

言語解析論

講師 竹内孔一

本日の内容

- 確率
 - 最尤推定
- レポートの話

確率の計算

- 正解付きコーパスによる学習

$$(\hat{W}, \hat{S}) = \arg \max_{W \in L, S \in W} \prod_{i=1}^n P(w_i | s_i) P(s_i | s_{i-1})$$

$$P(S) = \prod_{i=1}^n P(s_i | s_{i-1}) \quad P(W | S) = \prod_{i=1}^n P(w_i | s_i)$$

- 頻度からの確率計算

$$P(s_i | s_{i-1}) = \frac{C(s_{i-1}, s_i)}{C(s_{i-1})} \quad P(w_i | s_i) = \frac{C(s_i, w_i)}{C(s_i)}$$

タグ付きコーパス

最尤推定

私/名詞 の/接続助詞 友人/名詞 は/格助詞 吹奏楽/名詞
団/接尾辞 に/格助詞 所属/サ変 して/動詞 いる/助動詞

統計による学習

- 例題

- 次のタグ付きコーパスがあるとき $P(\text{格助詞} | \text{名詞})$
 $P(\text{"吹奏楽"} | \text{名詞})$ の確率を求めなさい

タグ付きコーパス

私/名詞 の/接続助詞 友人/名詞 は/格助詞 吹奏楽/名詞
団/接尾辞 に/格助詞 所属/サ変 して/動詞 いる/助動詞

解答

$$P(\text{格助詞} | \text{名詞}) = C(\text{名詞}, \text{格助詞}) / C(\text{名詞}) = 1/3$$

$$P(\text{"吹奏楽"} | \text{名詞}) = C(\text{名詞}, \text{"吹奏楽"}) / C(\text{名詞}) = 1/3$$

練習12

- 次のタグ付きコーパスがあるとき
 $P(\text{格助詞} | \text{名詞})$, $P(\text{"遺跡"} | \text{名詞})$,
 $P(\text{"は"} | \text{格助詞})$ の確率を求めなさい

コーパス

私/名詞 の/接続助詞 友人/名詞 は/格助詞 吹奏楽/名詞
団/接尾辞 に/格助詞 所属/サ変 して/動詞 いる/助動詞
彼/名詞 は/格助詞 津島/固有名詞 遺跡/名詞 に/格助詞
よく/副詞 家族/名詞 で/助詞 かける/動詞

最尤推定

- 以下の計算が最尤推定から得られることを証明する

$$P(s_i | s_{i-1}) = \frac{C(s_{i-1}, s_i)}{C(s_{i-1})} \quad P(w_i | s_i) = \frac{C(s_i, w_i)}{C(s_i)}$$

- 最尤推定(maximum likelihood estimation)とは

- 分布型Pに関する先験情報 $\{\theta_1, \dots, \theta_N\}$ があるとき
Y回の観測事象 $\{x_1, \dots, x_Y\}$ に対して生起確率

$$\prod_{k=1}^Y P(x_k; \theta_1, \dots, \theta_N)$$

またはその対数確率

$$\sum_{k=1}^Y \log P(x_k; \theta_1, \dots, \theta_N)$$

を最大にする $\{\theta_1, \dots, \theta_N\}$ を推定する

最尤推定

P(助詞|名詞)の場合

θ_i は 名詞→助詞

n_i は 名詞 から 助詞へ
の遷移の回数



名詞→助詞に関連する

全事象の集合:

$Z_i = \{\text{名詞} \rightarrow \text{助詞}, \text{名詞} \rightarrow \text{名詞}, \dots\}$

- 推定対象の確率変数
 $\{\theta_1, \dots, \theta_N\}$
- Y個の学習データ
- θ_i の出現回数が n_i
- θ_i に関係する全事象の集合 Z_i
- 事象の生起確率
 $\{P(\theta_1), \dots, P(\theta_N)\} \rightarrow$ これを最尤推定

事象Nの出現回数が n_i 回であるような標本集合の総確率値の最大化

$$L = \sum_{i=1}^N n_i \log P(\theta_i) \quad (\text{ただし底} < 1 \text{とする})$$

$$\text{制約} \quad \sum_{i \in Z_i} P(\theta_i) = 1 \quad \text{Lagrange 未定乗数法}$$

最尤推定

Lagrange 関数

$$F = \sum_{i=1}^N n_i \log P(\theta_i) - \lambda \left(\sum_{i \in Z_i} P(\theta_i) - 1 \right)$$

これを θ_i で偏微分する

$$\frac{\partial F}{\partial P(\theta_i)} = \frac{n_i}{P(\theta_i)} - \lambda = 0 \quad \text{よって} \quad P(\theta_i) = \frac{n_i}{\lambda} \quad \textcircled{1}$$

制約式に代入して

$$1 = \sum_{i \in Z_i} P(\theta_i) = \frac{1}{\lambda} \sum_{i \in Z_i} n_i \quad \lambda = \sum_{i \in Z_i} n_i$$

よって λ を $\textcircled{1}$ に代入して

$$P(\theta_i) = \frac{n_i}{\sum_{i \in Z_i} n_i}$$

参考: 奈良先端大:
音声情報処理講義録
鹿野先生 1994年

ラグランジュ未定乗数法

- 制約の中で式を解く
 - M個の制約条件 $g_i(x) = 0 \quad (i=1, 2, \dots, M)$
 - $f(x)$ の極値をとりたい (極大or極小)

ラグランジュ乗数 $\alpha = [\alpha_1, \dots, \alpha_M]^T$ を使って

$$L(X, \alpha) = f(X) - \sum_{i=1}^M \alpha_i g_i(X)$$

$$\frac{\partial L}{\partial x_k} = 0 \quad \frac{\partial L}{\partial \alpha_i} = 0$$

$$(k=1, 2, \dots, N) \quad (i=1, 2, \dots, M) \quad \text{を解けばよい.}$$

練習13

- Lagrange 未定乗数法を利用して次の問題を解きなさい
 - 立方体の辺の長さの合計が 24m のときその体積を最大化するには辺の長さをいくらずつにすればよいか

ヒント: 辺を a, b, C とおくと $4(a+b+c)=24$ が制約

形態素解析器

- 現状

Free の形態素解析システムがある

奈良先端大(松本研究室) ChaSen

<http://chasen.naist.jp/hiki/ChaSen/>
for linux and windows

京都大学(黒橋研究室) JUMAN

<http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html>
for linux