 化合物の分子構造をグラフとして表現し、その化学的性質を予測

出力が構造をもつ場合の例

- より複雑な状況としては、出力も構造をもつような場合がある
 - 一般的には、構造から構造に変換する問題として捉えられる
 - 配列ラベル付け問題（配列構造の各要素に対して出力値を与える）：
 - 入力 x : 配列構造
 - 出力 y : 同じ長さの配列構造
- | | | | |
|-------|-------|-----|-------|
| x_1 | x_2 | ... | x_T |
| y_1 | y_2 | ... | y_T |
- 多くの問題が配列のラベル付け問題として定式化される
 - 自然言語処理：品詞付けや固有表現抽出など
 - バイオインフォマティクス：遺伝子発見やタンパク質の2次構造予測など
 - 入出力の構造は必ずしも同じ種類である必要はない
 - 自然言語処理：構文解析（配列→木）
 - バイオインフォマティクス：RNA構造予測（配列→木）

配列データのラベリング問題

配列構造のラベル付け問題とは、入力配列と同じサイズの出力配列を予測する問題です

- 配列構造のラベル付け問題：出力が構造をもつ一番簡単な場合

- 入力：長さ T の配列 $x = (x_1, x_2, \dots, x_T)$

- 出力（予測）：同じ長さの配列 $y = (y_1, y_2, \dots, y_T)$

x_1	x_2	...	x_T
y_1	y_2	...	y_T

- これまでの枠組みでいえば、入力配列 x が与えられたもとでの出力配列 y の条件付き確率 $P(y | x)$ を得るのが目的
- 入力配列のそれぞれの要素 x_t ($t=1, \dots, T$) に対応する出力配列の要素 y_t ($t=1, \dots, T$) を予測するため、配列のラベル付けと呼ばれる
- 例えば、品詞付け問題は、文が単語の列として与えられたときに、各単語に対して適切な品詞を割り振る問題
 - 入力 x ：単語列（各 x_t は「私」「は」「走る」などの単語）
 - 出力 y ：品詞列（各 y_t は「名詞」「助詞」「動詞」などの品詞）

配列のラベル付け問題は、出力ラベル間の依存関係を考慮するという意味で（多クラス）分類問題の拡張となっています

- 原理上は、配列のラベル付け問題は、入力配列の要素それぞれについて独立した分類問題として考えることも可能
 - 文の品詞付けにおいて、ある単語のみ（例えば「私」）を見て、その品詞を予測するならば、これは通常の（多クラス）分類問題
- しかし、多くの場合、独立性の仮定は成り立たない
 - 上の例では、全体として整合性のある品詞列を予測する必要あり
 - 「名詞のあとには助詞が続きやすい」等の品詞間の依存関係を考慮し、各品詞でなく品詞「列」としてまとめて予測する必要あり
- 配列のラベル付け問題は、通常の分類問題の拡張となっている
 - 通常の分類問題と同じように訓練データ集合としては N 個の入出力の組 $\{(x^{(i)}, y^{(i)})\}_{i=1}^N$ が与えられる
 - このとき $x^{(i)}$ と $y^{(i)}$ はそれぞれ長さ $T^{(i)}$ の入力配列と出力配列

条件付き確率場 (CRF)

(配列に対する) 条件付き確率場のモデル

- 条件付き確率場 (Conditional random field; **CRF**) は、入力 x と出力 y が共に構造をもつ条件付き確率分布 $P(y|x)$ を表現するモデル
- ここでは、最も簡単なケースとして配列データのラベル付けのためのCRFを考える
 - 入出力はともに長さ T の配列 $x = (x_1, x_2, \dots, x_T)$ と $y = (y_1, y_2, \dots, y_T)$
 - 出力ラベルの取りうる集合 Σ を $\Sigma \equiv \{1, 2, \dots, C\}$ のように定義し、各 y_t は Σ の要素 ($y_t \in \Sigma$) とする
- このとき、配列 x が与えられたときに、これに同じ長さのラベル列 $y_t \in \Sigma^T$ を割り当てる確率を：
$$P(y|x; \omega) \equiv \frac{\exp(\omega^\top \varphi(x, y))}{\sum_{c \in \Sigma^T} \exp(\omega^\top \varphi(x, c))}$$
 - $\varphi(x, y)$ は入出力配列 x と y の両方を考慮した特徴ベクトル

条件付き確率場 (CRF) は、多クラスロジスティック回帰モデルのある種の拡張になっています

- CRFのモデルは、前にふれた多クラスロジスティック回帰モデルの別表記と同じ形をしている
- y に含まれる T 個のラベルを纏めて1つのクラスだと考えると、全部で C^T 個のクラスある（さらに、その数は入力の長さ T に依存）
- そのため、多クラスロジスティック回帰モデルの元々の形式でこれを表現しようとする、クラスによって異なるパラメータベクトル \mathbf{w}_y を用意する必要があるため、モデルの表現として適当でない
- 別形式では、パラメータの数は入出力の両方を考慮した特徴ベクトル $\boldsymbol{\phi}(x, y)$ の次元に等しいので、 $\boldsymbol{\phi}(x, y)$ をうまく設計してやれば、必ずしも指数個のパラメータベクトルを用意する必要は無い



別形式：

$$P(y|x; \omega) \equiv \frac{\exp(\omega^\top \boldsymbol{\phi}(x, y))}{\sum_{c \in \mathcal{Y}} \exp(\omega^\top \boldsymbol{\phi}(x, c))}$$

元々の形式：

$$P(y|x; \{\mathbf{w}^{(c)}\}_{c \in \mathcal{Y}}) \equiv \frac{\exp(\mathbf{w}^{(y)\top} \boldsymbol{\phi}(x))}{\sum_{c \in \mathcal{Y}} \exp(\mathbf{w}^{(c)\top} \boldsymbol{\phi}(x))}$$

条件付き確率場 (CRF) の特徴ベクトルの定義： 2種類の特徴を用います

- x と y の両方にまたがるCRFの特徴ベクトル $\phi(x, y)$ はどのように設計すればよいだろうか？
- 配列ラベル付けのためのCRFにおいてよく用いられるのは：
 - 配列中での位置 t における入力 x_t と出力 y_t の組み合わせによる特徴  と
 - ひとつ前の位置におけるラベル y_{t-1} と現在位置におけるラベル y_t の組み合わせによる特徴  である

x_1	x_2	...	x_{t-1}	x_t	...	x_T
y_1	y_2	...	y_{t-1}	y_t	...	y_T

位置 t における入力 x_t と出力 y_t の組み合わせによる特徴の定義は多クラスロジスティック回帰の別形式のものと同じです

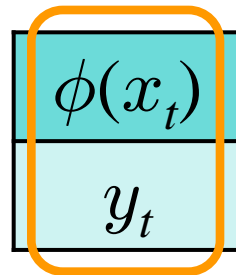
- 配列中での位置 t における入力 x_t と出力 y_t の組み合わせによる特徴
- ロジスティック回帰の別表現でも用いた入力と出力にまたがる特徴ベクトルの定義と同様に、各位置 t に対して以下を定義：

$$\varphi_t^{\mathcal{X}\Sigma}(x_t, y_t) \equiv (\delta(y_t = 1)\phi(x_t)^\top, \delta(y_t = 2)\phi(x_t)^\top, \dots, \delta(y_t = C)\phi(x_t)^\top)^\top$$

— $\delta(\cdot)$ ：括弧の中身が成立するなら1を、しないならば0をとる

- 各 x_t の特徴ベクトル $\phi(x_t)$ を D 次元とすると、これは DC 次元のベクトル

$$\varphi_t^{\mathcal{X}\Sigma}(x_t, y_t)$$



ひとつ前の位置のラベル y_{t-1} と現在位置のラベル y_t の組み合わせ特徴は、連続するラベルの組み合わせを用います

- ひとつ前の位置におけるラベル y_{t-1} と現在位置におけるラベル y_t の組み合わせによる特徴は、隣り合う2つの位置のラベルを各位置 t に対して以下を定義：

$$\varphi_t^{\Sigma\Sigma}(y_{t-1}, y_t) \equiv$$

$$(\delta(y_{t-1} = 1)\delta(y_t = 1), \delta(y_{t-1} = 1)\delta(y_t = 2), \dots, \delta(y_{t-1} = 1)\delta(y_t = C),$$

$$\delta(y_{t-1} = 2)\delta(y_t = 1), \delta(y_{t-1} = 2)\delta(y_t = 2), \dots, \delta(y_{t-1} = 2)\delta(y_t = C),$$

\vdots

$$\delta(y_{t-1} = C)\delta(y_t = 1), \delta(y_{t-1} = C)\delta(y_t = 2), \dots, \delta(y_{t-1} = C)\delta(y_t = C))^{\top}$$

— これは C^2 次元のベクトル

$$\varphi_t^{\Sigma\Sigma}(y_{t-1}, y_t) \begin{array}{|c|c|} \hline x_{t-1} & x_t \\ \hline y_{t-1} & y_t \\ \hline \end{array}$$

2種類の特徴ベクトルを組みあわせ、配列全体で足し合わせたものが配列全体の特徴ベクトルになります

- 以上の2種類の特徴ベクトル $\varphi_t^{\mathcal{X}\Sigma}(x_t, y_t)$ と $\varphi_t^{\Sigma\Sigma}(y_{t-1}, y_t)$ を並べた：

$$\varphi_t(x_t, y_{t-1}, y_t) \equiv (\varphi_t^{\mathcal{X}\Sigma}(x_t, y_t)^\top, \varphi_t^{\Sigma\Sigma}(y_{t-1}, y_t)^\top)^\top$$

を、位置 t における特徴ベクトルとする

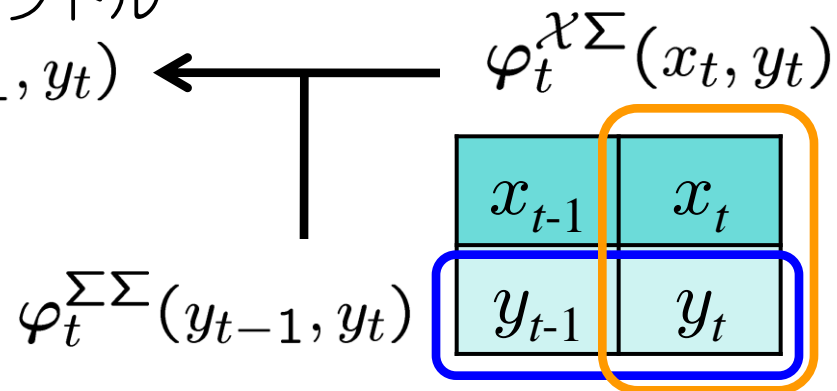
- そして、これを全ての位置について足し合わせたもの：

$$\varphi(x, y) \equiv \sum_{t=1}^T \varphi_t(x_t, y_{t-1}, y_t)$$

が、配列全体に対する特徴ベクトルの定義となる

位置 t における特徴ベクトル

$$\varphi_t(x_t, y_{t-1}, y_t)$$



CRFでは、配列上で隣り合う2つのラベルの組み合わせ特徴によって、ラベル間の依存関係を効率的に捉えています

- CRFの特徴ベクトルは全体として $CD + C^2$ の長さを持つ
- この特徴ベクトルの構成において特に重要なのは、隣り合う2つの位置のラベルの関係を考慮した特徴 $\varphi_t^{\Sigma\Sigma}(y_{t-1}, y_t)$ である
 - 各位置の入出力の特徴ベクトル $\varphi_t^{x\Sigma}(x_t, y_t)$ だけでは、位置間の関係を取り入れていないため、各位置で独立に分類問題を解くのと変わりがなくなってしまう
 - 一方で、長さ T の出力ラベル列に含まれる T 個のラベルの組み合わせを素朴に捉えてしまうと前述のように C^T 個のパラメータベクトル (DC^T 次元) を考えることになってしまう
- CRFの特徴ベクトルでは、ラベル間の関係を、配列上で隣り合う2つのラベルの関係にのみ限定して考えることによって、ラベル間の依存関係を効率的に捉えている

補足：端っこの処理

- 隣り合う2つのラベルについて定義された特徴の定義

$$\varphi_t^{\Sigma\Sigma}(y_{t-1}, y_t) \equiv$$

$$(\delta(y_{t-1} = 1)\delta(y_t = 1), \dots, \delta(y_{t-1} = 1)\delta(y_t = C),$$

...

$$\delta(y_{t-1} = C)\delta(y_t = 1), \dots, \delta(y_{t-1} = C)\delta(y_t = C))^{\top}$$

において、本来 $t=1$ のときには y_0 が定義されている必要がある

- 便宜的に、 y_0 は Σ のいずれのラベルももたない
つまり、任意の $c \in \Sigma$ について $\delta(y_0=c)=0$ であるものとする
- ただし、自然言語処理など、文中での位置が有用な情報を持っている場合には、 y_0 や y_{T+1} などに、文頭や文の末尾などを示す特殊なラベル（※）を特別に割り当てるという方法をとることもある

	x_1	x_2	...	x_T	
※	y_1	y_2	...	y_T	※

条件付き確率場における予測

条件付き確率場（CRF）の予測は、素朴に行おうとすると指数時間かかってしまうので、動的計画法を用います

- 予測：入力列 x が与えられたときに、最も高い条件付き確率 $P(y|x)$ を与える $y = \hat{y}$ を見つけること：

$$\hat{y} \equiv \operatorname{argmax}_{y \in \Sigma^T} \frac{\exp(\omega^\top \varphi(x, y))}{\sum_{c \in \Sigma^T} \exp(\omega^\top \varphi(x, c))}$$

- 可能な出力ラベル列 y の候補が C^T 個あるため、これを素朴に実行することは計算時間の面で現実的ではない。
- しかし、特徴の定義が隣り合う2つの変数に対してのみ定義されていることを利用すると、動的計画法によってこれを効率的に（具体的には $O(TC)$ の計算量で）行うことができる

予測の式は単純化できます

- 予測の式：
$$\hat{y} \equiv \operatorname{argmax}_{y \in \Sigma^T} \frac{\exp(\omega^\top \varphi(x, y))}{\sum_{c \in \Sigma^T} \exp(\omega^\top \varphi(x, c))}$$
- 分母は y に依存しない（すべての y について和をとっている）ため y についての最大化を行うにあたって分母は無視しても良い
- また、分子の \exp は単調増加関数であるため、 $\exp(\cdot)$ の中身を最大化する y がこれを最大化することがわかる。
- 従って、最大化問題は 以下のように書いて差支えない

$$\hat{y} = \operatorname{argmax}_{y \in \Sigma^T} \omega^\top \varphi(x, y)$$

- さらに条件付き確率場の特徴ベクトルの定義により：

$$\hat{y} = \operatorname{argmax}_{y \in \Sigma^T} \sum_{t=1}^T \omega^\top \varphi_t(x_t, y_{t-1}, y_t)$$

動的計画法を用いるために、新たな量 s_τ を導入します

- 最大化問題：
$$\hat{y} = \operatorname{argmax}_{y \in \Sigma^T} \sum_{t=1}^T \omega^\top \varphi_t(x_t, y_{t-1}, y_t)$$

を再帰によって効率的に解くために、以下の量 s_τ を定義する

$$s_\tau(y_\tau) \equiv \max_{(y_1, y_2, \dots, y_{\tau-1}) \in \Sigma^{\tau-1}} \sum_{t=1}^{\tau} \omega^\top \varphi_t(x_t, y_{t-1}, y_t)$$

- すると、次が成立する：

$$\max_{y \in \Sigma^T} \sum_{t=1}^T \omega^\top \varphi_t(x_t, y_{t-1}, y_t) = \max_{y_T \in \Sigma} s_T(y_T)$$

ので、 $s_T(y_T)$ がすべての $y_T \in \Sigma$ について得られれば最大スコアが得られる

s_τ について再帰式が成立します

- ここで先ほど定義した

$$s_\tau(y_\tau) \equiv \max_{(y_1, y_2, \dots, y_{\tau-1}) \in \Sigma^{\tau-1}} \sum_{t=1}^{\tau} \omega^\top \varphi_t(x_t, y_{t-1}, y_t)$$

を再帰的に、すなわち $s_\tau(y_{\tau+1})$ を $s_\tau(y_\tau)$ によって表現してみると：

$$\begin{aligned} s_{\tau+1}(y_{\tau+1}) &= \max_{(y_1, y_2, \dots, y_\tau) \in \Sigma^\tau} \sum_{t=1}^{\tau+1} \omega^\top \varphi_t(x_t, y_{t-1}, y_t) \\ &= \max_{y_\tau \in \Sigma} \max_{(y_1, y_2, \dots, y_{\tau-1}) \in \Sigma^{\tau-1}} \sum_{t=1}^{\tau} \omega^\top \varphi_t(x_t, y_{t-1}, y_t) + \omega^\top \varphi_{\tau+1}(x_{\tau+1}, y_\tau, y_{\tau+1}) \\ &= \max_{y_\tau \in \Sigma} s(y_\tau) + \omega^\top \varphi_{\tau+1}(x_{\tau+1}, y_\tau, y_{\tau+1}) \end{aligned}$$

- これを用いて、 $s_\tau(y_\tau)$ を $\tau=1, 2, \dots, T$ について再帰的に計算することですべての $y_T \in \Sigma$ について $s_T(y_T)$ が求まる
- この計算量は各 τ について $O(C)$ であるため、配列全体としては $O(CT)$

最適なラベル列はバックトラックによって求められます

- 再帰式：

$$s_{\tau+1}(y_{\tau+1}) = \max_{y_{\tau} \in \Sigma} s(y_{\tau}) + \omega^{\top} \varphi_{\tau+1}(x_{\tau+1}, y_{\tau}, y_{\tau+1})$$

- 再帰の各ステップ ($\tau=1,2,\dots,T$) において、 $s_{\tau}(y_{\tau})$ を再帰計算するときに、maxを与える $y_{\tau-1}$ を覚えておく
- $s_T(y_T)$ がすべての $y_T \in \Sigma$ について求め、最適な y_T がわかったら、覚えていた最適な y_{T-1} を取り出し...と逆向きに辿る（バックトラック）することで、最適なラベル列が求まる

条件付き確率場の事後確率最大化

事後確率の最大化（もしくは最尤推定）によってCRFを学習します

- N 個の入出力配列の組 $\{(x^{(i)}, y^{(i)})\}_{i=1}^N$ が訓練データ集合として与えられたとする。
 - i 番目の訓練データの長さを $T^{(i)}$ とする。
- 訓練データ集合に対する事後確率の対数を取ったものの和は：

$$\begin{aligned} L(\omega) &\equiv \sum_{i=1}^N \log P(y^{(i)} | x^{(i)}; \omega) - \lambda \|\omega\|_2^2 \\ &= \sum_{i=1}^N \log \frac{\exp(\omega^\top \varphi(x^{(i)}, y^{(i)}))}{\sum_{y \in \Sigma^{T^{(i)}}} \exp(\omega^\top \varphi(x^{(i)}, y))} - \lambda \|\omega\|_2^2 \\ &= \sum_{i=1}^N \omega^\top \varphi(x^{(i)}, y^{(i)}) - \sum_{i=1}^N \log \sum_{y \in \Sigma^{T^{(i)}}} \exp(\omega^\top \varphi(x^{(i)}, y)) - \lambda \|\omega\|_2^2 \end{aligned}$$

最急勾配法によって目的関数を最大化します

- これを最大化するために、最急勾配法を用いることにする
- 更新式（現在のパラメータ $\omega^{(t)}$ から新たなパラメータ $\omega^{(t+1)}$ へ）は：

$$\omega^{(t+1)} \leftarrow \omega^{(t)} + \eta^{(t)} \nabla(\omega^{(t)})$$

- $\eta^{(t)} > 0$ は更新の幅を決定する学習率と呼ばれるパラメータ
- $\nabla(\omega^{(t)})$ は現在のパラメータ $\omega = \omega^{(t)}$ における目的関数 $L(\omega)$ の勾配：

$$\nabla(\omega^{(t)}) \equiv \left. \frac{\partial L(\omega)}{\partial \omega} \right|_{\omega = \omega^{(t)}}$$

勾配を求める際に、計算の難しい箇所が2箇所あります

- パラメータの更新を行うためには $\nabla(\omega^{(t)})$ を計算する必要がある
- そのために目的関数 $L(\omega)$ を ω について偏微分したものを計算する

$$\frac{\partial L(\omega)}{\partial \omega} = \sum_{i=1}^N \varphi(x^{(i)}, y^{(i)}) - \sum_{i=1}^N \frac{\sum_{y \in \Sigma^{T(i)}} \exp(\omega^\top \varphi(x^{(i)}, y)) \varphi(x^{(i)}, y)}{\sum_{y \in \Sigma^{T(i)}} \exp(\omega^\top \varphi(x^{(i)}, y))} - 2\lambda \omega$$

- ここで問題となってくるのが2つの部分の計算：

$$\sum_{y \in \Sigma^{T(i)}} \exp(\omega^\top \varphi(x^{(i)}, y)) \quad \sum_{y \in \Sigma^{T(i)}} \exp(\omega^\top \varphi(x^{(i)}, y)) \varphi(x^{(i)}, y)$$

- 可能な全ての $y \in \Sigma^{T(i)}$ についての和を含むため、これを素朴に計算するのは、計算量の面で現実的ではない
- 予測のときと同じく動的計画法を利用できる

まずは簡単な方 $\sum_{y \in \Sigma^{T(i)}} \exp(\omega^\top \varphi(x^{(i)}, y))$ から計算します

- 条件付き確率場の特徴ベクトルの定義から:

$$\sum_{y \in \Sigma^{T(i)}} \exp(\omega^\top \varphi(x^{(i)}, y)) = \sum_{y \in \Sigma^{T(i)}} \exp \left(\sum_{t=1}^{T(i)} \omega^\top \varphi_t(x_t, y_{t-1}, y_t) \right)$$

- ここで以下の量を定義する:

$$u_\tau(y_\tau) \equiv \sum_{(y_1, y_2, \dots, y_{\tau-1}) \in \Sigma^{\tau-1}} \exp \left(\sum_{t=1}^{\tau} \omega^\top \varphi_t(x_t, y_{t-1}, y_t) \right)$$

- これを使うと我々の求めたいものは:

$$\sum_{y \in \Sigma^{T(i)}} \exp(\omega^\top \varphi(x^{(i)}, y)) = \sum_{y_{T(i)} \in \Sigma} u_{T(i)}(y_{T(i)})$$

- つまり、 $u_{T(i)}(y_{T(i)})$ を全ての $y_{T(i)} \in \Sigma$ について計算できればよい

動的計画法によって求めます

- $u_T^{(i)}(y_T^{(i)})$ を得るために、 $u_{\tau+1}(y_{\tau+1})$ を $u_\tau(y_\tau)$ を用いて再帰表現する：

$$\begin{aligned} u_{\tau+1}(y_{\tau+1}) &= \sum_{(y_1, y_2, \dots, y_\tau) \in \Sigma^\tau} \exp \left(\sum_{t=1}^{\tau+1} \omega^\top \varphi_t(x_t, y_{t-1}, y_t) \right) \\ &= \sum_{(y_1, y_2, \dots, y_\tau) \in \Sigma^\tau} \exp \left(\sum_{t=1}^{\tau} \omega^\top \varphi_t(x_t, y_{t-1}, y_t) + \omega^\top \varphi_\tau(x_{\tau+1}, y_\tau, y_{\tau+1}) \right) \\ &= \sum_{y_\tau \in \Sigma} \sum_{(y_1, y_2, \dots, y_{\tau-1}) \in \Sigma^{\tau-1}} \exp \left(\sum_{t=1}^{\tau} \omega^\top \varphi_t(x_t, y_{t-1}, y_t) \right) \exp \left(\omega^\top \varphi_\tau(x_{\tau+1}, y_\tau, y_{\tau+1}) \right) \\ &= \sum_{y_\tau \in \Sigma} u_\tau(y_\tau) \exp \left(\omega^\top \varphi_\tau(x_{\tau+1}, y_\tau, y_{\tau+1}) \right) \end{aligned}$$

- 予測のときと同様に、これを用いて $\tau = 1, 2, \dots, T^{(i)}$ について再帰的に計算することで、 $u_T^{(i)}(y_T^{(i)})$ が全ての $y_T^{(i)} \in \Sigma$ について求まる
- 計算量は各 τ について $O(C)$ であるため、 i 番目の訓練データについては $O(CT^{(i)})$ となる

もう一方 $\sum_{y \in \Sigma^{T(i)}} \exp(\omega^\top \varphi(x^{(i)}, y)) \varphi(x^{(i)}, y)$ の計算は面倒です

- 条件付き確率場の特徴ベクトルの定義から、以下のように書ける：

$$\sum_{y \in \Sigma^{T(i)}} \exp(\omega^\top \varphi(x^{(i)}, y)) \varphi(x^{(i)}, y) = \sum_{y \in \Sigma^{T(i)}} \exp\left(\sum_{t=1}^{T(i)} \omega^\top \varphi_t(x_t, y_{t-1}, y_t)\right) \sum_{t'=1}^{T(i)} \varphi_{t'}(x_{t'}, y_{t'-1}, y_{t'})$$

- さらに：

$$\begin{aligned} &= \sum_{t'=1}^{T(i)} \sum_{y \in \Sigma^{T(i)}} \exp\left(\sum_{t=1}^{T(i)} \omega^\top \varphi_t(x_t, y_{t-1}, y_t)\right) \varphi_{t'}(x_{t'}, y_{t'-1}, y_{t'}) \\ &= \sum_{t'=1}^{T(i)} \sum_{(y_{t'-1}, y_{t'}) \in \Sigma^2} \sum_{(y_1, y_2, \dots, y_{t'-2}, y_{t'+1}, \dots, y_{T(i)}) \in \Sigma^{T(i)-2}} \exp\left(\sum_{t=1}^{T(i)} \omega^\top \varphi_t(x_t, y_{t-1}, y_t)\right) \varphi_{t'}(x_{t'}, y_{t'-1}, y_{t'}) \\ &= \sum_{t'=1}^{T(i)} \sum_{(y_{t'-1}, y_{t'}) \in \Sigma^2} \varphi_{t'}(x_{t'}, y_{t'-1}, y_{t'}) \sum_{(y_1, y_2, \dots, y_{t'-2}, y_{t'+1}, \dots, y_{T(i)}) \in \Sigma^{T(i)-2}} \exp\left(\sum_{t=1}^{T(i)} \omega^\top \varphi_t(x_t, y_{t-1}, y_t)\right) \\ &= \sum_{t'=1}^{T(i)} \sum_{(y_{t'-1}, y_{t'}) \in \Sigma^2} \varphi_{t'}(x_{t'}, y_{t'-1}, y_{t'}) \exp(\omega^\top \varphi_{t'}(x_{t'}, y_{t'-1}, y_{t'})) \\ &\quad \sum_{(y_1, y_2, \dots, y_{t'-2}) \in \Sigma^{t'-2}} \exp\left(\sum_{t=1}^{t'-1} \omega^\top \varphi_t(x_t, y_{t-1}, y_t)\right) \sum_{(y_{t'+1}, \dots, y_{T(i)}) \in \Sigma^{T(i)-t'}} \exp\left(\sum_{t=t'+1}^{T(i)} \omega^\top \varphi_t(x_t, y_{t-1}, y_t)\right) \end{aligned}$$

ここで先ほど再帰計算によって求めた $u_{t'-1}(y_{t'-1})$ を再利用できます

- 先ほど再帰的に計算した $u_{t'-1}(y_{t'-1})$ が現れる

$$\begin{aligned}
 & \sum_{t'=1}^{T(i)} \sum_{(y_{t'-1}, y_{t'}) \in \Sigma^2} \varphi_{t'}(x_{t'}, y_{t'-1}, y_{t'}) \exp \left(\omega^\top \varphi_{t'}(x_{t'}, y_{t'-1}, y_{t'}) \right) \\
 &= \sum_{t'=1}^{T(i)} \sum_{(y_{t'-1}, y_{t'}) \in \Sigma^2} \varphi_{t'}(x_{t'}, y_{t'-1}, y_{t'}) \exp \left(\omega^\top \varphi_{t'}(x_{t'}, y_{t'-1}, y_{t'}) \right) \\
 & \quad \underbrace{\sum_{(y_1, y_2, \dots, y_{t'-2}) \in \Sigma^{t'-2}} \exp \left(\sum_{t=1}^{t'-1} \omega^\top \varphi_t(x_t, y_{t-1}, y_t) \right)}_{u_{t'-1}(y_{t'-1})} \sum_{(y_{t'+1}, \dots, y_{T(i)}) \in \Sigma^{T(i)-t'}} \exp \left(\sum_{t=t'+1}^{T(i)} \omega^\top \varphi_t(x_t, y_{t-1}, y_t) \right) \\
 & \quad \underbrace{\sum_{(y_{t'+1}, \dots, y_{T(i)}) \in \Sigma^{T(i)-t'}} \exp \left(\sum_{t=t'+1}^{T(i)} \omega^\top \varphi_t(x_t, y_{t-1}, y_t) \right)}_{v_{t'}(y_{t'})}
 \end{aligned}$$

これは次頁で求める $v_{t'}(y_{t'})$

動的計画法を用いるために新たな量 $v_\tau(y_\tau)$ を導入します

- 残るは、一番最後の：
$$\sum_{(y_{t'+1}, \dots, y_{T(i)}) \in \Sigma^{T(i)-t'}} \exp \left(\sum_{t=t'+1}^{T(i)} \omega^\top \varphi_t(x_t, y_{t-1}, y_t) \right)$$
- 新たな量 $v_\tau(y_\tau)$ を定義する：
$$v_\tau(y_\tau) \equiv \sum_{(y_{\tau+1}, \dots, y_{T(i)}) \in \Sigma^{T(i)-\tau}} \exp \left(\sum_{t=\tau+1}^{T(i)} \omega^\top \varphi_t(x_t, y_{t-1}, y_t) \right)$$
 - $v_{t'}(y_{t'})$ の値が全ての $y_{t'} \in \Sigma$ に対して分かればよい
- これがわかれば $\sum_{y \in \Sigma^{T(i)}} \exp(\omega^\top \varphi(x^{(i)}, y)) \varphi(x^{(i)}, y)$ の計算は $O(T^{(i)} C^2)$ ができる

v_τ もまた動的計画法によって効率的に計算できます

- この量にも以下のような再帰的關係が成り立つ：

$$v_\tau(y_\tau) = \sum_{y_{\tau+1} \in \Sigma} v_{\tau+1}(y_{\tau+1}) \exp \left(\omega^\top \varphi_\tau(x_\tau, y_{\tau-1}, y_\tau) \right)$$

- この關係を用いて今度は $\tau = T^{(i)}-1, T^{(i)}-2, \dots, 1$ のように逆向きに再歸式を適用していくことで、 $v_{t'}(y_{t'})$ が、全ての $y_{t'} \in \Sigma$ について求まる