人工知能特論II　第8回

二宮　崇

今日の講義の予定

- HMMの教師付学習
- HMMの教師無し学習
  - EMアルゴリズムの導入
- EMアルゴリズム

- 教科書
  - 北研二(著) 辻井潤一(編) 言語と計算4 確率的言語モデル 東大出版会
  - C. D. Manning & Hinrich Schütze “FOUNDATIONS OF STATISTICAL NATURAL LANGUAGE PROCESSING” MIT Press, 1999
  - Christopher M. Bishop “PATTERN RECOGNITION AND MACHINE LEARNING” Springer, 2006

隠れマルコフモデル
Hidden Markov Model (HMM)

- Q: 状態の有限集合
- Σ: 出力記号の有限集合
- π_q: 文頭が状態qになる確率
  - $\Sigma_{r \in Q} \pi_r = 1$
- $a_{q,r}$: 状態qから状態rへの遷移確率
  - $\Sigma_{r \in Q} a_{q,r} = 1$
- $b_{q,o}$: 状態qにおける記号oの出力確率
  - $\Sigma_{o \in \Sigma} b_{q,o} = 1$

状態記号列の確率と生成確率

- 状態と記号の列が与えられた時

  状態記号列: $q_1 o_1 q_2 o_2 \cdots q_T o_T$

$$p(q_1 o_1 q_2 o_2 \cdots q_T o_T) = \pi_{q_1} \prod_{t=2}^{T} a_{q_{t-1},q_t} \prod_{t=1}^{T} b_{q_t,o_t}$$

- 記号列のみが与えられた時

  記号列: $o_1 o_2 \cdots o_T$

$$p(o_1 o_2 \cdots o_T) = \sum_{q_1 \in Q, q_2 \in Q, \cdots q_T \in Q} p(q_1 o_1 q_2 o_2 \cdots q_T o_T)$$

（生成確率）

- 解析 (入力: $o_1 o_2 \ldots o_T$)

$$\tilde{q}_1 \tilde{q}_2 \cdots \tilde{q}_T = \arg\max_{q_1 \in Q, q_2 \in Q, \cdots q_T \in Q} p(q_1 o_1 q_2 o_2 \cdots q_T o_T)$$

（この問題はビタビアルゴリズムで効率的に解ける）

ラグランジュの未定乗数法

- ラグランジュの未定乗数法

$$\arg\max_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) \text{ただし} g_1(\boldsymbol{\theta}) = 0, \ldots, g_m(\boldsymbol{\theta}) = 0$$

$$\Rightarrow$$

$$L(\boldsymbol{\theta}) = f(\boldsymbol{\theta}) - \lambda_1 g_1(\boldsymbol{\theta}) - \ldots - \lambda_m g_m(\boldsymbol{\theta})$$

$$\frac{\partial L}{\partial \theta_1} = 0, \frac{\partial L}{\partial \theta_2} = 0, \ldots, \frac{\partial L}{\partial \theta_n} = 0$$

- $L(\boldsymbol{\theta})$ はラグランジュ関数と呼ばれる

学習 (パラメータ推定): HMMの教師付学習

HMMの教師付学習
Supervised Learning of HMMs

- パラメータ推定
  - 訓練データ (入力)

訓練データ: $x_1 x_2 \cdots x_n, y_1 y_2 \cdots y_n$

x_i : 文を表す記号列(単語列)。 $x_i = o_{i1} o_{i2} o_{i3} \cdots o_{iT_i}$とする。

T_i : x_iの記号列長

y_i : x_iに対応する正解状態列(正解品詞列)。 $y_i = q_{i1} q_{i2} q_{i3} \cdots q_{iT_i}$とする。

  - パラメータ (出力)

π_q … $|Q|$個の変数

$a_{q,r}$ … $|Q| \times |Q|$個の変数

$b_{q,o}$ … $|Q| \times |\Sigma|$個の変数

HMMの教師付学習
Supervised Learning of HMMs

- パラメータ推定

$$
\begin{aligned}
\pi, a, b &= \arg\max_{\pi,a,b} \prod_{i=1}^{n} p(x_i, y_i) \\
&= \arg\max_{\pi,a,b} \prod_{i=1}^{n} p(q_{i1} o_{i1} q_{i2} o_{i2} \cdots q_{iT_i} o_{iT_i}) \\
&= \arg\max_{\pi,a,b} \prod_{i=1}^{n} \pi_{q_{i1}} \prod_{t=2}^{T_i} a_{q_{i(t-1)}, q_{it}} \prod_{t=1}^{T_i} b_{q_{it}, o_{it}}
\end{aligned}
$$

HMMの教師付学習
Supervised Learning of HMMs

- パラメータ推定

$$l(\pi, a, b) = \prod_{i=1}^{n} \pi_{q_{i1}} \prod_{t=2}^{T_i} a_{q_{i(t-1)}, q_{it}} \prod_{t=1}^{T_i} b_{q_{it}, o_{it}}$$

$$\log l(\pi, a, b) = \sum_{i=1}^{n} \left(\log \pi_{q_{i1}} + \sum_{t=2}^{T_i} \log a_{q_{i(t-1)}, q_{it}} + \sum_{t=1}^{T_i} \log b_{q_{it}, o_{it}} \right)$$

制約付き最適化問題

$$\arg\max_{\pi, a, b} \log l(\pi, a, b)$$

$$\text{s.t.} \quad \sum_{q \in Q} \pi_q = 1$$

$$\sum_{r \in Q} a_{q,r} = 1 \quad (\text{for all } q)$$

$$\sum_{o \in \Sigma} b_{q,o} = 1 \quad (\text{for all } q)$$

HMMの教師付学習
Supervised Learning of HMMs

- ラグランジュの未定乗数法

ラグランジュ関数

$$L(\pi,a,b)$$

$$= \log l(\pi,a,b) - \rho\left(1-\sum_{q\in Q}\pi_q\right) - \sum_{q\in Q}\alpha_q\left(1-\sum_{r\in Q}a_{q,r}\right) - \sum_{q\in Q}\beta_q\left(1-\sum_{o\in\Sigma}b_{q,o}\right)$$

$$= \sum_{i=1}^{n}\left(\log\pi_{q_{i1}} + \sum_{t=2}^{T_i}\log a_{q_{i(t-1)},q_{it}} + \sum_{t=1}^{T_i}\log b_{q_{it},o_{it}}\right) - \rho\left(1-\sum_{q\in Q}\pi_q\right) - \sum_{q\in Q}\alpha_q\left(1-\sum_{r\in Q}a_{q,r}\right) - \sum_{q\in Q}\beta_q\left(1-\sum_{o\in\Sigma}b_{q,o}\right)$$

ラグランジュ乗数

ρ … 1個の変数

α_q … $|Q|$個の変数

β_q … $|Q|$個の変数

HMMの教師付学習
Supervised Learning of HMMs

- π_qを求める

$$\frac{\partial L}{\partial \pi_q} = \frac{C_1(q)}{\pi_q} + \rho = 0$$

ただし、$C_1(q) = \sum_{i=1}^{n} [q_{i1} = q]$

$$\pi_q = \frac{C_1(q)}{-\rho} \quad \cdots(1)$$

ところで$\sum_{q \in Q} \pi_q = 1$であるから

$$\sum_{q \in Q} \pi_q = \frac{\sum_{q \in Q} C_1(q)}{-\rho} = \frac{n}{-\rho} = 1$$

よって、$\rho = -n$。これを式(1)に代入して、

$$\pi_q = \frac{C_1(q)}{n}$$

アイバーソンの記法
(Iverson bracket)

$$[P] = \begin{cases} 1 & \text{if } P \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

$C_1(q)$は訓練データ中で、文の先頭がqになっている回数

HMMの教師付学習
Supervised Learning of HMMs

- $a_{q,r}$を求める

$$\frac{\partial L}{\partial a_{q,r}} = \frac{C(q,r)}{a_{q,r}} + \alpha_q = 0$$

ただし、$C(q,r) = \sum_{i=1}^{n} \sum_{t=2}^{T_i} [q_{i(t-1)} = q][q_{it} = r]$

$C(q,r)$は訓練データ中で、状態qから状態rに遷移した回数

$$a_{q,r} = \frac{C(q,r)}{-\alpha_q} \quad \cdots(1)$$

ところで$\sum_{r \in Q} a_{q,r} = 1$であるから

$$\sum_{r' \in Q} a_{q,r'} = \frac{\sum_{r' \in Q} C(q,r')}{-\alpha_q} = 1$$

よって、$\alpha_q = -\sum_{r' \in Q} C(q,r')$。これを式(1)に代入して、

$$a_{q,r} = \frac{C(q,r)}{\sum_{r' \in Q} C(q,r')}$$

HMMの教師付学習
Supervised Learning of HMMs

- $b_{q,o}$を求める

$$\frac{\partial L}{\partial b_{q,o}} = \frac{C(q,o)}{b_{q,o}} + \beta_q = 0$$

ただし、$C(q,o) = \sum_{i=1}^{n} \sum_{t=1}^{T_i} [q_{it} = q][o_{it} = o]$

$C(q,o)$は訓練データ中で、状態qから記号oを出力した回数

$$\mathrm{b}_{q,o} = \frac{C(q,o)}{-\beta_q} \quad \cdots(1)$$

ところで$\sum_{o \in \Sigma} b_{q,o} = 1$であるから

$$\sum_{o' \in \Sigma} b_{q,o'} = \frac{\sum_{o' \in \Sigma} C(q,o')}{-\beta_q} = 1$$

よって、$\beta_q = -\sum_{o' \in \Sigma} C(q,o')$。式(1)に代入して、

$$b_{q,o} = \frac{C(q,o)}{\sum_{o' \in \Sigma} C(q,o')} = \frac{C(q,o)}{C(q)}$$

ただし、$C(q) = \sum_{i=1}^{n} \sum_{t=1}^{T_i} [q_{it} = q]$

$C(q)$は訓練データ中で、状態qが出現した回数

HMMの教師付学習
Supervised Learning of HMMs

- パラメータ推定

$$\pi_q = \frac{C_1(q)}{n} = \frac{\text{先頭の状態が}q\text{になっている文の数}}{\text{全文数}}$$

$$a_{q,r} = \frac{C(q,r)}{\sum_{r' \in Q} C(q,r')} = \frac{q\text{から}r\text{に遷移した回数}}{q\text{から任意の}r\text{に遷移した回数}}$$

$$b_{q,o} = \frac{C(q,o)}{\sum_{o' \in \Sigma} C(q,o')} = \frac{C(q,o)}{C(q)} = \frac{q\text{から}o\text{を出力した回数}}{q\text{の出現回数}}$$

ただし、 $C_1(q) = \sum_{i=1}^{n} [q_{i1} = q]$

$$C(q,r) = \sum_{i=1}^{n} \sum_{t=2}^{T_i} [q_{i(t-1)} = q][q_{it} = r]$$

$$C(q,o) = \sum_{i=1}^{n} \sum_{t=1}^{T_i} [q_{it} = q][o_{it} = o]$$

$$C(q) = \sum_{i=1}^{n} \sum_{t=1}^{T_i} [q_{it} = q]$$

HMMの教師無し学習: EMアルゴリズムの導入

HMMの教師無し学習
Unsupervised Learning of HMMs

- パラメータ推定
  - 訓練データ (入力)

    訓練データ : $x_1 x_2 \cdots x_n$

    x_i : 文を表す記号列(単語列)。 $x_i = o_{i1} o_{i2} o_{i3} \cdots o_{iT_i}$とする。

    T_i : x_iの記号列長

  - パラメータ (出力)

$$\pi, a, b = \arg\max_{\pi,a,b} \prod_{i=1}^{n} p(x_i)$$

$$= \arg\max_{\pi,a,b} \prod_{i=1}^{n} p(o_{i1} o_{i2} \cdots o_{iT_i})$$

$$= \arg\max_{\pi,a,b} \prod_{i=1}^{n} \sum_{q_1 \in Q, q_2 \in Q, \cdots q_{T_i} \in Q} p(q_1 o_{i1} q_2 o_{i2} \cdots q_{T_i} o_{iT_i})$$

HMMの教師無し学習
Unsupervised Learning of HMMs

- EMアルゴリズムによる教師無し学習
  - 不完全データ（欠損や曖昧性のあるデータ）に対する有名な学習法
  - EMアルゴリズム＋前向き後向きアルゴリズム

EMアルゴリズム

EMアルゴリズムの問題設定 (1/2)

- 実際に観測されたデータ$x_1,\ldots,x_N$が存在
- それぞれのデータx_iは隠れ状態$y_{i1},\ldots,y_{iT}$のいずれかから生成されたと仮定
  - 隠れ状態の集合はデータ毎に変わっても良い
    (機械学習一般には隠れ状態集合は固定であることが多い)
- パラメータ集合θにより$p(x, y)$が計算される

$$x_1 \longrightarrow \left\{ \begin{array}{ccc} y_{11} & y_{12} & y_{13} \\ p(x_1, y_{11}) & p(x_1, y_{12}) & p(x_1, y_{13}) \end{array} \right\}$$

$$x_2 \longrightarrow \left\{ \begin{array}{c} y_{21} \\ p(x_2, y_{21}) \end{array} \right\}$$

$$x_3 \longrightarrow \left\{ \begin{array}{ccccc} y_{31} & y_{32} & y_{33} & y_{34} & y_{35} \\ p(x_3, y_{31}) & p(x_3, y_{32}) & p(x_3, y_{33}) & p(x_3, y_{34}) & p(x_3, y_{35}) \end{array} \right\}$$

$$\vdots \qquad \vdots \qquad \vdots$$

EMアルゴリズムの問題設定 (2/2)

- パラメータ推定
  - 訓練データ (入力)

    訓練データ : $x_1 x_2 \cdots x_n$

    $Y(x)$: xに対する隠れ状態集合
  - パラメータ (出力)

$$\tilde{\theta} = \arg\max_{\theta} \prod_{i=1}^{n} p(x_i;\theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} \sum_{y \in Y(x_i)} p(x_i, y;\theta)$$

$$= \arg\max_{\theta} \log \prod_{i=1}^{n} \sum_{y \in Y(x_i)} p(x_i, y;\theta)$$

$l(\theta) = \log \prod_{i=1}^{n} \sum_{y \in Y(x_i)} p(x_i, y;\theta)$ とおくと $\tilde{\theta} = \arg\max_{\theta} l(\theta)$

EMアルゴリズムの全体像

$$\tilde{\theta} = \arg\max_{\theta} l(\theta)$$

問題変形

$$\theta^{(\tau+1)} = \arg\max_{\theta} Q(\theta^{(\tau)}, \theta)$$

個々の問題に応じて決まるQ関数の極値を解析的に求める

個々の問題によって決まるパラメータ更新式

[Eステップ] $p(y \mid x ; \theta)$ を計算

[Mステップ]

$$\theta^{(\tau+1)} = \arg\max_{\theta} Q(\theta^{(\tau)}, \theta)$$

によりパラメータ更新

Q関数の導出 (1)

- 問題: 実際に観測されたデータ$x_1,\ldots,x_n$が存在して、それに対して、対数尤度を最大化するパラメータを求める

$$\tilde{\theta} = \arg\max_{\theta} l(\theta) = \arg\max_{\theta} \sum_{i=1}^{n} \log p(x_i;\theta)$$

- 問題チェンジ: パラメータをθからθ'にした時の対数尤度の差を最大化することを繰り返せば極大値が求まる

$$\arg\max_{\theta'} \sum_{i=1}^{n} \{\log p(x_i;\theta') - \log p(x_i;\theta)\}$$

argmaxを求めているが、ようは正の値になればより尤度の高いパラメータが得られることに注意

Q関数の導出 (2)

- 個々の事象の対数尤度の差

$$\log p(x_i;\theta') - \log p(x_i;\theta) = \log \frac{p(x_i;\theta')}{p(x_i;\theta)} = \log \frac{p(x_i;\theta')}{p(x_i;\theta)} \sum_y p(y \mid x_i;\theta)$$

$$= \sum_y p(y \mid x_i;\theta) \log \frac{p(x_i;\theta')}{p(x_i;\theta)}$$

$$= \sum_y p(y \mid x_i;\theta) \log \left[\frac{p(x_i,y;\theta')}{p(x_i,y;\theta)} \frac{p(y \mid x_i;\theta)}{p(y \mid x_i;\theta')} \right]$$

$$= \sum_y p(y \mid x_i;\theta) \log \frac{p(x_i,y;\theta')}{p(x_i,y;\theta)} + \sum_y p(y \mid x_i;\theta) \log \frac{p(y \mid x_i;\theta)}{p(y \mid x_i;\theta')}$$

↑

ジェンセンの不等式より、常に≧0

Q関数の導出 (3)

- 個々の事象の対数尤度の差

$$\log p(x_i;\theta') - \log p(x_i;\theta) = \sum_y p(y \mid x_i;\theta) \log \frac{p(x_i, y;\theta')}{p(x_i, y;\theta)} + \sum_y p(y \mid x_i;\theta) \log \frac{p(y \mid x_i;\theta)}{p(y \mid x_i;\theta')}$$

$$\geq \sum_y p(y \mid x_i;\theta) \log \frac{p(x_i, y;\theta')}{p(x_i, y;\theta)}$$

$$= \sum_y p(y \mid x_i;\theta) \log p(x_i, y;\theta') - \sum_y p(y \mid x_i;\theta) \log p(x_i, y;\theta)$$

ここをQ関数とみなす

$$Q(\theta, \theta') = \sum_y p(y \mid x_i;\theta) \log p(x_i, y;\theta')$$

すると、ここは、

$$Q(\theta, \theta)$$

Q関数の導出 (4)

- まとめ
  - 対数尤度の差は次のようにおける

$$\log p(x_i;\theta') - \log p(x_i;\theta) \geq Q(\theta,\theta') - Q(\theta,\theta)$$

ただし $Q(\theta,\theta') = \sum_y p(y \mid x_i;\theta) \log p(x_i, y;\theta')$

- より良いパラメータθ'を見つけるためには、
  - $Q(\theta, \theta') - Q(\theta, \theta) \geqq 0$となれば良いが、
  - 効率を考えると、対数尤度の差が最大になるほうが良い
  - $Q(\theta, \theta)$はθ'に関わりなく一定なので、対数尤度の最大化するには、$Q(\theta, \theta')$を最大化すれば良い
  - $\theta' = \theta$とおくと(古いパラメータと同じにすると)Q関数の差は0になる⇒argmaxをとれば、常に$Q(\theta, \theta')$-$Q(\theta, \theta) \geqq 0$

EMアルゴリズム: Q関数の最大化

- 次のパラメータ更新を繰り返すアルゴリズム

$$\theta^{(\tau+1)} = \arg\max_{\theta} Q(\theta^{(\tau)}, \theta)$$

ただし、 $Q(\theta, \theta') = \sum_{y} p(y \mid x_i; \theta) \log p(x_i, y; \theta')$

全ての観測データ $x_1, x_2, \ldots, x_n$ に対しては、

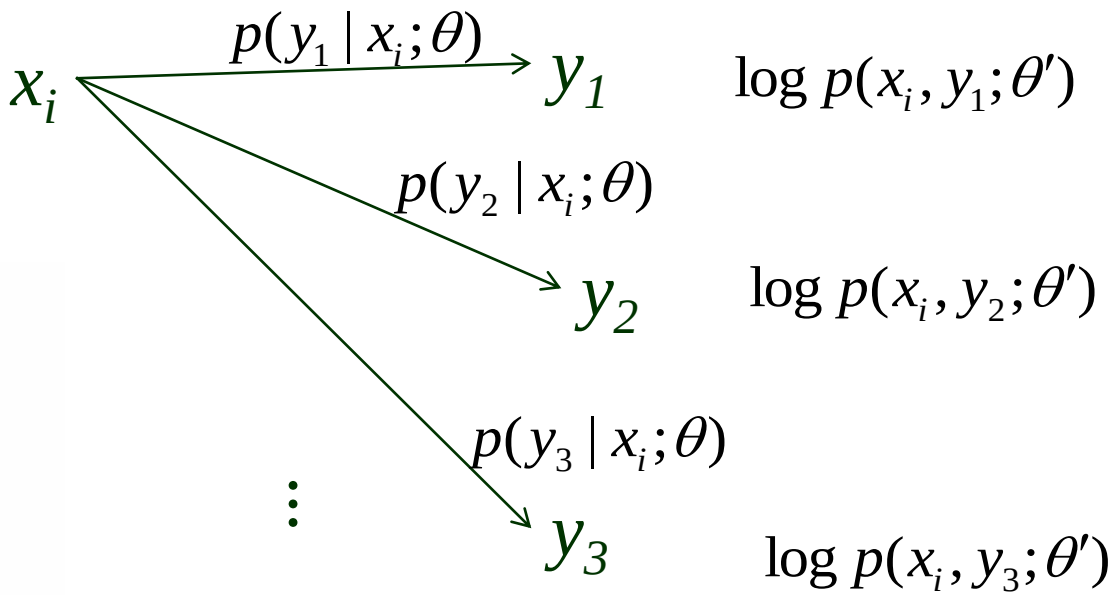
$$Q(\theta, \theta') = \sum_{i=1}^{n} \sum_{y} p(y \mid x_i; \theta) \log p(x_i, y; \theta')$$

とすればよい

しかし、まだ問題は解けていない！
argmax Qをどうやって求めるか？？

休憩: Q関数の直感的な意味 (1)

- Q関数 $Q(\theta,\theta') = \sum_y p(y \mid x_i;\theta) \log p(x_i, y;\theta')$
  - (古いパラメータθで計算した隠れ状態の条件付き確率)× (新しいパラメータθ'によるx_iとyの同時確率の対数)≒ x_iとyの同時確率の対数の期待値



休憩: Q関数の直感的な意味 (2)

- そもそもなぜ直接θを最大化しないのか？

$$
\begin{aligned}
l(\theta) &= \log \prod_{i=1}^{n} \sum_{y} p(x_i, y; \theta) \\
&= \sum_{i=1}^{n} \log \sum_{y} p(x_i, y; \theta) \\
&= ????
\end{aligned}
$$

⇒パラメータ更新式にすれば、実はこのsumをlogの外にだすことができるのであった

休憩: ジェンセンの不等式

ジェンセンの不等式

- 凸関数 $f(x)$ は区間 I 上の実数値関数
- $p_1,p_2,\ldots,p_n$は$p_1+p_2+\ldots+p_n=1$を満たす非負の実数
- 任意の$x_1,x_2,\ldots,x_n \in$ I に対し次の不等式が成り立つ

$$p_1 f(x_1)+p_2 f(x_2)+\cdots+p_n f(x_n) \geq f(p_1 x_1+p_2 x_2+\cdots+p_n x_n)$$

- $f(x)=-\log(x)$、$x_i=q_i/p_i$とおくと

$$\sum_i p_i \log \frac{p_i}{q_i} \geq -\log\left(\sum_i p_i \frac{q_i}{p_i}\right) = -\log \sum_i q_i = 0$$

まとめ

- HMMの教師付学習
- HMMの教師無し学習
  - EMアルゴリズムの導入
- EMアルゴリズム
  - Q関数の導出
- 資料

  http://aiweb.cs.ehime-u.ac.jp/~ninomiya/ai2/