1 Introduction

(省略)

2 The Structure of a CART Model

CART モデルとは、p 次元入力変数 $\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}$ に対し、目的変数 y の状態分布のパラメータ θ を対応する決定木の末端ノードから求める、という構造をしている。
末端ノードを b 個持つ二分木を T とし、それぞれのパラメータの集合を $\Theta = (\theta_1, ...\theta_b)$ と置く。二分木 T は入力変数 $\boldsymbol{x}$ を内部ノードで $\{x_i < s\}$ or $\{x_i \geq s\}$ や $\{x_i \in C\}$ or $\{x_i \notin C\}$ のような条件で分岐させていき、末端ノードへと割り当てる。$\boldsymbol{x}$ が対応する末端ノードが θ_i である時、y の事後分布は θ_i で特徴付けられる分布 $f(y|\theta_i)$ によって

$$p(y|\boldsymbol{x}) = f(y|\theta_i)$$

と表される。f は回帰問題であれば $\mathcal{N}(y|\theta_i, \sigma^2)$、分類問題であれば $Ber(\theta_i)$ などが用いられる。
i 番目の末端ノードに割り当てられるデータのうち j 番目を $y_{i,j}$ $(i = 1, ..., b, j = 1, ..., n_i)$ とし、

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_b \end{pmatrix},\ Y_i = \begin{pmatrix} y_{i,1} \\ \vdots \\ y_{i,n_i} \end{pmatrix}$$

と定義すれば、Θ と T が与えられたときのデータ n 個に対する尤度は

$$\begin{aligned} p(Y|X, \Theta, T) &= \Pi_{i=1}^{b} f(Y_i|\theta_i) \\ &= \Pi_{i=1}^{b} \Pi_{j=1}^{n_i} f(y_{i,j}|\theta_i) \quad (\sum_{i=1}^{b} n_i = n) \end{aligned} \tag{1}$$

と書ける (f は有名分布なので求めるのが簡単)。木の事後分布とは

$$\begin{aligned} p(T|X, Y) &= p(T)p(Y|X, T) \\ &= p(T) \int p(Y|X, \Theta, T)p(\Theta|T)d\Theta \end{aligned} \tag{2}$$

と書けるため、あとは木の事前分布 $p(T)$、T が与えられたときのパラメータの事前分布 $p(\Theta|T)$ を設定すれば良い。

3 Specification of the Tree Prior $p(T)$

$p(T)$ とは木の事前分布というより、木を構成する確率過程である。すなわち、

1. T を一つの末端ノード η のみから成る木とする。
2. 確率 $p_{SPLIT}(\eta, T)$ で末端ノード η を分割する。
3. 2. でノード η が分割された場合、分割ルール ρ を確率 $p_{RULE}(\rho|\eta, T)$ で割り当てる。新たにこれを T とし、左右の子ノードをそれぞれ新しく η として 2,3 を行う。

続いて $p_{SPLIT},\ p_{RULE}$ の適切な事前分布を設定することを考える。

3.1 Determination of Tree Size and Shape by p_{SPLIT}

例えば $p_{SPLIT}(\eta, T)$ を定数 α と置くと、末端ノードの数が b と成る確率は単に

$$\alpha^{b-1}(1-\alpha)^b$$

となる。これは、末端ノードの数が b の二分木の内部ノードの数が $b-1$ となることから明らか。一般に、木の複雑さを抑えるため、次のような p_{SPLIT} が用いられる。ノードの深さを d_η と置くと、

$$p_{SPLIT}(\eta, T) = \alpha(1+d_\eta)^{-\beta}$$

α, β を様々に変えた時の p_{SPLIT} の挙動が論文 5 ページの Figure3 に描かれている。

3.2 Specification of the Splitting Rule Assignment by p_{RULE}

分割ルールは「p 通りの x_i のうちどれを選ぶか」「閾値をどう設定するか」に規定される。p_{RULE} の事前分布には、x_i を $1/p$ の確率で選び、観測された x_i の値域のなかからランダムに閾値を選ぶ」こと (uniform specification of p_{RULE}) が推奨される。

4 Specification of the Parameter Prior $p(\Theta|T)$

(2) 式の積分が解析的に解けるように $p(\Theta|T)$ を設定したい。

4.1 Parameter Priors for Regression Trees

ベイズ混合モデルを次のように設定する。

$$\begin{cases} y_{i,1},...,y_{i,n_i}|\mu_i : iid \sim \mathcal{N}(\mu_i, \sigma^2),\ i=1,...,b \\ \mu_1,...,\mu_b|\sigma^2, T : iid \sim \mathcal{N}(\bar{\mu}, \frac{\sigma^2}{a}) \\ \sigma^2|T \sim IG(\frac{\nu}{2}, \frac{\nu\lambda}{2}) \end{cases}$$

すると、木 T についてのデータ n 個の尤度は

$$\begin{aligned} p(Y|X,T) &= \int p(Y|X,\Theta,T)p(\Theta|T)d\Theta \\ &= \int [\ \Pi_{i=1}^{b}\ [\Pi_{j=1}^{n_i}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{(y_{i,j}-\mu_i)^2}{2\sigma^2}\}] \\ &\qquad \times\sqrt{\frac{a}{2\pi\sigma^2}}\exp\{-\frac{a(\mu_i-\bar{\mu})^2}{2\sigma^2}\}\ ] \\ &\quad \times\frac{(\nu\lambda/2)^{\nu/2}}{\Gamma(\frac{\nu}{2})}(\sigma^2)^{-(\frac{\nu}{2}+1)}\exp(-\frac{\nu\lambda}{2\sigma^2})d\Theta \\ &= \int \Pi_{i=1}^{b}\int a^{1/2}(2\pi\sigma^2)^{-\frac{n_i+1}{2}}\exp[-\frac{1}{2\sigma^2}\{a(\mu_i-\bar{\mu})^2+\sum_{j=1}^{n_i}(y_{i,j}-\mu_i)^2\}]d\mu_i \\ &\quad \times\frac{(\nu\lambda/2)^{\nu/2}}{\Gamma(\frac{\nu}{2})}(\sigma^2)^{-(\frac{\nu}{2}+1)}\exp(-\frac{\nu\lambda}{2\sigma^2})d\sigma^2 \end{aligned}$$

μ_i についての積分計算を求める。指数部分の中身を $(*)$ と置くと、

$$(*) = (n_i+a)\mu_i^2 - 2(a\bar{\mu}+\sum_{j=1}^{n_i}y_{i,j})\mu_i + \sum_{j=1}^{n_i}y_{i,j}^2 + a\bar{\mu}^2$$

ここで、

$$\begin{aligned} \bar{y}_i &= \frac{1}{n_i}\sum_{j=1}^{n_i}y_{i,j} \\ s_i &= \sum_{j=1}^{n_i}(y_{i,j}-\bar{y})^2 \\ &= \sum_{j=1}^{n_i}y_{i,j}^2 - n_i\bar{y}^2 \end{aligned}$$

と置けば、

$$\sum_{j=1}^{n_i} y_{i,j} = n_i \bar{y}_i, \quad \sum_{j=1}^{n_i} y_{i,j}^2 = s_i + n_i \bar{y}_i^2$$

なので、

$$\begin{aligned}
(*) &= (n_i + a)\mu_i^2 - 2(a\bar{\mu} + n_i\bar{y}_i)\mu_i + s_i + a\bar{\mu}^2 + n_i\bar{y}_i^2 \\
&= (n_i + a)(\mu_i - \frac{a\bar{\mu} + n_i\bar{y}_i}{n_i + a})^2 - \frac{(a\bar{\mu} + n_i\bar{y}_i)^2}{n_i + a} + s_i + a\bar{\mu}^2 + n_i\bar{y}_i^2 \\
&= (n_i + a)(\mu_i - \frac{a\bar{\mu} + n_i\bar{y}_i}{n_i + a})^2 + \frac{n_i a}{n_i + a}(\bar{y}_i - \bar{\mu})^2 + s_i
\end{aligned}$$

ここで、

$$t_i = \frac{n_i a}{n_i + a}(\bar{y}_i - \bar{\mu})^2$$

と置けば、

$$(*) = (n_i + a)(\mu_i - \frac{a\bar{\mu} + n_i\bar{y}_i}{n_i + a})^2 + t_i + s_i$$

ゆえに、μ_i についての積分部分は

$$\begin{aligned}
& a^{1/2}(2\pi\sigma^2)^{-\frac{n_i+1}{2}} \exp\{-\frac{1}{2\sigma^2}(t_i + s_i)\} \int \exp\{-\frac{(n_i + a)}{2\sigma^2}(\mu_i - \frac{a\bar{\mu} + n_i\bar{y}_i}{n_i + a})^2\} d\mu_i \\
= \;& a^{1/2}(2\pi\sigma^2)^{-\frac{n_i+1}{2}} \exp\{-\frac{1}{2\sigma^2}(t_i + s_i)\}(\frac{2\pi\sigma^2}{n_i + a})^{1/2} \\
= \;& (2\pi\sigma^2)^{-n_i/2} \frac{a^{1/2}}{(n_i + a)^{1/2}} \exp\{-\frac{1}{2\sigma^2}(t_i + s_i)\}
\end{aligned}$$

したがって、

$$
\begin{aligned}
p(Y|X,T) &= \int \Pi_{i=1}^{b}(2\pi\sigma^2)^{-n_i/2}\frac{a^{1/2}}{(n_i+a)^{1/2}}\exp\{-\frac{1}{2\sigma^2}(t_i+s_i)\} \\
&\quad \frac{(\nu\lambda/2)^{\nu/2}}{\Gamma(\frac{\nu}{2})}(\sigma^2)^{-(\frac{\nu}{2}+1)}\exp(-\frac{\nu\lambda}{2\sigma^2})d\Theta \\
&= \int (2\pi\sigma^2)^{-n/2}\frac{a^{b/2}}{\Pi_{i=1}^{b}(n_i+a)^{1/2}}\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{b}(t_i+s_i)\} \\
&\quad \frac{(\nu\lambda/2)^{\nu/2}}{\Gamma(\frac{\nu}{2})}(\sigma^2)^{-(\frac{\nu}{2}+1)}\exp(-\frac{\nu\lambda}{2\sigma^2})d\Theta \\
&= \frac{ca^{b/2}}{\Pi_{i=1}^{b}(n_i+a)^{1/2}}\int (\sigma^2)^{-(\frac{n+\nu}{2}+1)}\exp[-\frac{1}{2\sigma^2}\{\nu\lambda+\sum_{i=1}^{b}(t_i+s_i)\}d\sigma^2 \\
&= \frac{ca^{b/2}}{\Pi_{i=1}^{b}(n_i+a)^{1/2}}(\sum_{i=1}^{b}(s_i+t_i)+\nu\lambda)^{-(n+\nu)/2}
\end{aligned}
$$

ただし、c は正規化定数。対数を取って、

$$
\log p(Y|X,T) = \frac{b}{2}\log a - \frac{1}{2}\sum_{i=1}^{b}\log(n_i+a) - \frac{n+\nu}{2}\log\{\sum_{i=1}^{b}(s_i+t_i)+\nu\lambda\} + const
$$

また MH アルゴリズムによって T が得られれば、Θ の事後分布は

$$
\begin{aligned}
p(\Theta|T,Y) &\propto \Pi_{i=1}^{b}\ [\Pi_{j=1}^{n_i}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{(y_{i,j}-\mu_i)^2}{2\sigma^2}\}]\sqrt{\frac{a}{2\pi\sigma^2}}\exp\{-\frac{a(\mu_i-\bar{\mu})^2}{2\pi\sigma^2}\} \\
&\propto \Pi_{i=1}^{b}\sqrt{a}(2\pi\sigma^2)^{-\frac{n_i+1}{2}}\exp\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n_i}(y_{i,j}-\mu_i)^2-\frac{a}{2\sigma^2}(\mu_i-\bar{\mu})^2\}
\end{aligned}
$$

指数関数の中身を μ_i で平方完成する。

$$
\begin{aligned}
& -\frac{1}{2\sigma^2}\sum_{j=1}^{n_i}(y_{i,j}-\mu_i)^2 - \frac{a}{2\sigma^2}(\mu_i-\bar{\mu})^2 \\
= & -\frac{1}{2\sigma^2}\sum_{j=1}^{n_i} y_{i,j}^2 + \frac{\mu_i}{\sigma^2}\sum_{j=1}^{n_i} y_{i,j} - \frac{n_i\mu_i^2}{2\sigma^2} - \frac{a\mu_i^2}{2\sigma^2} + \frac{a\bar{\mu}}{\sigma^2}\mu_i + \frac{a}{2\sigma^2}\bar{\mu}^2 \\
= & -\frac{1}{2\sigma^2}\{(n_i+a)\mu_i^2 - 2(n_i+a\bar{\mu})\mu_i\} + const \\
= & -\frac{1}{2\sigma^2}\{\mu_i^2 - \frac{2(n_i\bar{y}+a\bar{\mu})}{n_i+a}\mu_i\} + const \\
= & -\frac{n_i+a}{2\sigma^2}(\mu_i - \frac{n_i\bar{y}+a\bar{\mu}}{n_i+a})^2 + const
\end{aligned}
$$

したがって

$$\mu_i|T,Y \sim \mathcal{N}(\frac{n_i\bar{y}}{n_i+a}, \frac{\sigma^2}{n_i+a})$$

となる。

4.2 Parameter Priors for Classification Trees

二クラス分類を考える (多クラス分類への拡張は容易)。ベイズ混合モデルを次のように設定する。

$$
\begin{cases}
y_{i,1},...,y_{i,n_i}|p_i : iid \sim Ber(p_i), & i=1,...,b \\
p_1,...,p_b : iid \sim Beta(\alpha,\beta)
\end{cases}
$$

すると、

$$
\begin{aligned}
p(Y|X,T) &= \int \Pi_{i=1}^b \{\Pi_{j=1}^{n_i} p_i^{y_{i,j}}(1-p_i)^{1-y_{i,j}}\}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p_i^{\alpha-1}(1-p_i)^{\beta-1}d\Theta \\
&= (\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)})^b \ \Pi_{i=1}^b \int p_i^{\sum_{i=1}^{n_i} y_{i,j}+\alpha-1}(1-p_i)^{\sum_{i=1}^{n_i}(1-y_{i,j})+\beta-1}dp_i \\
&= (\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)})^b \ \Pi_{i=1}^b \frac{\Gamma(\sum_{j=1}^{n_i} y_{i,j}+\alpha)\Gamma(\sum_{j=1}^{n_i}(1-y_{i,j})+\beta)}{\Gamma(n_i+\alpha+\beta)}
\end{aligned}
$$

対数をとって、

$$
\begin{aligned}
\log p(Y|X,T) &= b\{\log\Gamma(\alpha+\beta) - \log\Gamma(\alpha) - \log\Gamma(\beta)\} \\
&\quad + \sum_{i=1}^{b}\{\log\Gamma(\sum_{j=1}^{n_i} y_{i,j}+\alpha) + \log(\sum_{j=1}^{n_i}(1-y_{i,j})+\beta) - \log\Gamma(n_i+\alpha+\beta)\}
\end{aligned}
$$

また MH アルゴリズムによって T が得られれば、Θ の事後分布は

$$\begin{aligned} p(\Theta|T,Y) &\propto \Pi_{i=1}^{b}\{\Pi_{i=1}^{n_i} p_i^{y_{i,j}}(1-p_i)^{1-y_{i,j}}\}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p_i^{\alpha-1}(1-p_i)^{\beta-1} \\ &\propto \Pi_{i=1}^{b} p_i^{\sum_{i=1}^{n_i} y_{i,j}+\alpha-1}(1-p_i)^{\sum_{i=1}^{n_i}(1-y_{i,j})+\beta-1} \\ \therefore p_i|T,Y &\sim Beta(\sum_{i=1}^{n_i} y_{i,j}+\alpha, \sum_{i=1}^{n_i}(1-y_{i,j})+\beta) \end{aligned}$$

となる。

5 Stochastic Search of the Posterior

T の事後分布からの確率標本を MH アルゴリズムによって発生する。MH アルゴリズムでは、t 番目までの T のサンプルが得られているとき、$t+1$ 番目のサンプルを次のように発生する。

1. 提案分布 $q(T^t, T^*)$ から候補点 T^* を得る。
2. 候補点 T^* を、確率

$$\alpha(T^t, T^*) = \min\{\frac{q(T^*, T^t)p(Y|X,T^*)p(T^*)}{q(T^t, T^*)p(Y|X,T^t)p(T^t)}, 1\} \tag{3}$$

で受容し、T^{t+1} とする。T^* が棄却された場合は $T^{t+1} = T^t$ とする。

T^t から候補点 T^* をランダムに発生させる $q(T^t, T^*)$ を次のように定義する。$q(T^t, T^*)$ からの確率標本とは、次の 4 ステップのうちいずれかをランダムに選んで実行することとする。

- GROW：末端ノードをランダムに一つ選ぶ。選んだノードを分割し、p_{RULE} に従い分割ルールを割り当てる。
- PRUNE：末端ノードの親ノードを 2 つランダムに選び、それを末端ノードとする。
- CHANGE：内部ノードをランダムに一つ選び、分割ルールを p_{RULE} にしたがって割り当て直す。
- SWAP：共に内部ノードである親子ノードのペアを取り出し、分割ルールを交換する。もし 2 つの子ノードの分割ルールが同一だった場合、まとめて交換する。

このように、提案分布を対称になるように取ることで通常の酔歩連鎖 MH アルゴリズムが実行できる。