ベイズ推測（ベイズ推論）
Bayesian inference

正田 備也

masada@rikkyo.ac.jp

Contents

統計的推測 statistical inference

- 観測データとして、d 次元ユークリッド空間 $\mathbb{R}^d$ 上の n 個の点の集合 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ が与えられているとする
- 観測データをひとつの記号で $\mathcal{D}$ と書くことにする
  - つまり、$\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$
- $\boldsymbol{x}_i$ は、独立に同じ分布 $q(\boldsymbol{x})$ にしたがうと考えることにする
  - つまり、$q(\mathcal{D}) = \prod_{i=1}^n q(\boldsymbol{x}_i)$
- しかし、この分布 $q(\boldsymbol{x})$ を直接知る方法はない
- そこで、$\mathcal{D}$ から $q(\boldsymbol{x})$ を推測することを、統計的推測ないし統計的学習という

確率モデル probabilistic model

- 真の分布を推測するとき、私たちは、あるパラメータ $\boldsymbol{\eta}$ をもつ確率分布 $p(\boldsymbol{x}|\boldsymbol{\eta})$ を準備する
  - $p(\boldsymbol{x}|\boldsymbol{\eta})$ を確率モデルと呼ぶ
- また、パラメータを決めること自体にも不確かさがあるとする場合、パラメータがしたがう事前分布 $p(\boldsymbol{\eta})$ も準備する
  - 事前分布のパラメータ（ハイパーパラメータ）はまだ書かずにおく
- このとき事後分布は以下のように書ける

$$p(\boldsymbol{\eta}|\mathcal{D}) = \frac{1}{Z_n} p(\boldsymbol{\eta}) p(\mathcal{D}|\boldsymbol{\eta}) = \frac{1}{Z_n} p(\boldsymbol{\eta}) \prod_{i=1}^{n} p(\boldsymbol{x}_i|\boldsymbol{\eta}) \qquad (1)$$

- Z_n については次のスライドで

周辺尤度

- 式 (1) の事後分布の規格化定数 Z_n を分配関数とも呼ぶ
  - 分配関数 partition function は物理方面から来ている用語
- Z_n は規格化定数なので、$Z_n = \int p(\boldsymbol{\eta})p(\mathcal{D}|\boldsymbol{\eta})d\boldsymbol{\eta}$ を満たす
- つまり、$Z_n = p(\mathcal{D})$
- $p(\mathcal{D})$ を、周辺尤度、もしくは、エビデンス evidence と呼ぶ
  - 論文では marginal likelihood より evidence のほうをよく見かける
- すなわち、事後分布は以下のようにも書ける

$$p(\boldsymbol{\eta}|\mathcal{D}) = \frac{p(\boldsymbol{\eta})p(\mathcal{D}|\boldsymbol{\eta})}{p(\mathcal{D})} \tag{2}$$

観測データを生成する分布の推定(1/2)

- 観測データ $\mathcal{D}$ から、それを生成する分布 $\hat{p}(\boldsymbol{x})$ を推定する方法はいろいろある
- 観測データ $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ が与えられているとき、パラメータ $\boldsymbol{\eta}$ の関数 $\prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{\eta})$ を尤度関数と呼ぶ

1. 尤度関数を最大にするパラメータ $\boldsymbol{\eta}_{\mathrm{ML}}$ を最尤推定量といい、$p(\boldsymbol{x}_i|\boldsymbol{\eta}_{\mathrm{ML}})$ を推測の結果 $\hat{p}(\boldsymbol{x})$ とする方法を、最尤推定という
2. 事後分布の最大値を与えるパラメータ $\boldsymbol{\eta}_{\mathrm{MAP}}$ を事後確率最大化推定量といい、$p(\boldsymbol{x}_i|\boldsymbol{\eta}_{\mathrm{MAP}})$ を推測の結果 $\hat{p}(\boldsymbol{x})$ とする方法を、事後確率最大化推定（もしくはMAP推定）という

観測データを生成する分布の推定(2/2)

3. ベイズ的なモデリングの課題は事後分布を求めることだったが、事後分布を使うと、以下で定義する予測分布 $p(\boldsymbol{x}|\mathcal{D})$ をもって推測の結果 $\hat{p}(\boldsymbol{x})$ とすることができる

- 事後分布 $p(\boldsymbol{\eta}|\mathcal{D})$ によって確率モデル $p(\boldsymbol{x}|\boldsymbol{\eta})$ を平均化したもの、つまり、$p(\boldsymbol{\eta}|\mathcal{D})$ に関する $p(\boldsymbol{x}|\boldsymbol{\eta})$ の期待値を、予測分布と呼ぶ

$$p(\boldsymbol{x}|\mathcal{D}) = \int p(\boldsymbol{x}|\boldsymbol{\eta})p(\boldsymbol{\eta}|\mathcal{D})d\boldsymbol{\eta} \tag{3}$$

ベイズ推測

- ベイズ推測 Bayesian inference とは、「真の分布は、おおよそ予測分布だろう」と推測することである
  - cf. 渡辺澄夫『ベイズ統計の理論と方法』コロナ社、5 頁
- 以下、正規分布と多項分布の場合について、共役事前分布を使ったときに予測分布がどのような分布になるかを説明する

Contents

単変量正規分布を使ったベイズ的モデリング

- 観測データ $\mathcal{D} = \{x_1, \ldots, x_N\}$ が独立に同じ正規分布 $\mathcal{N}(\mu, \tau^{-1})$ にしたがうと仮定
- 平均 μ と精度 τ の事前分布として正規ガンマ分布 $\mathrm{NG}(\mu, \tau; \mu_0, \lambda_0, \alpha, \beta)$ を使う
- 正規ガンマ分布の確率密度関数は

$$
\begin{aligned}
p(\mu, \tau; \mu_0, \lambda_0, \alpha, \beta) &= p(\mu|\tau; \mu_0, \lambda_0, \alpha, \beta) p(\tau; \alpha, \beta) \\
&= \frac{\beta^\alpha \sqrt{\lambda_0}}{\Gamma(\alpha)\sqrt{2\pi}} \tau^{\alpha - \frac{1}{2}} e^{-\beta\tau} e^{-\frac{\lambda_0 \tau (\mu - \mu_0)^2}{2}} \qquad (4)
\end{aligned}
$$

共役事前分布としての正規ガンマ分布

- 正規ガンマ分布は共役事前分布
- よって事後分布も正規ガンマ分布となる
- 事後分布 $p(\mu, \tau | \mathcal{D}; \mu_0, \lambda_0, \alpha, \beta)$ は以下のように書ける
  - 前回の講義資料を参照

$$
\begin{aligned}
&p(\mu, \tau | \mathcal{D}; \mu_0, \lambda_0, \alpha, \beta) \\
&\propto \tau^{\alpha + \frac{N}{2} - \frac{1}{2}} \exp\left[- \tau \left(\beta + \frac{Ns}{2} + \frac{\lambda_0 N (\bar{x} - \mu_0)^2}{2(\lambda_0 + N)} \right) \right] \\
&\quad \times \exp\left[- \frac{\tau}{2} (\lambda_0 + N) \left(\mu - \frac{\lambda_0 \mu_0 + N \bar{x}}{\lambda_0 + N} \right)^2 \right]
\end{aligned} \tag{5}
$$

予測分布

- 上記の設定のもとでの予測分布 $p(x|\mathcal{D})$ は $p(x|\mathcal{D}) = \int p(x|\mu,\tau)p(\mu,\tau|\mathcal{D})d\mu d\tau$ によって求められる
- 事後分布は、$p(\mu,\tau|\mathcal{D}) = p(\mu|\tau,\mathcal{D})p(\tau|\mathcal{D})$ と、正規分布とガンマ分布の積で書ける（正規ガンマ分布だから）。よって

$$p(x|\mathcal{D}) = \int p(x|\mu,\tau)p(\mu|\tau,\mathcal{D})p(\tau|\mathcal{D})d\mu d\tau \tag{6}$$

  - $p(x|\mu,\tau) \propto \exp[-\frac{\tau}{2}(x-\mu)^2]$
  - $p(\mu|\tau,\mathcal{D}) \propto \exp[-\frac{\tau(\lambda_0+N)}{2}(\mu - \frac{\lambda_0\mu_0+N\bar{x}}{\lambda_0+N})^2]$
  - $p(\tau|\mathcal{D}) \propto \tau^{\alpha+\frac{N}{2}-1}\exp[-\tau(\beta+\frac{Ns}{2}+\frac{\lambda_0 N(\bar{x}-\mu_0)^2}{2(\lambda_0+N)})]$
- この予測分布がどういう分布になるかを、以下で示す

t-分布 Student's t-distribution

- $X_1, \ldots, X_n$ を平均 μ、標準偏差 σ の正規分布に独立にしたがう確率変数とする
- 標本平均を $\bar{X} = \frac{\sum_i X_i}{n}$、不偏分散を $S^2 = \frac{\sum_i (X_i - \bar{X})^2}{n-1}$ とする
- このとき、$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$ と定義される値は、自由度 $\nu = n - 1$ の t-分布にしたがう
- t-分布の確率密度関数は、以下のようになる

$$p(t; \nu) = \frac{\Gamma((\nu + 1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)} (1 + t^2/\nu)^{-(\nu+1)/2} \tag{7}$$

t location-scale distribution

- 平均が μ、スケールが σ^2、自由度が ν の t location-scale distribution の確率密度関数は、以下のとおり

$$p(x;\mu,\sigma^2) = \frac{\Gamma((\nu+1)/2)}{\sigma\sqrt{\nu\pi}\Gamma(\nu/2)}\left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2} \tag{8}$$

- $\mu=0$、$\sigma=1$ のとき、t-分布に一致する
  - 英語版 Wikipedia「t-分布」の「4.2 Bayesian Inference」を参照
  - MATLAB の tLocationScaleDistribution の項も参照
- 以下では、予測分布がこの t location-scale distribution になることを示す

準備として、正規ガンマ分布 NG$(\mu,\tau|\mu_0,\lambda_0,\alpha,\beta)$ において $\alpha=\beta=\frac{\nu}{2}$ である場合、以下が成立することを確認しておく（証明はしない）。(cf. CS340 (Machine learning) Fall 2007 @ UBC)

$$t_\nu(\mu|\mu_0,\lambda_0^{-1}) = \int_0^\infty \mathrm{NG}(\mu,\tau|\mu_0,\lambda_0,\frac{\nu}{2},\frac{\nu}{2})d\tau = \int_0^\infty \mathcal{N}(\mu|\mu_0,(\lambda_0\tau)^{-1})\mathrm{Ga}(\tau|\frac{\nu}{2},\frac{\nu}{2})d\tau \tag{9}$$

ただし、$t_\nu(x|\mu,\sigma^2)$ は、平均 μ、スケール σ^2、自由度 ν の t location-scale distribution である。
$x\sim\mathrm{Ga}(\alpha,\beta)$ のとき、$cx\sim\mathrm{Ga}(\alpha,\frac{\beta}{c})$ となる。よって

$$\begin{aligned}
\int_0^\infty \mathcal{N}(\mu|\mu_0,(\lambda_0\tau)^{-1})\mathrm{Ga}(\tau|\alpha,\beta)d\tau &= \int_0^\infty \mathcal{N}(\mu|\mu_0,(\lambda_0\tau)^{-1})\mathrm{Ga}(\tau|\alpha,\frac{\beta}{\alpha}\alpha)d\tau \\
&= \int_0^\infty \mathcal{N}(\mu|\mu_0,(\lambda_0\tau)^{-1})\mathrm{Ga}(\frac{\alpha}{\beta}\tau|\alpha,\alpha)d\tau \\
&= \int_0^\infty \mathcal{N}(\mu|\mu_0,(\frac{\beta}{\alpha}\lambda_0\tau')^{-1})\mathrm{Ga}(\tau'|\alpha,\alpha)d\tau' \\
&= t_{2\alpha}(\mu|\mu_0,\frac{\beta}{\alpha\lambda_0})
\end{aligned} \tag{10}$$

もうひとつの準備として、異なる二つの単変量正規分布の密度関数の積について、以下の結果を確認しておく。

この結果は、このブログ記事において提示されているものだが、The Matrix Cookbook の 7.2.6. で提示されている結果の特殊例でもある。

平均が m で標準偏差が s の単変量正規分布の密度関数を $f(x; m, s)$ と書くことにすると

$$f(x; \mu_1, \sigma_1) f(x; \mu_2, \sigma_2) = f\left(\mu_1; \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right) f(x; \mu, \sigma) \tag{11}$$

ただし、

$$\mu = \frac{\sigma_1^{-2}\mu_1 + \sigma_2^{-2}\mu_2}{\sigma_1^{-2} + \sigma_2^{-2}} \tag{12}$$

$$\sigma^2 = \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2} \tag{13}$$

$p(\mu|\tau,\mathcal{D}) \propto \exp[-\frac{\tau(\lambda_0+N)}{2}(\mu - \frac{\lambda_0\mu_0+N\bar{x}}{\lambda_0+N})^2]$ において $\mu_N = \frac{\lambda_0\mu_0+N\bar{x}}{\lambda_0+N}, \tau_N = \tau(\lambda_0+N)$ とおく。前のスライドの結果を使うべく、$\mu_1 = x, \sigma_1 = \tau^{-1/2}, \mu_2 = \mu_N, \sigma_2 = \tau_N^{-1/2}$ とおくと

$$
\begin{aligned}
p(x|\mu,\tau)p(\mu|\tau,\mathcal{D}) &\propto \exp\left[-\frac{\tau\tau_N}{2(\tau+\tau_N)}(x-\mu_N)^2\right]\exp\left[-\frac{\tau+\tau_N}{2}\Big(\mu - \frac{\tau_N x + \tau\mu_N}{\tau+\tau_N}\Big)^2\right] \\
&= \exp\left[-\frac{\tau(\lambda_0+N)}{2(\lambda_0+N+1)}(x-\mu_N)^2\right]\exp\left[-\frac{\tau(\lambda_0+N+1)}{2}\Big(\mu - \frac{(\lambda_0+N)x+\mu_N}{\lambda_0+N+1}\Big)^2\right]
\end{aligned}
\tag{14}
$$

よって、$\alpha_N = \alpha + \frac{N}{2}, \beta_N = \beta + \frac{Ns}{2} + \frac{\lambda_0 N(\bar{x}-\mu_0)^2}{2(\lambda_0+N)}$、さらに $\lambda_N = \lambda_0 + N$ とおくと、

$$
\begin{aligned}
p(x|\mu,\tau)p(\mu|\tau,\mathcal{D})p(\tau|\mathcal{D}) &\propto \exp\left[-\frac{\tau(\lambda_0+N)}{2(\lambda_0+N+1)}(x-\mu_N)^2\right] \\
&\times \exp\left[-\frac{\tau(\lambda_0+N+1)}{2}\Big(\mu - \frac{(\lambda_0+N)x+\mu_N}{\lambda_0+N+1}\Big)^2\right] \times \tau^{\alpha_N-1}e^{-\tau\beta_N} \\
&= \tau^{\alpha_N-1}e^{-\tau\beta_N}\exp\left[-\frac{\tau\lambda_N}{2(\lambda_N+1)}(x-\mu_N)^2\right] \times \exp\left[-\frac{\tau(\lambda_N+1)}{2}\Big(\mu - \frac{\lambda_N x+\mu_N}{\lambda_N+1}\Big)^2\right]
\end{aligned}
\tag{15}
$$

μ を積分消去すると

$$p(x,\tau|\mathcal{D}) = \int_{-\infty}^{\infty} p(x|\mu,\tau)p(\mu|\tau,\mathcal{D})p(\tau|\mathcal{D})d\mu \propto \tau^{\alpha_N - \frac{1}{2}} e^{-\tau\beta_N} \exp\left[-\frac{\tau\lambda_N}{2(\lambda_N+1)}(x-\mu_N)^2\right] \tag{16}$$

この式は $p(x,\tau|\mathcal{D})$ が正規ガンマ分布の密度関数であることを示している。
そこで、τ を積分消去するために式 (10) の結果を使うと、

$$p(x|\mathcal{D}) = \int p(x,\tau|\mathcal{D})d\tau = t_{2\alpha_N}(x|\mu_N, \frac{\beta_N(\lambda_N+1)}{\alpha_N\lambda_N}) \tag{17}$$

したがって、予測分布は、平均 μ_N、スケール $\frac{\beta_N(\lambda_N+1)}{\alpha_N\lambda_N}$、自由度 $2\alpha_N$ の t location-scale distribution である。

Contents

多項分布を使ったベイズ的モデリング

- 観測データ $\mathcal{D} = \{x_1, \ldots, x_N\}$ が独立に同じカテゴリカル分布 Cat(ϕ) にしたがうと仮定
  - 例えば、$x_{62} =$ “apple” は、62 番目に出現した単語が “apple” だ、という意味
- ϕ の事前分布としてディリクレ分布 Dir($\boldsymbol{\beta}$) を使う
- ディリクレ分布の確率密度関数は

$$p(\boldsymbol{\phi}; \boldsymbol{\beta}) = \frac{\Gamma(\sum_{w=1}^{W} \beta_w)}{\prod_{w=1}^{W} \Gamma(\beta_w)} \prod_{w=1}^{W} \phi_W^{\beta_w - 1} \tag{18}$$

共役事前分布としてのディリクレ分布

- 正規ガンマ分布は共役事前分布
- よって事後分布も正規ガンマ分布となる
- 事後分布 $p(\boldsymbol{\phi}|\mathcal{D};\boldsymbol{\beta})$ は以下のように書ける

$$p(\boldsymbol{\phi}|\mathcal{D};\boldsymbol{\beta}) = \frac{\Gamma(\sum_{w=1}^{W}(c_w+\beta_w))}{\prod_{w=1}^{W}\Gamma(c_w+\beta_w)}\prod_{w=1}^{W}\phi_W^{c_w+\beta_w-1} \tag{19}$$

予測分布

- 多項分布の場合、予測分布についても、多数の単語出現からなる観測データの予測分布を考えることが多い
- その予測分布を求めたい単語列（つまり文書）を $\boldsymbol{x}_0$ とする
- $\boldsymbol{x}_0 = \{x_{0,1}, \ldots, x_{0,n_0}\}$ とする。
- 上記の設定のもとでの予測分布 $p(\boldsymbol{x}_0|\mathcal{D};\boldsymbol{\beta})$ は $p(\boldsymbol{x}_0|\mathcal{D};\boldsymbol{\beta}) = \int p(\boldsymbol{x}_0|\boldsymbol{\phi})p(\boldsymbol{\phi}|\mathcal{D};\boldsymbol{\beta})d\boldsymbol{\phi}$ によって求められる
- この予測分布がどういう分布になるかを、以下で示す

$$
\begin{aligned}
p(\boldsymbol{x}_0|\mathcal{D};\boldsymbol{\beta}) &= \int p(\boldsymbol{x}_0|\boldsymbol{\phi})p(\boldsymbol{\phi}|\mathcal{D};\boldsymbol{\beta})d\boldsymbol{\phi} \\
&= \int \left(\frac{n_0!}{\prod_{w=1}^W c_{0,w}!} \prod_{w=1}^W \phi_w^{c_{0,w}} \times \frac{\Gamma(\sum_{w=1}^W (c_w+\beta_w))}{\prod_{w=1}^W \Gamma(c_w+\beta_w)} \prod_{w=1}^W \phi_W^{c_w+\beta_w-1} \right) d\boldsymbol{\phi} \\
&= \frac{n_0!}{\prod_{w=1}^W c_{0,w}!} \frac{\Gamma(\sum_{w=1}^W (c_w+\beta_w))}{\prod_{w=1}^W \Gamma(c_w+\beta_w)} \int \prod_{w=1}^W \phi_W^{c_w+c_{0,w}+\beta_w-1} d\boldsymbol{\phi} \\
&= \frac{n_0!}{\prod_{w=1}^W c_{0,w}!} \frac{\Gamma(\sum_{w=1}^W (c_w+\beta_w))}{\prod_{w=1}^W \Gamma(c_w+\beta_w)} \frac{\prod_{w=1}^W \Gamma(c_w+c_{0,w}+\beta_w)}{\Gamma(\sum_{w=1}^W (c_w+c_{0,w}+\beta_w))} \\
&= \frac{n_0!\Gamma(\sum_{w=1}^W (c_w+\beta_w))}{\Gamma(\sum_{w=1}^W (c_w+c_{0,w}+\beta_w))} \prod_{w=1}^W \frac{\Gamma(c_w+c_{0,w}+\beta_w)}{c_{0,w}!\Gamma(c_w+\beta_w)} \\
&= \frac{n_0(n_0-1)\cdots 2\cdot 1}{(n+n_0-1+\beta_\Sigma)(n+n_0-2+\beta_\Sigma)\cdots(n+1+\beta_\Sigma)(n+\beta_\Sigma)} \\
&\quad \times \prod_{w=1}^W \frac{(c_w+c_{0,w}-1+\beta_w)(c_w+c_{0,w}-2+\beta_w)\cdots(c_w+1+\beta_w)(c_w+\beta_w)}{c_{0,w}(c_{0,w}-1)\cdots 2\cdot 1}
\end{aligned} \tag{20}
$$

ディリクレ多項分布（Polya 分布）

- パラメータ
  - $\alpha_1, \dots, \alpha_K)$ s.t. $\alpha_k > 0$ for $1 \le k \le K$
- 確率密度関数

$$p(\boldsymbol{x}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \frac{\prod_k \Gamma(\alpha_k + n_k)}{\Gamma(\sum_k \alpha_k + n_k)} \tag{21}$$

- 導出

$$p(\boldsymbol{x}|\boldsymbol{\alpha}) = \int p(\boldsymbol{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\alpha}) d\boldsymbol{\theta} \tag{22}$$

  - $p(\boldsymbol{x}|\boldsymbol{\theta})$ はパラメータが $\boldsymbol{\theta}$ の多項分布の確率質量関数
  - $p(\boldsymbol{\theta}|\boldsymbol{\alpha})$ はパラメータが $\boldsymbol{\alpha}$ のディリクレ分布の確率密度関数

Contents

まとめ：予測分布とは何だったか

- 事前分布 $p(\boldsymbol{\eta})$ と尤度 $p(\mathcal{D}|\boldsymbol{\eta})$ から、事後分布 $p(\boldsymbol{\eta}|\mathcal{D})$ を求める

$$p(\boldsymbol{\eta}|\mathcal{D}) = \frac{1}{Z_n} p(\boldsymbol{\eta}) p(\mathcal{D}|\boldsymbol{\eta}) = \frac{1}{Z_n} p(\boldsymbol{\eta}) \prod_{i=1}^{n} p(\boldsymbol{x}_i|\boldsymbol{\eta}) \tag{23}$$

  - 事後分布 $p(\boldsymbol{\eta}|\mathcal{D})$ は、モデルパラメータが取りうるあらゆる値に重み付けをしている、と見なせる
- 下記の予測分布 $p(\boldsymbol{x}_0|\mathcal{D})$ を使うと、事後分布 $p(\boldsymbol{\eta}|\mathcal{D})$ によるパラメータへの重み付けを反映するかたちで、未知データ $\boldsymbol{x}_0$ の確率を計算できる

$$p(\boldsymbol{x}_0|\mathcal{D}) = \int p(\boldsymbol{x}_0|\boldsymbol{\eta}) p(\boldsymbol{\eta}|\mathcal{D}) d\boldsymbol{\eta} \tag{24}$$

予測分布の別の見方

- 予測分布を求めることは、観測データがしたがう新しい分布を作り出すこと、とも言える
- 正規分布と正規ガンマ分布から、t location-scale 分布
- 多項分布とディリクレ分布から、ディリクレ多項分布
  - ただし、今回の２つの例のように、予測分布の確率密度関数（or 確率質量関数）を式の計算で求めることができてしまうケースは、少ない。