知的情報処理システム特論　第11回

二宮　崇

今日の講義の予定

- HMMの教師なし学習
- EMアルゴリズム
- EMアルゴリズムの別の導出法と理解

- 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

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、

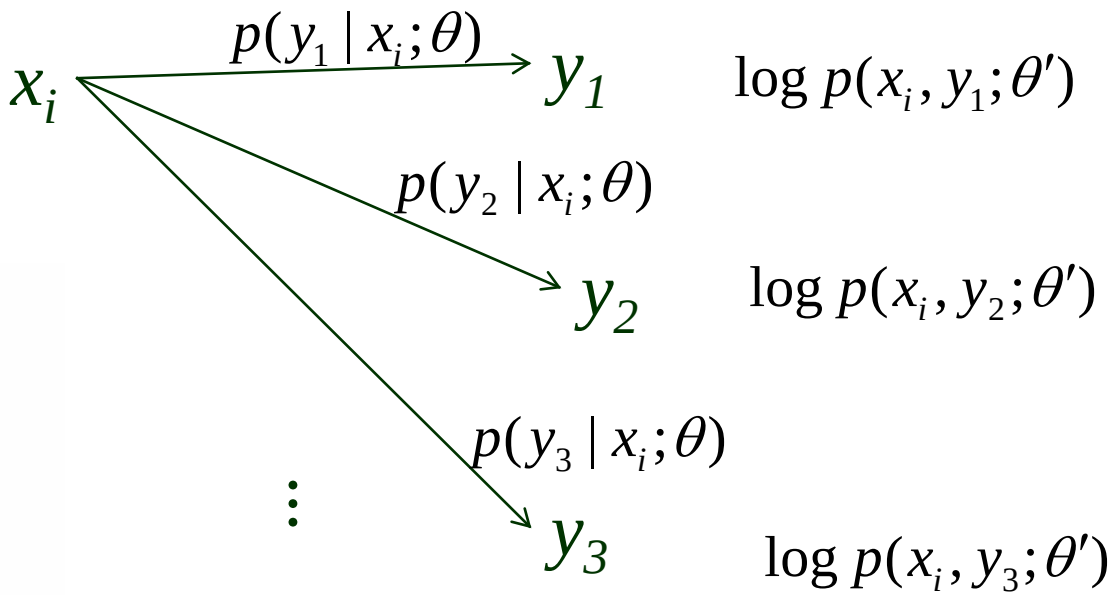
$$Q(\theta, \theta') = \sum_{i=1}^{n} \sum_{y} p(y \mid x_i; \theta) \log p(x_i, y; \theta')$$

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

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

EMアルゴリズムの別の導出法と理解

EMアルゴリズムの別の導出法と理解 1/3

- パラメータ: $\boldsymbol{\theta}$
- 入力: $\mathbf{x}$
- 隠れ状態: $\mathbf{y}$
- 観測データ: $X=\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$
- 対数尤度: $\log p_{\boldsymbol{\theta}}(X)$

$$
\begin{aligned}
\log p_{\boldsymbol{\theta}}(X) &= \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i) = \sum_{i=1}^{n} \sum_{\mathbf{y}_i \in Y(\mathbf{x}_i)} q(\mathbf{y}_i \mid \mathbf{x}_i) \log p_{\boldsymbol{\theta}}(\mathbf{x}_i) \\
&= \sum_{i=1}^{n} \sum_{\mathbf{y}_i \in Y(\mathbf{x}_i)} q(\mathbf{y}_i \mid \mathbf{x}_i) \left\{ \log p_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{y}_i) - \log p_{\boldsymbol{\theta}}(\mathbf{y}_i \mid \mathbf{x}_i) \right\} \\
&= \sum_{i=1}^{n} \sum_{\mathbf{y}_i \in Y(\mathbf{x}_i)} q(\mathbf{y}_i \mid \mathbf{x}_i) \left\{ \log p_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{y}_i) - \log q(\mathbf{y}_i \mid \mathbf{x}_i) - \log p_{\boldsymbol{\theta}}(\mathbf{y}_i \mid \mathbf{x}_i) + \log q(\mathbf{y}_i \mid \mathbf{x}_i) \right\} \\
&= \sum_{i=1}^{n} \sum_{\mathbf{y}_i \in Y(\mathbf{x}_i)} q(\mathbf{y}_i \mid \mathbf{x}_i) \left\{ \log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{y}_i)}{q(\mathbf{y}_i \mid \mathbf{x}_i)} - \log \frac{p_{\boldsymbol{\theta}}(\mathbf{y}_i \mid \mathbf{x}_i)}{q(\mathbf{y}_i \mid \mathbf{x}_i)} \right\} \\
&= \sum_{i=1}^{n} \sum_{\mathbf{y}_i \in Y(\mathbf{x}_i)} q(\mathbf{y}_i \mid \mathbf{x}_i) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{y}_i)}{q(\mathbf{y}_i \mid \mathbf{x}_i)} - \sum_{i=1}^{n} \sum_{\mathbf{y}_i \in Y(\mathbf{x}_i)} q(\mathbf{y}_i \mid \mathbf{x}_i) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{y}_i \mid \mathbf{x}_i)}{q(\mathbf{y}_i \mid \mathbf{x}_i)}
\end{aligned}
$$

EMアルゴリズムの別の導出法と理解　2/3

- パラメータ: $\boldsymbol{\theta}$
- 入力: $\mathbf{x}$
- 隠れ状態: $\mathbf{y}$
- 観測データ: $X=\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$
- 対数尤度: $\log p_{\boldsymbol{\theta}}(X)$

$$L(q,\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{y_i \in Y(x_i)} q(y_i \mid x_i) \log \frac{p_{\boldsymbol{\theta}}(x_i, y_i)}{q(y_i \mid x_i)}$$

$$KL(q \,\|\, p) = -\sum_{i=1}^{n} \sum_{y_i \in Y(x_i)} q(y_i \mid x_i) \log \frac{p_{\boldsymbol{\theta}}(y_i \mid x_i)}{q(y_i \mid x_i)} \geq 0$$

とおくと

$$\begin{aligned} \log p_{\boldsymbol{\theta}}(X) &= L(q,\boldsymbol{\theta}) + KL(q \,\|\, p) \\ &\geq L(q,\boldsymbol{\theta}) \end{aligned}$$

ポイント!
前ページのように式を展開するよりもここに

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{y}_i) = \log p_{\boldsymbol{\theta}}(\mathbf{y}_i \mid \mathbf{x}_i) + \log p_{\boldsymbol{\theta}}(\mathbf{x}_i)$$

を代入して等式が成り立つことを確認するほうがわかりやすい

EMアルゴリズムの別の導出法と理解　3/3

- Eステップ

$$q^{(\tau+1)}(\mathbf{y}_i \mid \mathbf{x}_i) = \arg\max_{q(\mathbf{y}_i|\mathbf{x}_i)} L(q(\mathbf{y}_i \mid \mathbf{x}_i), \boldsymbol{\theta}^{(\tau)}) = \arg\min_{q(\mathbf{y}_i|\mathbf{x}_i)} KL(q(\mathbf{y}_i \mid \mathbf{x}_i) \| p_{\boldsymbol{\theta}^{(\tau)}}(\mathbf{y}_i \mid \mathbf{x}_i)) = p_{\boldsymbol{\theta}^{(\tau)}}(\mathbf{y}_i \mid \mathbf{x}_i)$$

パラメータが固定されているので、$p_{\boldsymbol{\theta}}(X)$ は変わらない
→ Lを最大化 ⇔ KLを最小化

- Mステップ

$$\boldsymbol{\theta}^{(\tau+1)} = \arg\max_{\boldsymbol{\theta}} L(q^{(\tau+1)}, \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \sum_{\mathbf{y}_i \in Y(\mathbf{x}_i)} q^{(\tau+1)}(\mathbf{y}_i \mid \mathbf{x}_i) \log p_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{y}_i)$$

Q関数

隠れ状態の確率とパラメータを交互に動かして、
Lを最大化

まとめ

- HMMの教師無し学習
- EMアルゴリズム
  - Q関数の導出
- EMアルゴリズムの別の導出法と理解
- 資料

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