第２章　ベイズ機械学習

2-1 確率統計

$Column \sim$ ベイスの工学的応用（カルマンフィルタ）$\sim$

$Column \sim$ グラフィカルモデル（マルコフブランケット）$\sim$

$Column \sim$ 情報理論とエントロピー　$\sim$

2-2 ベイズ推論の基礎

$Column \sim$ 指数型分布族 $\sim$

本章で取り上げる尤度関数、共役事前分布には一般形が存在する。まず、尤度関数については一般的に次のような形のものを想定すると都合がいい。この形で表される分布の族のことを、「指数型分布族」と呼ぶ。

$$p(\vec{x}|\vec{\eta}) = h(\vec{x}) \exp(\vec{\eta}^T \vec{t}(\vec{x}) - a(\vec{\eta}))$$

η は自然パラメータ、$\vec{t}(\vec{x}_n)$ は十分統計量、$h(\vec{x})$ は基底測度、$a(\vec{\eta})$ は対数分配関数と呼ばれている。対数分配関数 $a(\vec{\eta})$ は $p(\vec{x}|\vec{\eta})$ を積分して 1 になるように保証してくれるもので、

$$\int h(\vec{x}) \exp(\vec{\eta}^T \vec{t}(\vec{x}) - a(\vec{\eta}))\ dx = 1$$

$$a(\vec{\eta}) = ln \int h(\vec{x}) \exp(\vec{\eta}^T \vec{t}(\vec{x}))\ dx$$

この指数型分布族に対して都合のいい共役事前分布は次のようなものが知られている。

$$p_\lambda(\vec{\eta}) = h(\vec{\eta}) \exp(\vec{\eta}^T \vec{\lambda_1} - a(\vec{\eta})\lambda_2 - a_c(\vec{\lambda}))$$

この共役事前分布を用いて事後分布を計算すると、

$$p(\vec{\eta}|X) \propto p_\lambda(\vec{\eta}) \prod_{n=1}^{N} p(\vec{x}_n|\vec{\eta})$$

$$= h(\vec{\eta}) \exp(\vec{\eta}^T \vec{\lambda_1} - a(\vec{\eta})\lambda_2 - a_c(\vec{\lambda})) \prod_{n=1}^{N} h(\vec{x}) \exp(\vec{\eta}^T \vec{t}(\vec{x}) - a(\vec{\eta}))$$

$$= h(\vec{\eta}) \exp(\vec{\eta}^T \vec{\lambda_1} - a(\vec{\eta})\lambda_2 - a_c(\vec{\lambda})) \left\{ \prod_{n=1}^{N} h(\vec{x}) \right\} \exp\left(\vec{\eta}^T \sum_{n=1}^{N} (\vec{t}(\vec{x}) - N a(\vec{\eta})) \right)$$

$$\propto h(\vec{\eta})\exp\left(\vec{\eta}^T\left(\vec{\lambda_1}+\sum_{n=1}^{N}\vec{t}(\vec{x}_n)\right)-a(\vec{\eta})(\lambda_2+N)\right)$$

このように事後分布も事前分布と同じ形になる。事後分布のパラメータは、

$$\vec{\lambda_1}=\vec{\lambda_1}+\sum_{n=1}^{N}\vec{t}(\vec{x}_n),\quad \lambda_2=\lambda_2+N$$

予測分布についても同じように指数分布族で表すことができ、

$$p(\vec{x}_*|X)=\int p(\vec{x}_*|\vec{\eta})p(\vec{\eta}|X)d\vec{\eta}$$

$$=\int d\vec{\eta}\ \{h(\vec{x})\exp(\vec{\eta}^T\vec{t}(\vec{x})-a(\vec{\eta}))h_c(\vec{\eta})\exp(\vec{\eta}^T\vec{\lambda_1}-a(\vec{\eta})\lambda_2-a_c(\vec{\lambda}))\}$$

$$=\cdots\cdots\cdots=h(\vec{x}_*)\frac{\exp(a_c(\vec{\lambda_1}+\vec{t}(\vec{x_*}),\lambda_2+1))}{\exp(a_c(\vec{\lambda_1},\lambda_2))}$$

となり、一般的には指数型分布族にはならないことがわかる。ベルヌーイ分布、ガウス分布など、多くの分布が指数型分布族として表せることが知られており、この形で表すことができれば、あとは式（）や（）に代入すれば事後分布、予測分布を計算することができる。

また、計算の都合上、$g(\vec{\eta})=\exp(-a(\vec{\eta}))$ と置き、（）の両辺に関して $\vec{\eta}$ について勾配を取ると、

$$\nabla_{\vec{\eta}}\int h(\vec{x})\exp(\vec{\eta}^T\vec{t}(\vec{x})-a(\vec{\eta}))\ dx=0$$

$$\nabla_{\vec{\eta}}g(\vec{\eta})\int h(\vec{x})\exp(\vec{\eta}^T\vec{t}(\vec{x}))\ dx$$

$$+\int d\vec{x}\ \{h(\vec{x})\exp(\vec{\eta}^T\vec{t}(\vec{x}))\vec{t}(\vec{x})\}=0$$

$$-\frac{1}{g(\vec{\eta})}\nabla_{\vec{\eta}}=g(\vec{\eta})\int d\vec{x}\ \{h(\vec{x})\exp(\vec{\eta}^T\vec{t}(\vec{x}))\vec{t}(\vec{x})\}=E[\vec{t}(\vec{x})]$$

よって次の結果が得られる。

$$-\nabla_{\eta}ln\ g(\vec{\eta})=\nabla_{\vec{\eta}}a(\vec{\eta})=E[\vec{t}(\vec{x})]$$

このことから、対数分配関数 $a(\vec{\eta})$ の $\vec{\eta}$ に関する勾配は十分統計量 $\vec{t}(\vec{x})$ の期待値になる。

また 2 階の偏微分は十分統計量の共分散になることもわかる。

$$\frac{\partial^2 a(\vec{\eta})}{\partial \eta_i \partial \eta_j} = E[\vec{t}_i(\vec{x})\vec{t}_j(\vec{x})] - E[\vec{t}_i(\vec{x})]E[\vec{t}_j(\vec{x})]$$

$$= Cov[\vec{t}_i(\vec{x}), \vec{t}_j(\vec{x})]$$

$Column \sim$ モーメントマッチング（指数関数族と仮定密度フィルタリング）$\sim$

2-3 確率的生成モデル

現実の問題では、データを生成する分布は複雑で 1 つの確率分布で取り扱えるケースは多くない。複数の分布をデータの生成過程を仮定しながら組み合わせて全体のモデル（同時分布）を作り、そこから事後分布、予測分布を計算する手法を「確率的生成モデル」と呼び、確率分布を複数組み合わせてできたモデルを「混合モデル」と呼ぶ。

2-3-1 混合モデルの構築

多峰性をもつデータに関してのクラスタリングを考える。データを表現するためのモデルを構築する要件定義として例えば以下の過程を考える。

1. K 個のクラスタは混合比率 $\pi = (\pi_1, ..., \pi_K)$ で分布上に存在し、π は事前分布 $p(\pi)$ から生成される。
2. それぞれのクラスタ自身の持つパラメータ θ_k が事前分布 $p(\theta_k)$ から生成される。
3. データ点 x_n が K 個ある分布うちのどれかから生成されるとし、x_n に対応するクラスタの割り当てを s_n をする。この s_n は比率 π によって決まるとし、s_n の生成する分布を $p(s_n|\vec{\pi})$ とする。
4. s_n によって選択された k 番目の確率分布 $p(\vec{x_n}|\vec{\theta_k})$ からデータ x_n が生成される。

これら全ての確率分布をデータ生成順に組み合わせ、N 個のデータに関して同時分布を考えると以下のようになる。

$$p(X, S, \Theta, \vec{\pi}) = p(X|S, \Theta)p(S|\vec{\pi})p(\Theta)p(\vec{\pi}) = \{\prod_{n=1}^{N} p(\vec{x_n}|\vec{s_n}, \Theta)p(\vec{s_n}|\vec{\pi})\}\{\prod_{k=1}^{K} p(\vec{\theta_k})\}p(\vec{\pi})$$

実際に問題を解く際には、$p(X|S, \Theta)$,$p(\Theta)$ は問題設定に応じて決め、（クラスタリングの場合は）s_n をサンプリングする分布として以下のカテゴリ分布、

$$p(\vec{s_n}|\vec{\pi}) = Cat(\vec{s_n}|\vec{\pi}) = \prod_{k=1}^{K} \pi_k^{s_{n,k}}$$

π をサンプリングする分布としてカテゴリ分布の共役事前分布である $Dirichlet$ 分布を選ぶことが多い。

$$p(\vec{\pi}) = Dir(\vec{\pi}|\vec{\alpha})$$

また s_n は直接は観測されないが、x_n を生成する K 個の分布のうち 1 つを選択するという意味で、x_n を発生させる確率分布を潜在的に決めている確率変数であると言える。このため s_n は潜在変数と呼ばれている。

2-3-2 混合モデルの推論

この同時分布から事後分布 $p(S, \Theta, \vec{\pi}|X)$, クラスタ S の推定 $p(S|X)$ が可能であるが、いずれの計算も

$$p(X) = \sum_S \iint p(X, S, \Theta, \pi) d\Theta d\pi = \sum_s p(X, S)$$

$$p(S|X) = \iint p(S, \Theta, \pi|X) d\Theta d\pi$$

の計算が発生してしまい、解析的に解くことがほぼ不可能になる。次章で、この問題をある程度解消して近似的に解を出す方法を説明する。

$Column \sim$ ノンパラメトリックモデル（カーネル密度推定と最近傍法）$\sim$

$$P = \frac{K}{N}$$

$$P \approx V p(x)$$

よって上式より

$$p(x) \approx \frac{K}{NV}$$

ここで２つのアプローチが考えられる。

1.Vを先に決めて K を数える。（カーネル密度推定）

2.K を与えて、Vを決定する。（最近傍法、K 近傍法）

1. カーネル密度推定

$$V = h^D$$

$$W(u) = \begin{cases} 1 & |u_i| \leq 1/2 \\ 0 & else \end{cases}$$

$$K = \sum_{i=1}^{n} W\Big[\frac{x - x_i}{h}\Big]$$

よって推定密度は以下のようになる。

$$p(x) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{h^D} W\Big[\frac{x - x_i}{h}\Big]$$

一般的にはこのように立方体の中にデータ点をいれるのではなく、中心がデータ点 x_n にある N 個の立方体を重ね合わせるという解釈する。

このカーネルは立方体の縁で不連続になるので、ガウスカーネルを用いて以下のような確率密度モデルを考えることが多い。（中心がデータ点 x_n にある N 個のガウス分布を重ね合わせる。）

$$p(x) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{(2\pi h^2)^{D/2}} \exp\Big(-\frac{||x - x_n||^2}{2h^2}\Big)$$

2. 最近傍法

今度は先にデータ数 K を決めて、それに応じて Vを決定する。ここでは立方体ではなく d 次元の球（超球）を考える。

超球の体積は

$$V = \frac{\pi^{\frac{d}{2}} r^d}{\Gamma(\frac{d}{2} + 1)}$$

よって（）より確率密度は、

$$p(x) = \frac{K\Gamma(\frac{d}{2} + 1)}{N_k \pi^{\frac{d}{2}} r^d}$$

ここで $K = 1$, 最も近いデータ点までの距離を Δ とすると、

$$p(x|C_k) = \frac{\Gamma(\frac{d}{2} + 1)}{N\pi^{\frac{d}{2}} \Delta^d}$$

これにベイスの定理を適用してクラスの事後分布を求めると、

$$p(C_k|x) = \frac{p(x|C_k)p(C_k)}{p(x)} \propto p(x|C_k)p(C_k)$$

$$\approx \frac{\Gamma(\frac{d}{2}+1)}{N_k \pi^{\frac{d}{2}} \Delta^d} \frac{N_k}{N} \propto \frac{1}{\Delta^d}$$

よって、データ点からの距離が近いほどクラスの事後分布が大きくなる。このことから、データ点から一番近いデータのクラスを選択すればいいことがわかる。よってこの手法を最近傍法と呼ぶ。これを一般の K に拡張したものが K 近傍法である。（PRML 上巻第 2 章参照）

$Column \sim \sim$

2-4 **近似推論**

$Column \sim$ **分配関数と統計物理** $\sim$

$Column \sim$ **いろいろな近似推論** $\sim$

2-5 **ガウス混合モデルと教師なし学習**

$Column \sim$ **解析力学と汎関数（変分法）** $\sim$

$Column \sim$ **線形次元削減と変分自己符号器（VAE）** $\sim$