データサイエンス・シリーズ 6

データ学習アルゴリズム

(渡辺澄夫著, 共立出版 2001)

発表日: 平成 15 年 8 月 1 日
担当者: 岩崎 唯史
担当箇所: 3.2 競合学習
3.2.1 確率競合モデル
3.2.2 混合正規モデルの推論
3.2.3 混合分布の最急降下法

3.2 競合学習

3.2.1 確率競合モデル

■競合的な確率変数とは? [p.12, 例 15]

H 個の有限集合

$$\mathcal{C}^H = \{\underbrace{(1,0,0,\ldots,0)}_{H\text{個}}, (0,1,0,\ldots,0), \ldots (0,0,0,\ldots,1),\}$$

上の確率変数 $U = (U_1, U_2, \ldots, U_H)$ の内，あるユニットだけが $1(U_i = 1)$ で，残りは全て $0(U_h = 0, h \neq i)$.

⬇ 例えて言うと...

多数の要素 (解の候補) が限られた資源 (“もっともらしさ” の確率) をめぐって “All or Nothing” 方式で競い合う.

■入力と出力をもつ確率競合モデル

入力 $x \in R^M$, 出力 $y \in R^N$, 競合的確率変数 (中間ユニット) $u \in \mathcal{C}^H$.
直積集合 $R^M \times R^N \times \mathcal{C}^H$ 上の確率変数 (x, y, u) の密度関数 ※

$$p(x, y, u|w) = \prod_{h=1}^{H} \{a_h \, q(x|b_h) \, r(y|c_h)\}^{u_h},$$

をもつもの (競合的な確率変数を推論の途中にもつ学習モデル) を**確率競合モデル**という. ここで,

$$w = \{\{a_h, b_h, c_h\}; h = 1, 2, \ldots, H, \sum_{h=1}^{H} a_h = 1, a_h > 0\},$$

はモデル中のパラメータ.

※ [p.12]: 確率 a_h で h 番目のユニットが 1 になり, 残りのユニットが 0 になる競合的な確率変数 u の密度関数は $p(u|a) = \prod_{h=1}^{H} (a_h)^{u_h}$.

特に $q(x|b_h)$, $r(y|c_h)$ が正規分布である場合,

$$q(x|b_h) = \frac{1}{(2\pi\sigma_h^2)^{M/2}} \exp\left(-\frac{\|x-\xi_h\|^2}{2\sigma_h^2}\right),$$

$$r(y|c_h) = \frac{1}{(2\pi\rho_h^2)^{N/2}} \exp\left(-\frac{\|y-\eta_h\|^2}{2\rho_h^2}\right).$$

モデル中のパラメータは

$$b_h = \{\xi_h \in R^M, \sigma_h \in R; \sigma_h > 0\},$$

$$c_h = \{\eta_h \in R^N, \rho_h \in R; \rho_h > 0\}.$$

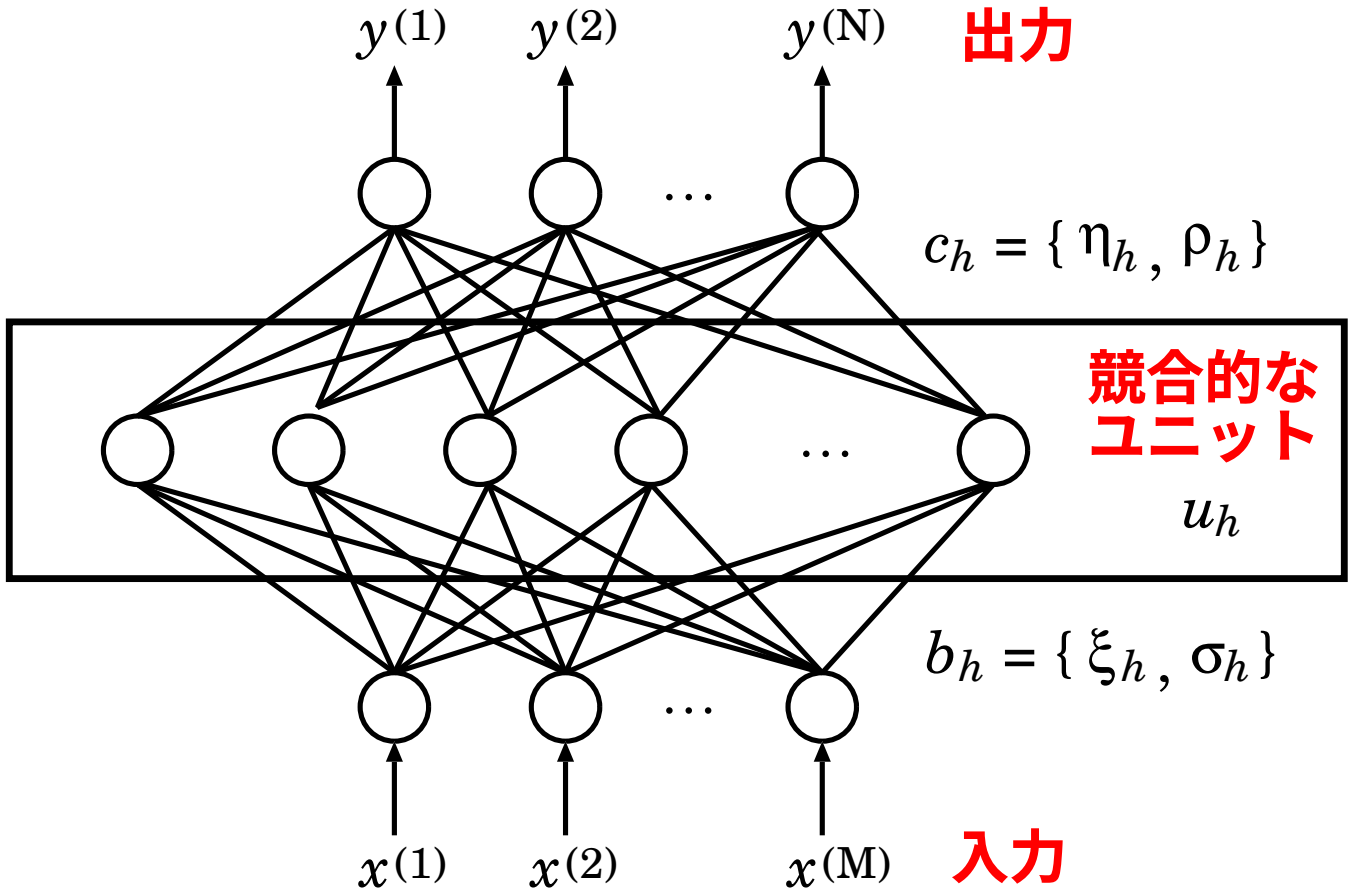

図. 確率競合モデル. 入力 x が与えられたとき, 中間層のユニットのうちの 1 つだけが確率的に 1 となり, 他のユニットは 0 になる.

入力と出力だけに着目し, 中間層 (競合的確率変数 u_h) について平均すると, 密度関数は混合 (正規) 分布となる.

$$
\begin{aligned}
p(x, y|w) &= \sum_{\{u\}} \prod_{h=1}^{H} \{a_h\, q(x|b_h)\, r(y|c_h)\}^{u_h} \\
&= \sum_{h=1}^{H} a_h\, q(x|b_h)\, r(y|c_h).
\end{aligned}
$$

入力と出力だけに着目すると, 中間層の出力を平均値で置き換えた確率競合モデルは混合分布と等価.

■教師なし学習

入力 X と出力 Y を合わせて 1 つの入力と考え, X と Y の組を改めて X とすると, 確率変数 (x, u) の密度関数は

$$p(x, u|w) = \prod_{h=1}^{H} \{a_h \, q(x|b_h)\}^{u_h}.$$

入力 X の密度関数を学習するモデルであり, 入力を確率的な H 個の異なる概念に自動分類するものと考えることができる.

「教師なし学習」※ または「自己組織化」

※正解を直接与えず, 自分で基準を設定しながらデータを分類・整理する方法.

3.2.2 混合正規モデルの推論

■ 条件付き確率とその注意点

入力 x が与えられたときの出力 Y の条件付き確率密度関数

$$p(y|x,w) = \frac{p(x,y|w)}{p(x|w)} = \frac{\sum_{h=1}^{H} a_h\, q(x|b_h)\, r(y|c_h)}{\sum_{h=1}^{H} a_h\, q(x|b_h)}.$$

分母 (入力 x をどの程度よく知っているか) が $p(x|w) \approx 0$ となる x に対しては, 推論 $p(y|x,w)$ が不定値に近くなるため, 推論を行うべきではない.

「知らない入力には答えられない (答えるべきでない)」

■ 入力から出力の推論と出力から入力の推論

入力 x が与えられとき, 条件付き確率 $p(y|x,w)$ に従う確率変数 y の生成手順:

(手順 1) $h = 1, 2, \ldots, H$ の中から, 次の確率でユニット h を選択.

$$p_h = \frac{a_h\, q(x|b_h)}{\sum_{k=1}^{H} a_k\, q(x|b_k)}.$$

(手順 2) $r(y|c_h)$ に従う確率変数として出力 Y を生成.

- **入力 x が与えられときの出力 Y の平均 $E(Y|x)$:**

$$E(Y|x) = \int y\, p(y|x,w)dy = \frac{\sum_{h=1}^{H} a_h\, \eta_h\, q(x|b_h)}{\sum_{h=1}^{H} a_h\, q(x|b_h)}.$$

- **上記手順を逆に使うと, 出力 y が与えられたとき, 入力 X を生成するモデルを構成できる.**

3.2.3 混合分布の最急降下法

■ 競合モデルにおける同時確率の学習アルゴリズム

入力と出力は同等であるので, 入力だけがある場合を考えれば十分.

入力 x に対する損失関数

$$\mathcal{E}(x,w) = -\log p(x|w) = -\log\left(\sum_{h=1}^{H} a_h\, q(x|b_h)\right),$$

を最小にするようなパラメータ $w = \{a_h, b_h\}$ を探す.

最急降下法: 初期値 $w(0)$ から始め, $w(t+1) = w(t) - \lambda\nabla\mathcal{E}(x, w(t))$ を繰り返す. ただし, $\nabla\mathcal{E}(x,w(t)) = (\partial\mathcal{E}/\partial a_h, \partial\mathcal{E}/\partial b_h)\,\mathbf{at}\, w = w(t)$.

特に

$$a_h = \frac{\exp(\alpha_h)}{\sum_{h'=1}^{H} \exp(\alpha_{h'})},$$

で $q(x|b_h)$ **が平均** ξ_h**, 分散** σ_h^2 **の正規分布の場合, パラメータ** $w = \{\alpha_h, \xi_h, \sigma_h\}$ **に関する最急降下ベクトル** $\nabla \mathcal{E}(x, w)$ **の各成分は**

$$\begin{aligned}
\frac{\partial \mathcal{E}(x,w)}{\partial \alpha_h} &= -A_h(x,w) + \frac{\exp(\alpha_h)}{\sum_{h'=1}^{H} \exp(\alpha_{h'})}, \\
\frac{\partial \mathcal{E}(x,w)}{\partial \xi_h} &= A_h(x,w) \left(\frac{\xi_h - x}{\sigma_h^2} \right), \\
\frac{\partial \mathcal{E}(x,w)}{\partial \sigma_h} &= A_h(x,w) \left(\frac{M}{\sigma_h} - \frac{\|x - \xi_h\|^2}{\sigma_h^3} \right), \\
A_h(x,w) &\stackrel{\text{def}}{=} \frac{a_h \, q(x|b_h)}{p(x|w)} = \frac{\exp(\alpha_h) q(x|b_h)}{p(x|w) \sum_{h'=1}^{H} \exp(\alpha_{h'})}.
\end{aligned}$$

■ 学習アルゴリズムの注意点

- 入出力をもつ学習モデルでは, 同時確率を用いた学習則と条件付き確率を用いた学習則は異なる.
- 過学習に注意.
- アルゴリズムの実装に際しては, $w(t)$ の初期値と変更量の設定に微妙な調整が必要.
- σ_h が 0 になりにくくする必要あり ($\nabla\mathcal{E}(x,w)$ の分母にあるため, $w(t)$ の変更量に対する影響大).

◆ [注 24]: 混合正規分布では $H \geq 2$ で最尤推定量が存在しない

学習用サンプルを $\{(x_i, y_i); i = 1, 2, \ldots, n\}$ とし，モデル

$$p(x|w) = \sum_{h=1}^{2} a_h \, q(x|b_h),$$

による密度関数の学習を考える．最尤推定を行う場合の損失関数は

$$L_n(w) = -\frac{1}{n} \sum_{i=1}^{n} \log p(x_i|w).$$

ここで，a_2, b_2 を固定，$\xi_1 = x_1$ としたとき，$\sigma_1 \to 0$ で $L_n(w) \to -\infty$ ($\sigma_h = 0$ で損失関数 $L_n(w)$ が解析性を失う)．

混合正規分布では最尤推定は意味を持たない．

- 性質のよい局所最尤解を見つけることが大切．
- 非線形なパラメータをもつ学習系では同様の問題が頻繁に生じる．

◆ [注 25]: 混合正規分布の近似学習則

パラメータ σ_h, a_h が一定 ($\sigma_h = 1$, $a_h = 1/H$ $(h = 1, 2, \ldots, H)$) の場合, モデルのパラメータは ξ_h(正規分布の平均) のみ.

混合正規分布の学習式 (最急降下法) の近似アルゴリズム:

入力例 x が与えられたとき, $\|\xi_h - x\|$ $(h = 1, 2, \ldots, H)$ を最小にするものを初期値 $\xi_h(0)$ に選び, 次式を繰り返す.

$$\xi_h(t+1) = \xi_h(t) - A \cdot (\xi_h(t) - x), \qquad (A \text{ は正の小さな定数}).$$

混合正規分布の学習則は, 上記の学習則 (与えられた入力に一番近いベクトルを初期値に選び, 入力の方角に少しずつ移動させていく) を一般化したもの.