多変数ロジスティック分類

分類問題とは何か

典型的な分類の問題の例として以下が挙げられる

- 手書きの文字を認識する
- 罹患している病気を症状から推定する
- メールのスパムかどうかを判定する

このような問題をどのようにして確率的にモデリングするのか?
今回はそのような点について焦点をあてて考察していく

以下ではエントロピー最大化を第一原理としてロジスティック分類を導き、トレーニングデータが与えられたときの最適解の導出まで行う

ロジスティック分布の導出

確率モデルで分類問題を表す

以下では一般的な形でモデリング化するため、入力の変数を $\mathbf{x} \in \mathbf{R^D}$ として出力のラベルを $C \in \{C_1, C_2, \ldots, C_K\}$ とする。今回は確率モデルとして定式化するため、ある入力が与えられたときの特定のラベルが出現する確率を考えればよい。

これは条件付き確率 $P(C|\mathbf{x})$ として表せる。したがって、ある入力値 $\mathbf{x}$ が与えられときの最適なラベルは、この条件付き確率を最大化させるようなラベルである。
この最適なラベルを $C^\star$ とすれば以下のようにして

$$C^\star = \underset{k}{\operatorname{argmax}}\, P(C_k|\mathbf{x}) \tag{1}$$

未知の入力値に対するラベルの予測を与えることができる。

以後は一般化も踏まえて、入力ベクトルを固有の特徴を抽出する変換 $\boldsymbol{\phi}(\mathbf{x}) \in \mathbf{R^K}$ を加えたものを前提に考える。

最大エントロピー原理

さて、上記の議論はあくまで何らかのパラメーター $\boldsymbol{\omega}$ を用いて上記の条件付き確率を制限して $P(C|\mathbf{x};\boldsymbol{\omega})$ のように表せないと、議論がこれ以上先に進めない。

そのため最大エントロピー原理を用いて、トレーニングデータが与えられときに、尤もらしい確率分布がどのように表せるかを考えたい。

以後は計算を簡単にするために、出力のラベルはすべて 1 of K 符号化で表されているとしよう。1 of K 符号化でラベルがエンコードされる場合、正解データが C_k とすると L 次元ベクトル $\mathbf{t} \in \mathbf{R^L}$ として表せる。このとき $t_i \in \{0, 1\}$ かつ $t_i = \delta_{i,k}\ (i = 1, 2, \ldots, L)$ が成り立っている。
デルタ関数の定義は以下の通りである。

$$\delta_{i,k} = \begin{cases} 1 & (i = k) \\ 0 & (i \neq k) \end{cases}$$

定義がややこしそうだが、結局ラベルが L 個あったら L 次元ベクトルとして表し、C_k が正解データであれば k 番目の要素を 1 としてそれ以外を 0 とするベクトルである。

ここで、トレーニングデータとそのラベルをそれぞれ

$$(\boldsymbol{\phi}(\mathbf{x^{(1)}}), \mathbf{t^{(1)}}), (\boldsymbol{\phi}(\mathbf{x^{(2)}}), \mathbf{t^{(2)}}), \ldots\ (\boldsymbol{\phi}(\mathbf{x^{(N)}}), \mathbf{t^{(N)}}) \tag{2}$$

が与えられたとする。
このとき

$$\sum_{k=1}^{K} P(C_k|\mathbf{x^{(n)}}) = 1 \qquad (n = 1, 2, \ldots, N) \tag{3}$$

$$\sum_{n=1}^{N} P(C_k|\mathbf{x^{(n)}})\, \boldsymbol{\phi}(\mathbf{x^{(n)}}) = \sum_{n=1}^{N} t_k^{(n)} \boldsymbol{\phi}(\mathbf{x^{(n)}}) \qquad (k = 1, 2, \ldots, K) \tag{4}$$

が満たさなければならないと仮定しよう。

(3) は確率の定義より明らかに満たさなければならない。
(4) についてはいわゆる $P(C_k|\mathbf{x^{(n)}})$ が十分 $t_k^{(n)}$ をよく表さなければならない、という制約である。条件付きエントロピーは $-\sum_{k=1}^{L} P(C_k|\mathbf{x^{(n)}}) \ln P(C_k|\mathbf{x^{(n)}})$ より、これを (3), (4) の制約の元で最大化すればよい。
$P_k^{(n)} = P(C_k|\mathbf{x^{(n)}})$ のように簡易的に表すことにすれば、ラグランジュの未定乗数法より

$$\begin{aligned} H(p) = &\sum_{n=1}^{N} \sum_{k=1}^{K} -P_k^{(n)} \ln P_k^{(n)} \\ &+ \sum_{n=1}^{N} \lambda^{(n)} \{ \sum_{k=1}^{K} P_k^{(n)} - 1 \} \\ &+ \sum_{k=1}^{K} {\omega_{\mathbf{k}}}^t \{ \sum_{n=1}^{N} \boldsymbol{\phi}(\mathbf{x^{(n)}})(P_k^{(n)} - t_k^{(n)}) \} \end{aligned} \tag{5}$$

を最大にするような $P_k^{(n)}$ を求めればよいことがわかる。
ここでスラッグ変数 $\lambda^{(n)}$ と ${\omega_{\mathbf{k}}}^t$ を導入した。

$P_k^{(n)}$ を求める

式の定式化までは行えたのであとは $H(p)$ を単純に $P_d^{(m)}$ で微分すればよい。

$$\begin{aligned}\frac{\partial H(p)}{\partial P_d^{(m)}} &= \sum_{n,k}\left\{-\frac{\partial P_k^{(n)}}{\partial P_d^{(m)}}\{\ln P_d^{(m)}+1\}+\lambda^n\{\frac{\partial P_k^{(n)}}{\partial P_d^{(m)}}\}+\omega_{\mathbf{k}}^{\mathbf{t}}\phi(\mathbf{x^{(n)}})\frac{\partial P_k^{(n)}}{\partial P_d^{(m)}}\right\}\\ &= -lnP_d^{(m)}-1+\lambda^{(m)}+\omega_{\mathbf{d}}^{\mathbf{t}}\phi(\mathbf{x^{(m)}})\end{aligned}$$

のように求まるので、以下のように微分をゼロをおけば

$$\frac{\partial H(p)}{\partial P_d^{(m)}} = 0 \tag{6}$$

$$P_d^{(m)} = \exp\left\{\lambda^{(m)}-1+\omega_{\mathbf{d}}^{\mathbf{t}}\phi(\mathbf{x^{(m)}})\right\} \tag{7}$$

(7) を (3) に代入すれば

$$\exp\left\{\lambda^{(m)}-1\right\} = \exp\left(\omega_{\mathbf{d}}^{\mathbf{t}}\phi(\mathbf{x^{(m)}})\right) \tag{8}$$

より (8) を (7) に代入して添字を整理すれば、

$$P_k^{(n)} = P(C_k|\mathbf{x^{(n)}}) = \frac{\exp\left(\omega_{\mathbf{k}}^{\mathbf{t}}\phi(\mathbf{x^{(m)}})\right)}{\sum_{d=1}^{K}\exp\left(\omega_{\mathbf{d}}^{\mathbf{t}}\phi(\mathbf{x^{(m)}})\right)} \tag{9}$$

と表せる。
このようにして目的であった条件付き確率分布がパラメーター $\omega_{\mathbf{d}}^{\mathbf{t}}$ を用いて表せるところまで求めることができた

(9) は多変数のロジスティック分布である

多変数ロジスティック分布の最尤推定

条件付き確率分布が得られたので (2) のトレーニングデータが与えられたときに負の対数尤度は以下のように表せる。

$$H(\mathbf{W}) = -\ln\left\{\prod_{n=1}^{N}\prod_{k=1}^{K}P(C_k|\mathbf{x^{(n)}})^{t_k^{(n)}}\right\} \tag{10}$$

$$= -\sum_{k,n}t_k^{(n)}\ln P_k^{(n)} \tag{11}$$

この対数尤度を最大化するような $\mathbf{W} = \omega_{\mathbf{k}}^{\mathbf{t}}\ \ (k=1,2,\ldots K)$ を最急勾配法によって表せればよい。

ただしここで

$$P_k^{(n)} = \frac{\exp\left\{a_k^{(n)}\right\}}{\sum_{d=1}^{K} \exp\left\{a_d^{(n)}\right\}} \tag{12}$$

ただし

$$\begin{aligned} a_k^{(n)} &= a_k(\mathbf{x^{(n)}}) \\ &= \omega_{\mathbf{k}}^{\mathbf{t}} \phi(\mathbf{x^{(m)}}) \\ &= \sum_{d=1}^{D} w_{kd} \phi_d(\mathbf{x^{(n)}}) \\ &= \sum_{d=1}^{D} w_{kd} \phi_{dn} \quad (\phi_d(\mathbf{x^{(n)}}) = \phi_{dn} \text{とした}) \end{aligned}$$

とする。
(9) の関係式を変数の依存関係で分割しただけである。

このとき P_k に対して a_j の微分を考えると

$$\begin{aligned} \frac{\partial P_k}{\partial a_j} &= \frac{\partial}{\partial a_j} \left\{ \frac{\exp\{a_k\}}{\sum_{d=1}^{K} \exp\{a_d\}} \right\} \\ &= \left\{ \frac{\partial}{\partial a_j} (\exp\{a_k\}) \right\} \frac{1}{\sum_{d=1}^{K} \exp\{a_d\}} + \\ &\exp\{a_k\} (-\frac{1}{(\sum_{d=1}^{K} \exp\{a_d\})^2}) \frac{\partial}{\partial a_j} \left\{ \sum_{d=1}^{K} \exp\{a_d\} \right\} \\ &= \frac{\exp\{a_k\}}{(\sum_{d=1}^{K} \exp\{a_d\})} (\delta_{jk} - \frac{\exp\left\{a_j\right\}}{(\sum_{d=1}^{K} \exp\{a_d\})}) \\ &= P_k(\delta_{kj} - P_j) \end{aligned}$$

より

$$\frac{\partial P_k}{\partial a_j} = P_k(\delta_{kj} - P_j) \tag{13}$$

が成り立つため (11) を ω_{mj} に対して微分すると

$$
\begin{aligned}
\frac{\partial H(\mathbf{W})}{\partial \omega_{mj}} &= \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{l=1}^{K} t_k^{(n)} \left\{ \frac{\partial}{\partial a_l^{(n)}} ln P_k^{(n)} \right\} \frac{\partial a_l^{(n)}}{\partial \omega_{mj}} \\
&= -\sum_{n,k,l} t_k^{(n)} P_k (\delta_{kl} - P_j) \frac{\partial}{\partial \omega_{mj}} \left\{ \sum_{d=1}^{D} \omega_{ld} \phi_{dn} \right\} \\
&= -\sum_{n,k,l} t_k^{(n)} P_k (\delta_{kl} - P_j) \delta_{ml} \phi_{jn} \\
&= \sum_{n=1}^{N} \left\{ P_m^{(n)} - t_m^{(n)} \right\} \phi_{jn}
\end{aligned}
$$

添字を差し替えて

$$
\frac{\partial H(\mathbf{W})}{\partial \omega_{kd}} = \sum_{n=1}^{N} \left\{ P_k^{(n)} - t_k^{(n)} \right\} \phi_{dn} \tag{14}
$$

と与えられることがわかる。

最急降下法の規則まとめ

微分が得られたので最急降下法で停留解を得ることが可能になる。
これまでの議論をまとめると、規則は以下で与えられる

$$
\begin{aligned}
\omega_{kd}^{(new)} &= \omega_{kd} + \eta \frac{\partial H}{\partial \omega_{kd}} \\
&= \omega_{kd} + \eta \sum_{n=1}^{N} \left\{ P_k^{(n)} - t_k^{(n)} \right\} \phi_{dn} \\
P_k^{(n)} &= \frac{\exp\{a_k\}}{\sum_{d=1}^{K} \exp\{a_d\}} \\
\exp\{a_k\} &= \sum_{d=1}^{D} w_{kd} \phi_{dn} \\
\phi_{dn} &= \phi_d(\mathbf{x^{(n)}})
\end{aligned}
$$

またこのようにして求められた最適解 $\omega_{dk}^{\star}$ にを用いて (1) ラベルの予測を行えることができる

おまけ（二次の微小量）

ちなみに最急降下法ではなく二次の微小量を用いてニュートンラフソン法を使うことも可能である

(14) をさらに ω_{st} で微分すると

$$
\begin{aligned}
\frac{\partial^2 H}{\partial \omega_{st} \partial \omega_{kd}} &= \sum_{n=1}^{N} \frac{\partial P_k^{(n)}}{\partial \omega_{st}} \phi_{dn} \\
&= \sum_{n=1}^{N} \sum_{l=1}^{K} \frac{\partial P_k^{(n)}}{\partial a_l^{(n)}} \frac{\partial a_l^{(n)}}{\partial \omega_{st}} \phi_{dn} \\
&= \sum_{n=1}^{N} \sum_{l=1}^{K} P_k^{(n)} \left\{ \delta_{kl} - P_l^{(n)} \right\} \delta_{sl} \phi_{tn} \phi_{dn} \\
&= \sum_{n=1}^{N} P_k^{(n)} \left\{ \delta_{ks} - P_s^{(n)} \right\} \phi_{tn} \phi_{dn}
\end{aligned}
$$

と得られるため