06t4027h 斎藤高行

11章 ベイズ推定法における モデル選択

11.1 事前確率の設定とモデル選択

11.2 周辺尤度のラプラス近似

ベイズ推定法での事前確率の設定

- ベイズ推定法は、パラメータの事前確率を用いて確率密度関数を推定する
- 未知の事前確率は自分で設定する必要がある
- 得られる推定量は事前確率に依存する

妥当な推定結果を得るためには、事前確率を客観的に決定する必要がある

ベイズ推定法での事前確率の設定

あるパラメータ β によって制御できる事前確率 $p(\theta)$ があるとき、それを $p(\theta;\beta)$ で表す

パラメータ β をハイパーパラメータと呼び、
これを訓練標本 $\{x_i\}_{i=1}^{n}$ によって設定する
(設定するのであって、推定するのではない)

ベイズ推定法での事前確率の設定

最尤推定法のように、ハイパーパラメータ β によって訓練標本 χ が最も生起されやすい場合を考える

訓練標本 χ が生起される確率 $p(\chi;\beta)$ は

$$p(\chi;\beta)=\int\prod_{i=1}^{n}q(x_i\mid\theta)p(\theta;\beta)d\theta$$

ベイズ推定法での事前確率の設定

- 先の式は周辺尤度と呼び、βの関数と見なす
- 周辺尤度の対数をとって負にしたものを自由エネルギーという
- 周辺尤度を最大にするハイパーパラメータ β を用いて事前確率を設定する方法は

  経験ベイズ法、第2種最尤推定法、証拠近似

  と呼ばれる

ベイズ推定法での事前確率の設定

周辺尤度を最大にするハイパーパラメータを
β_{EB} とすると

$$\beta_{EB} := \arg\max_{\beta} \int \prod_{i=1}^{n} q(x_i \mid \theta) p(\theta; \beta) d\theta$$

であり、このときの事前確率は $p(\theta; \beta_{EB})$
と表される

ベイズ推定法でのモデル選択

パラメトリックモデル $q(x|\theta)$ も同様に、周辺尤度を最大にするものを選ぶ

すなわち、モデルの集合の中から
訓練標本 $\{x_i\}_{i=1}^n$ が最も生起されやすいモデルを選ぶ

周辺尤度のラプラス近似

ラプラス近似を用いることで周辺尤度を解析的に近似することができる

$f(\theta)=\prod_{i=1}^{n} q(x_i \mid \theta)p(\theta;\beta)$ とおき

$\int f(\theta)d\theta$ について考える

周辺尤度のラプラス近似

$f(\theta)$ を最大にする θ を $\hat{\theta}_{MAP}$ で表す

$$\hat{\theta}_{MAP} := \arg\max_{\theta} f(\theta)$$

これは9章3節で示した最大事後確率推定量である (→123ページ)

周辺尤度のラプラス近似

$\log f(\theta)$ を $\hat{\theta}_{MAP}$ の周りでテイラー展開する

$$\log f(\theta) = \log f\left(\hat{\theta}_{MAP}\right) + \left(\theta - \hat{\theta}_{MAP}\right)^T \frac{\partial}{\partial \theta} \log f(\theta)\Big|_{\theta=\hat{\theta}MAP}$$

$$+ \frac{1}{2}\left(\theta - \hat{\theta}_{MAP}\right)^T H\left(\theta - \hat{\theta}_{MAP}\right) + \cdots$$

(11.2)

周辺尤度のラプラス近似

ここで、H は $\log f(\theta)$ での $\hat{\theta}_{MAP}$ のヘッセ行列である (ヘッセ行列 : →77ページ)

$$H_{i,j} = \frac{\partial^2}{\partial\theta^{(i)}\partial\theta^{(j)}} \log f(\theta)\Big|_{\theta=\hat{\theta}MAP}$$

$$= \frac{\partial^2}{\partial\theta^{(i)}\partial\theta^{(j)}} \left(\sum_{i=1}^{n} \log q(x_i \mid \theta) + \log p(\theta;\beta) \right) \Big|_{\theta=\hat{\theta}MAP}$$

周辺尤度のラプラス近似

$\hat{\theta}_{MAP}$ が $f(\theta)$ を最大にすることと、log関数の単調性から次が成り立つ

$$\frac{\partial}{\partial \theta} \log f(\theta)\Big|_{\theta=\hat{\theta}MAP} = 0_t$$

したがって、式11.2の1次の項はゼロである

周辺尤度のラプラス近似

式11.2の2次までの項を $\log \hat{f}(\theta)$ とおく

$$\log \hat{f}(\theta) := \log f\left(\hat{\theta}_{MAP}\right) + \frac{1}{2}\left(\theta - \hat{\theta}_{MAP}\right)^T H\left(\theta - \hat{\theta}_{MAP}\right)$$

両辺の指数をとると

$$\hat{f}(\theta) = f\left(\hat{\theta}_{MAP}\right)\exp\left(\frac{1}{2}\left(\theta - \hat{\theta}_{MAP}\right)^T H\left(\theta - \hat{\theta}_{MAP}\right)\right)$$

周辺尤度のラプラス近似

正規分布の確率密度関数の積分は1になる

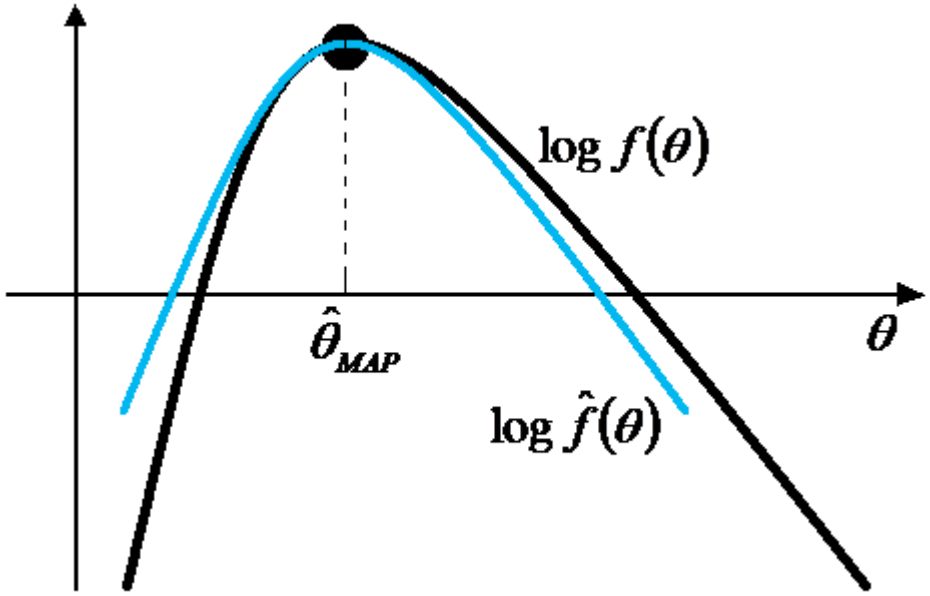
$$\frac{1}{(2\pi)^{\frac{t}{2}}\det(-H)^{\frac{1}{2}}}\int\exp\left(-\frac{1}{2}\left(\theta-\hat{\theta}_{MAP}\right)^{T}(-H)\left(\theta-\hat{\theta}_{MAP}\right)\right)d\theta=1$$

このことを利用することにより、$\hat{f}(\theta)$の積分を
次のように表すことができる (t：θの次元数)

$$\int\hat{f}(\theta)d\theta=f\left(\hat{\theta}_{MAP}\right)\sqrt{\frac{(2\pi)^{t}}{\det(-H)}}$$

周辺尤度のラプラス近似

この $\int \hat{f}(\theta)d\theta$ を $\int f(\theta)d\theta$ の近似とする方法をラプラス近似という



周辺尤度のラプラス近似

ラプラス近似により、周辺尤度の近似が次式で得られる

$$p(\chi;\beta) \approx \sqrt{\frac{(2\pi)^t}{\det(-H)}} \prod_{i=1}^{n} q\left(x_i \mid \hat{\theta}_{MAP}\right) p\left(\hat{\theta}_{MAP};\beta\right)$$

実際は、対数を取ったものを用いることが多い

$$\log p(\chi;\beta) \approx \sum_{i=1}^{n} \log q\left(x_i \mid \hat{\theta}_{MAP}\right) + \log\left(\hat{\theta}_{MAP};\beta\right)$$

$$+ \frac{t}{2}\log(2\pi) - \frac{1}{2}\log(\det(-H))$$

ラプラス近似の特長

事後確率 $p(\chi|\theta)$ がガウス分布に近いとき、ラプラス近似は精度が高くなる

訓練標本数 n が十分に多いとき、中心極限定理により事後確率がガウス分布に収束されることが保証されるため、ラプラス近似による周辺尤度の近似の精度は高い