最大エントロピー法についてのメモ

NAIST 自然言語処理学講座 M2 林部 祐太

DMLA@2010.4.23

1 このメモについて

私が DMLA 勉強会にて論文「A simple introduction to maximum entropy models for natural language processing」を紹介した時に，配布した資料に，紹介中に頂いた貴重なコメントを元に加筆修正を加えたものです．頂いたコメントに感謝しつつ，どなたかの参考になれば幸いと思い，公開させていただきます．

2 最大エントロピー法

2.1 概要

以下，文献[2]の説明を抜粋しつつ説明を行う．基本的な考え方は「データは特徴をすべて反映してるものだから，データから分からない所の分布は一様分布にしよう」という「最大エントロピー原理」を用いて，データから確率分布を求める手法で，最尤法の一種である．最尤法とは観測されたデータが理論的に最も起こりやすいようにパラメータを決める手法である．関連概念としてベイジアンがあるがここでは述べない．なお，最尤法をベイジアンで解釈することも可能であり，事前分布に一様分布を仮定し，事後分布のモードを計算することと同値である．

最大エントロピー法は，対数線形 (log-lenear) モデルとも言われる．スコアの対数 (log をとったもの) は，各素性の (値×重み) の線形和となる．最大エントロピー法は，「素性に対する制約 + エントロピー最大化によるモデル推定」で，対数線形モデルは「log-lenear で表現される確率モデルの最尤推定」であるが，これら 2 つは実は同じ結果になることが最近分かったようだ[1]．実際，エントロピー最大原理に基づき，それをラグランジュの未定乗数法で解くと対数線形モデルの式が出てくる．

なお，識別学習の手法で，近年注目を浴びている SVM(Support Vector Machine) があり，これは「関数」でデータを識別するものである．したがって外れ値 (outliner) に弱いというでメリットがあるが，カーネルを決めさえすれば，あとは適当に素性を入れていくだけで，人手で素性選択をしなくても，自動で素性の組み合わせを考慮してくれるというメリットがある．

一方，最大エントロピー法は，確率分布を考えることで，未知のデータを「識別」する．人手で素性選択をしなくてはならないのがデメリットであるが，多少の外れ値には影響を受けにくいのはメリットである．明らかに役立つ素性が決まっている場合は最大エントロピー法で十分であろう．素性の組み合わせが膨大になり，学習が困難になるため，素性の頻度での足切り (一定回数以上出てこない素性は学習時に考えない) を行ったり，人手で有効な素性の組み合わせを考えたり (素性選択) する．

また，過学習を避けるため，エントロピー計算時に正規化項を入れたりする．すなわち，些細な素性に重みを付けることを避けるため (頻度の低い素性の影響力を弱め，モデルの過度の複雑化

を避ける)，重みの 2 乗したものの和であるガウス[*1]正規化項 (=L2 正則化) や，重みの絶対値の和であるラプラシアン正規化項 (=L1 正則化) など) を導入したりする．

2.2 導出

最大エントロピー原理 (principle of maximum entropy) では，素性関数で表された制約の下で，エントロピーを最大にする，すなわち，最も均一分布に近い分布を選ぶ．入力ベクトル $\mathbf{x}_j$ と出力クラス y_j を考える．最大エントロピー法とは，「最大エントロピー原理」に基づいて確率分布を推定することである．

素性関数 (feature function) は $\mathbf{x}_j$ と y_j が特定の条件を満たすとき 1，そうでないとき 0 をとる．例えば，文書を $\mathbf{x}_j$ で表すとき，文書に「機械学習」の語が含まれていて，クラス y_j が「機械学習」のカテゴリなら $f_i(\mathbf{x}, y)$ は 1 をとり，そうでなければ 0 をとる．

真の分布 $P(\mathbf{x}, y)$ と経験分布 $\hat{P}(\mathbf{x}, y)$ ，それぞれの下での素性関数の期待値が等しいという制約

$$\mathrm{E}_P[f_i] = \mathrm{E}_{\hat{P}}[f_i]$$

の下で，次のエントロピーを最大にする $P(y|\mathbf{x})$ を求める．

$$P^* = \arg\max_P H(P) = -\arg\max_P \sum_{\mathbf{x},y} \hat{P}(y)P(y|\mathbf{x}) \log P(y|\mathbf{x})$$

これは，ラグランジュの未定乗数法を用いて求める．すなわち，

- 各 i について，$\mathrm{E}_P[f_i] = \mathrm{E}_{\hat{P}}[f_i]$
- $\sum_i p(x_i) = 1$

という制約を満たすように極値 P^* を求めると，

$$L = H(p) + \sum_i \lambda_i (p(x)f_i(x) - \hat{p}(x)f_i(x)) + \lambda_0 \left(\sum_i p(x_i) - 1 \right)$$

に関して，

$$\frac{\partial L}{\partial x} = 0, \frac{\partial L}{\partial \lambda} = 0$$

を解くと良い．

$$\begin{aligned}
\frac{\partial L}{\partial x} &= (-\log p(x) - 1) + \sum \lambda_i f_i(x) + \lambda_0 = 0 \\
\log p(x) &= -1 + \sum \lambda_i f_i(x) + \lambda_0 \\
p(x) &= \exp\left(\sum \lambda_i f_i(x) + C \right) \quad \{C = \lambda_0 - 1\} \\
p(x) &= \pi \exp\left(\sum \lambda_i f_i(x) \right) \quad \{\pi = e^{\lambda_0 - 1}\} \\
&= \pi \prod \exp\left(\lambda_i f_i(x)\right) \\
&= \pi \prod \alpha_i^{f_i(x)} \quad \{\alpha_i = e^{\lambda_i}\}
\end{aligned}$$

ここで λ_i は素性関数 f_i の重み，π は正規化項，α_i はモデルのパラメータと見ることができる．パラメータの学習法は色々あり，本文献では GIS を紹介してあるが，現在は (準) ニュートン法が主流である．

2.3 KL 距離

本文献の 4 章から 6 章ではそのような P^* が存在し，かつ唯一であることを，エントロピーの立場からと，最尤法の立場からの 2 種類の方法で証明している．そのために KL ダイバージェンス (Kullback-Leibler divergence) を使って説明している．KL 距離とは 2 つの確率分布の違いを計る尺度であるが，その定義は

$$\begin{aligned}
D(p, q) &= \sum p(x) \log \frac{p(x)}{q(x)} \\
&= -\sum p(x) \log q(x) + \sum p(x) \log p(x) \\
&= H(p, q) - H(p)
\end{aligned}$$

であり，「p と q のクロスエントロピー」と「p のエントロピー」との差で表すことができる．すなわち，KL ダイバージェンスは、真の確率分布 p であるような符号に比較して，ある間違った最適化をされた確率分布 q の符号化をしたときに，余分にかかる符号長の予測値を表していると見ることができる．

[*1] 平均 0 のガウス分布を重みの事前分布に考え，事後分布を考えてる

いま考えるのは，エントロピー最小化で，それは一様分布との KL 距離最小化と同値である．

参考文献

[1] 二宮崇, 東京大学 数理言語情報論 2010 年第 12 回講義, http://www.r.dl.itc.u-tokyo.ac.jp/~ninomi/mistH21w/cl/mistH21w-ninomi-12.pdf

[2] 朱鷺の杜, 最大エントロピー, http://ibisforest.org/index.php?最大エントロピー

[3] Ratnaparkhi A., "A simple introduction to maximum entropy models for natural language processing", Technical Report 97-08, Institute for Research in Cognitive Science, University of Pennsylvania, 1997