第２章　ベイズ機械学習

2-1 確率統計の基礎

確率と確率変数
確率分布
周辺確率
期待値、分散
独立性
条件付き確率
ベイズの定理
確率密度関数
有名な確率分布

$Column \sim$ ベイズの工学的応用（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）$\sim$

大きなグラフィカルモデルが与えられたときに、注目したノードを独立 (条件付き独立) にする最小単位を決めるときに便利な道具が、「マルコフブランケット」と呼ばれるものである。

$Column \sim$ 情報理論とエントロピー　$\sim$

2-2 ベイズ推論の基礎

2-2-1 ベイズ推論における学習、予測

ベイス推論では、パラメータ（重み）も不確実性をもつ確率変数として捉え、次の手順で問題を解くことが多い。

1. 問題に合わせて、適切な尤度関数 $p(D|\vec{w})$ を設定する。（モデル化）
2. 尤度関数にあった共役事前分布 $p(\vec{w})$ を選ぶ。
3. ベイスの定理を用いて、事後分布 $p(\vec{w}|D)$ を解析的に求める。（学習）
4. 事後分布を用いて、予測分布 $p(x_*|D)$ を計算する。（予測）

$Step1.$　**尤度関数 $p(D|\vec{x})$ の設計（モデル化）**

解きたい問題に対して可視化などを行い、ある程度妥当であると思われる尤度関数を設定する。線形回帰など一般的にはガウス分布を用いることが多いが、カウントデータ（非負）ならポアソン分布、周期性をもつ分布はフォンミーゼス分布（PRML2 章）など、データの分布に合わせた確率分布を選ぶことが望ましい。

*Step*2. **共役事前分布 $p(\vec{w})$ の設定**

設定した尤度関数の共役事前分布 $p(\vec{w})$ を選ぶ。「共役事前分布」はベイズの定理ととても相性がよく、ベイズの定理を適用してもその分布の形が変わらない分布である。大抵は尤度関数と 1 対 1 の関係で決まっているので、この *Step*2. はすぐに終わる。

*Step*3. **学習**

ベイスの定理を用いて以下の事後分布 $p(\vec{w}|D)$ を計算する

$$p(\vec{w}|D) = \frac{p(D|\vec{w})p(\vec{w})}{p(D)}$$

*Step*4. **予測**

未観測のデータ x_* に対して以下の予測分布を計算する。

$$p(x_*|D) = \int p(x_*|\vec{w})p(\vec{w}|D)\ d\vec{w}$$

これは予測に際して必要ない $\vec{w}$ について積分除去を行ったものと考えることができる。また、事後分布とは異なり、一般的には予測分布は共役事前分布の形になるとは限らない。

(1) 尤度関数としてベルヌーイ分布

$$p(x|\mu) = Bern(x|\mu)$$

でモデル化できる問題において、μ の分布を訓練データ x_n から推論せよ。また未観測の値 $x_* \in 0, 1$ に対する予測分布を計算せよ。

(2) 線形回帰 $y_n = \vec{w}^T x_n + \epsilon_n$ についてモデル $p(y_n|\vec{x}_n, \vec{w})$ の構築を行い、事後分布、予測分布を計算せよ。

2-2-2 モデルエビデンス（周辺尤度）

ベイズの定理を変形して、

$$p(D) = \frac{p(D|\vec{w})p(\vec{w})}{p(\vec{w}|D)}$$

と表す。このとき、$p(D)$ を周辺尤度（モデルエビデンス）と呼ぶ。これはモデルのデータ生成確率と解釈することができ、この値を複数のモデル間で比較することで最適なモデルの選択を行うことができる。

$Column \sim$ 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（）に代入するだけで事後分布、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。

（1）あるクラスタ k に対する観測モデルとしてポアソン分布を採用し、混合モデルを構築せよ。

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）$\sim$

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

$$P \approx Vp(x)$$

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

事後分布、周辺尤度、予測分布など問題によっては解析的に解くことが難しいものに関しては、近似的に解を求めることが多い。近似手法は大きく分けると、サンプリング、変分法に大別される。

2-4-1 ギブスサンプリング

分布全体の解析的な把握が難しい場合、期待値等の分布に関する部分的な統計量を解析することは重要である。そのような各種統計量を得たい場合、分布から複数の実現値をサンプリングし、その実現値を元に計算を行うことが有効的である。

$$z_1^{(i)}, z_2^{(i)}, z_3^{(i)} \sim p(z_1, z_2, z_3)$$

混合モデル等、複雑なモデルに関しては全てのサンプルを上記のように同時にサンプルすることは難しいため、ギブスサンプリングという手法を用いて以下のようにサンプリングを行う。

$$z_1^{(i)} \sim p(z_1 | z_2^{(i-1)}, z_3^{(i-1)})$$
$$z_2^{(i)} \sim p(z_2 | z_1^{(i)}, z_3^{(i-1)})$$
$$z_3^{(i)} \sim p(z_3 | z_1^{(i)}, z_2^{(i)})$$

この手法は MCMC (マルコフ連鎖モンテカルロ法) の手法の一つに分類されており、サンプル数が十分に多い場合、繰り返しで得られた z_k は真の事後分布から得られたものであると理論的に保証されている。($Column$ 参照)

(1) ギブスサンプリングを用いて、() で求めたポアソン混合モデルの事後分布 $p(S, \vec{\lambda}, \vec{\pi}|X)$ からサンプリングを行うアルゴリズムを導け。混合分布では以下のように、潜在変数とパラメータを次のように分けてサンプリングすると簡単な確率分布が得られることが知られている。

$$S \sim p(S|X, \vec{\lambda}, \vec{\pi}), \quad \vec{\lambda},\ \vec{\pi} \sim p(\vec{\lambda}, \vec{\pi}|X, S)$$

2-4-2 平均場近似 (変分推論)

複雑な分布を最適化問題を解くことによってより簡単な近似分布で表現する手法を「変分推論」、「変分近似」と呼ぶ。事後分布は解析的に解けなくなる状況に陥ることがあるため、確率変数に特定の制約を付けた上で事後分布を近似する。

最適化には KL ダイバージェンスを使い、最小化問題として以下のように定式化される。

$$q_{opt} = argmin_q KL[q(z_1, z_2, z_3)|\ p(z_1, z_2, z_3)]$$

ここで、解が $q_{opt}(z_1, z_2, z_3) = p(z_1, z_2, z_3)$ とならないように q に制約をつける手法として、各確率変数に独立性の仮定をおく。

$$p(z_1, z_2, z_3) \approx q(z_1)q(z_2)q(z_3)$$

これを「平均場近似」と呼ぶ。

(2) 平均場近似を用いて、() で求めたポアソン混合モデルの変分推論アルゴリズムを導出せよ。ただし、事後分布 $p(S, \vec{\lambda}, \vec{\pi}|X)$ の潜在変数とパラメータを以下のように分けて近似せよ。

$$p(S, \vec{\lambda}, \vec{\pi}|X) \approx q(S)q(\vec{\lambda}, \vec{\pi})$$

$Column \sim$ MCMC とその他サンプリング手法 $\sim$

$Column \sim$ 周辺尤度とエビデンス下界（ELBO）$\sim$

2-5 ガウス混合モデルと教師なし学習

2-5-1 K-means アルゴリズム

クラスタリング (教師なし学習) によく使われる手法に「K-means アルゴリズム」がある。K-means アルゴリズムでは以下の目的関数を最小化する。

$$J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}||\vec{x}_n - \vec{\mu}_n||^2$$

$\vec{\mu}_k$ について解くと、

$$\vec{\mu}_k = \frac{\sum_n r_{nk}\vec{x}_n}{\sum_n r_{nk}}$$

この式の分母は k 番目のクラスタに割り当てられたデータの数に等しいので、$\vec{\mu}_k$ は、k 番目のクラスタに割り当てられた全てのデータ点 $\vec{x_n}$ の平均となっている。これが K-means アルゴリズムと呼ばれている理由である。なお、K-means アルゴリズムは次節の混合ガウス分布に対する EM アルゴリズムの非確率的極限となっている。

2-5-2 EM アルゴリズム (Expectation-Maximization Algorithm)

潜在変数が含まれる最尤推定の問題で使われる最適化アルゴリズム。

$$\ln p(X|\pi,\mu,\Sigma) = \sum_{n=1}^{N} \ln\Big\{\sum_{n=1}^{N} \pi_k N(x_n|\mu_k,\Sigma_k)\Big\}$$

混合ガウス分布で表された尤度関数 () の最大化は、EM アルゴリズムを用いると以下の 4 ステップによって最適化される。

1. μ_k, Σ_k, π_k を初期化し、対数尤度 () の初期値を計算する。
2. (E ステップ) 1 の値を用いて以下の負担率を計算する。

$$\gamma_k = \frac{\pi_k N(x_n|\mu_k,\Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n|\mu_j,\Sigma_j)}$$

3. (M ステップ) 2 で求めた負担率を用いて、次式で μ_k, Σ_k, π_k を再計算する。

$$\mu_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \vec{x}_n$$

$$\Sigma_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\vec{x}_n - \mu_k^{new})(x_n - \mu_k^{new})^T$$

$$\pi_k^{new} = \frac{N_k}{N}, \quad N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

4. $\mu_k^{new}, \Sigma_k^{new}, \pi_k^{new}$ で対数尤度 () を計算。対数尤度、もしくはパラメータの値の変化を見て収束性を確認。収束していなければ、2 に戻る。

上記の更新ステップで、対数尤度関数は必ず増加することが保証されている。(Column)

$Column \sim$ 対数尤度関数の最大化と KL-divergence $\sim$

$Column \sim$ 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）$\sim$